Steffen Hölldobler et al. (Hrsg.)

# Ausgezeichnete Informatikdissertationen 2018

**Im Auftrag der GI herausgegeben durch die Mitglieder des Nominierungsausschusses**

Sven Apel, Universität des Saarlandes
Abraham Bernstein, Universität Zürich
Felix Freiling, Universität Erlangen-Nürnberg
Steffen Hölldobler (Vorsitzender), Technische Universität Dresden
Hans-Peter Lenhof, Universität des Saarlandes
Paul Molitor, Martin-Luther-Universität Halle-Wittenberg
Gustaf Neumann, Wirtschaftsuniversität Wien
Rüdiger Reischuk, Universität zu Lübeck
Björn Scheuermann, Humbold-Universität zu Berlin
Nicole Schweikardt, Humbold-Universität zu Berlin
Myra Spiliopoulou, Otto-von-Guericke-Universität Magdeburg
Sabine Süsstrunk, École Polytechnique Fédérale de Lausanne
Klaus Wehrle, RWTH Aachen

Gesellschaft für Informatik e.V. (GI)

**Lecture Notes in Informatics (LNI) - Proceedings**
Series of the Gesellschaft für Informatik (GI)

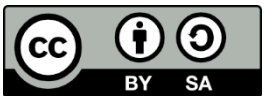
Volume D-19

ISBN 978-3-88579-978-8

**Volume Editors**
Prof. Dr. Steffen Hölldobler
Technische Universität Dresden
01062 Dresden, Deutschland
sh@iccl.tu-dresden.de




**printed by** Köllen Druck+Verlag GmbH, Bonn



# Vorwort

Die Gesellschaft für Informatik e.V. (GI) vergibt gemeinsam mit der Schweizer Informatik Gesellschaft (SI) und der Österreichischen Computergesellschaft (OCG) jährlich einen Preis für eine hervorragende Dissertation im Bereich der Informatik. Hierzu zählen nicht nur Arbeiten, die einen Fortschritt in der Informatik bedeuten, sondern auch Arbeiten aus dem Bereich der Anwendungen in anderen Disziplinen und Arbeiten, die die Wechselwirkungen zwischen Informatik und Gesellschaft untersuchen. Die Auswahl dieser Dissertationen stützt sich auf die von den Universitäten und Hochschulen für diesen Preis vorgeschlagenen Dissertationen. Jede dieser Hochschulen kann jedes Jahr nur eine Dissertation vorschlagen. Somit sind die im Auswahlverfahren vorgeschlagenen Kandidatinnen und Kandidaten bereits „Preisträger“ ihrer Hochschule.

Die 28 Einreichungen zum Dissertationspreis 2018 belegen die Bedeutung und auch die Bekanntheit des Dissertationspreises. Wie jedes Jahr wurden die vorgeschlagenen Arbeiten im Rahmen eines Kolloquiums im Leibniz-Zentrum für Informatik Schloss Dagstuhl von den Nominierten vorgestellt. Für die Mitglieder des Nominierungsausschusses war das persönliche Zusammentreffen mit den Nominierten der Höhepunkt der Auswahlarbeit, und für die Nominierten hat das Kolloquium sicher eine Reihe neuer Erfahrungen und wissenschaftlicher Kontakte geboten. Das wissenschaftlich sehr hohe Niveau der Vorträge, die regen Diskussionen und die angenehme Atmosphäre in Schloss Dagstuhl wurde von allen Teilnehmerinnen und Teilnehmern des Kolloquiums sehr begrüßt.

Wie in jedem Jahr fiel es dem Nominierungsausschuss sehr schwer, eine Dissertation auszuwählen, die durch den Preis besonders gewürdigt wird. Mit der Präsentation aller vorgeschlagenen Dissertationen in diesem Band wird die Ungerechtigkeit, eine aus mehreren ebenbürtigen Dissertationen hervorzuheben, etwas ausgeglichen. Dieser Band soll zudem einen Beitrag zum Wissenstransfer innerhalb der Informatik und von den Universitäten und Hochschulen in die Bereiche Technik, Wirtschaft und Gesellschaft leisten.

Die beteiligten Gesellschaften zeichnen Herrn Dr. Yannic Maus für seine Dissertation „The Power of Locality: Exploring the Limits of Randomnesss in Distributed Computing“ mit dem Dissertationspreis 2018 aus.

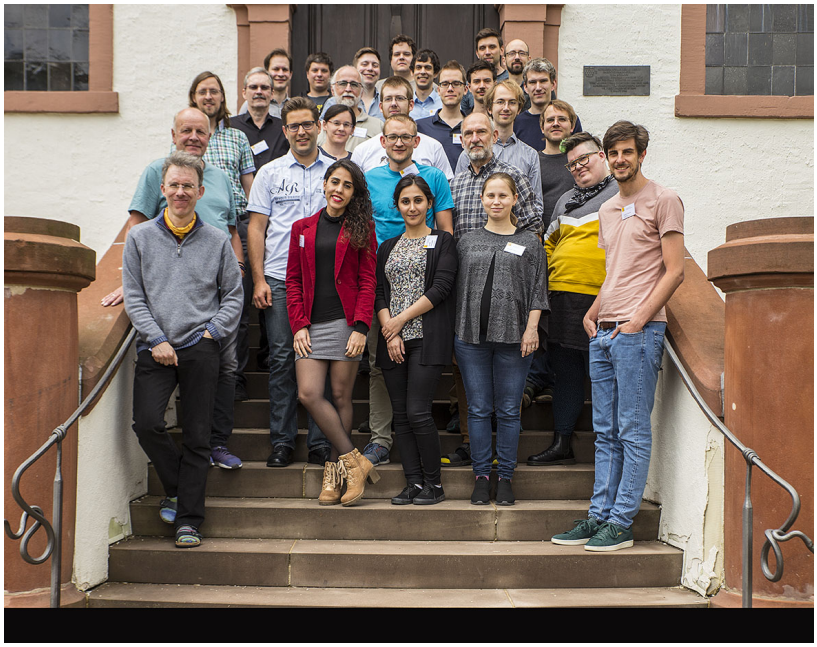
Im Zentrum der Arbeit von Herr Maus steht die Frage, warum die Laufzeiten der schnellsten randomisierten Algorithmen in verteilten Systemen exponentiell schneller sind als die der schnellsten deterministische Algorithmen. Er hat dazu neue Klassen und Techniken eingeführt, die einen signifikanten Fortschritt hin zur Beantwortung der Frage darstellen.

Mit dieser Preisverleihung würdigen die beteiligten Gesellschaften – die Gesellschaft für Informatik e.V. (GI), die Schweizer Informatik Gesellschaft (SI) und die Österreichische Computergesellschaft (OCG) – eine herausragende theoretische Arbeit, mit deren Hilfe randomisierte Algorithmen in verteilten Systemen deutlich besser verstanden werden.

Ein besonderer Dank gilt dem Nominierungsausschuss, der sehr effizient und konstruktiv zusammengearbeitet hat. Bei Frau Emmanuelle-Anna Dietz Saldanha möchte ich mich für die Unterstützung bei der Entgegennahme der vorgeschlagenen Dissertationen, für die Organisation des Kolloquiums sowie für die Zusammenstellung und Anpassung der Beiträge

an das Format der GI-Edition Lecture Notes in Informatik (LNI) bedanken. Für die finanzielle Unterstützung des Nominierungskolloquiums sei den beteiligten Gesellschaften gedankt. Die Gastfreundlichkeit und die hervorragende Bewirtung in Dagstuhl trugen zum Erfolg des Kolloquiums bei, wofür ich mich an dieser Stelle ebenfalls herzlich bedanke.

Steffen Hölldobler
Dresden im August 2019



## Kandidat*innen für den GI-Dissertationspreis 2018

| | |
|---|---|
| Dr.-Ing. Arriel, Juliana | Otto-von-Guericke-Universität Magdeburg |
| Dr. Berrang, Pascal | Universität des Saarlandes |
| Dr. techn. Böhmer, Kristof | Universität Wien / Fakultät für Informatik |
| Dr. rer. nat. Björkelund, Anders | Universität Stuttgart |
| Dr. Bloessl, Bastian | Universität Paderborn |
| Dr. rer. nat. Brachmann, Eric | Technische Universität Dresden |
| Dr. rer. nat. Buschek, Daniel | Ludwig-Maximilians-Universität |
| Dr. Dührkop, Kai | Friedrich-Schiller-Universität Jena |
| Dr. Eichlseder, Maria | Technische Universität Graz |
| Dr. Faessler, Matthias | Universität Zürich |
| Dr. Frömmgen, Alexander | Technische Universität Darmstadt |
| Dr. rer. nat. Herbst, Nikolas | Julius-Maximilians-Universität Würzburg |
| Dr. rer. nat. Immler, Fabian | Technische Universität München |
| Dr.-Ing. Krüger, Julia | Universität zu Lübeck |
| Dr. rer. nat. Maus, Yannic | Albert-Ludwigs-Universität Freiburg |
| Dr. Maystre, Lucas | EPFL - Ecole Polytechnique Federale de Lausanne |
| Dr. rer. nat. Nicolaescu, Ana | RWTH Aachen University |
| Dr.-Ing. Pfaff, Florian | Karlsruher Institut für Technologie |
| Dr. rer. nat. Piatkowski, Nico Philipp | TU Dortmund |
| Dr. Rehr, Robert | Universität Hamburg |
| DI Dr. Rigger, Manuel | Johannes Kepler Universität Linz |
| Dr. rer. nat. Sacha, Dominik | Universität Konstanz |
| Dr. Schlachter, Uli Christian | Carl von Ossietzky Universität Oldenburg |
| Dr. Schütt, Kristof | Technische Universität Berlin |
| Dr. Shirinzadeh, Saeideh | Universität Bremen |
| Dr. techn. Spiel, Katta | TU Wien |
| Dr. Urbat, Henning | TU Braunschweig |
| Dr. van der Heijden, Rens | Universität Ulm |

## Mitglieder des Nominierungsausschusses für den GI-Dissertationspreis 2018

Von links nach rechts:

| | |
|---|---|
| Prof. Dr. Rüdiger Reischuk | Universität zu Lübeck |
| Prof. Dr.-Ing. Felix Freiling | Universität Erlangen-Nürnberg |
| Prof. Dr. Björn Scheuermann | Humbold-Universität zu Berlin |
| Prof. Dr. Abraham Bernstein | Universität Zürich |
| Prof. Dr. Steffen Hölldobler (Vorsitzender) | Technische Universität Dresden |
| Prof. Dr. Gustaf Neumann | Wirtschaftsuniversität Wien |
| Prof. Dr. Hans-Peter Lenhof | Universität des Saarlandes |

Nicht im Bild:

| | |
|---|---|
| Prof. Dr. Sven Apel | Universität des Saarlandes |
| Prof. Dr. Paul Molitor | Martin-Luther-Universität Halle-Wittenberg |
| Prof. Dr. Nicole Schweikardt | Humbold-Universität zu Berlin |
| Prof. Dr. Myra Spiliopoulou | Otto-von-Guericke-Universität Magdeburg |
| Prof. Dr. Sabine Süsstrunk | École Polytechnique Fédérale de Lausanne |
| Prof. Dr. Klaus Wehrle | RWTH Aachen |

# Inhaltsverzeichnis

Steffen Hölldobler et. al. (Hrsg.): Ausgezeichnete Informatikdissertationen 2018,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2019 11

# Personalisierte Recommender Systems für Software-Produktlinienkonfigurationen[1]

Juliana Arriel[2]

**Abstract:** *Software-Produktlinien* (SPLs) werden in der Industrie zur Massenproduktion individualisierter Produkte eingesetzt, um Produktionskosten und Time-to-Market zu reduzieren. Die inhärente Komplexität und Variabilität von SPLs führt jedoch zu einer exponentiell anwachsenden Menge an möglichen Produkten. Gerade bei großen SPLs sind daher Skalierbarkeit und Performanz ein Herausforderung und macht spezialisierte Tool-Unterstützung entscheidend, um Entscheidungsträger bei der Produktkonfiguration zu unterstützen. In diesem Zusammenhang war die Konfiguration von SPLs in den letzten Jahren ein wichtiges Forschungsthema. In dieser Arbeit geben wir einen Überblick über die SPL-Konfigurationstechniken und schlagen einen effizienten, durch Recommender-Systeme unterstützten SPL-Konfigurationsprozess vor, um Entscheidungsträgern genaue und skalierbare Lösungen anzubieten. Dazu passen wir moderne, kollaborative Recommender-Algorithmen an den SPL-Konfigurationskontext an. Zusätzlich führen wir visuelle Unterstützung ein, um Entscheidungsträger durch den Konfigurationsprozess zu führen, indem wir ihnen ermöglichen, sich auf eine begrenzte Anzahl von validen und relevanten Teilen des Konfigurationsraums zu konzentrieren. Wir demonstrieren empirisch die Anwendbarkeit der implementierten Algorithmen und Werkzeuge in verschiedenen realen Szenarien.

## 1 Einführung

Im heutigen kompetitiven Softwaremarkt ist es von größter Wichtigkeit auf die Bedürfnisse einzelnen Kunden einzugehen. Dabei reichen diese Bedürfnisse von funktionalen bis zu nicht-funktionalen Eigenschaften (z.B. Performanz und Kosten) eines Produktes. *Software-Produktlinien* (SPL) haben sich durch ihre systematische Wiederverwendung als effizientes und effektives Mittel für kundenindividuelle Massenproduktion durchgesetzt. SPL beschreiben eine Familie von Softwareprodukten mit einer gemeinsamen Menge von Features. Durch die Kombinationen verschiedener Features wird die Individualisierung von Softwareprodukten für jeden Kunden ermöglicht. Trotz ausführlicher Forschung in den letzten Jahrzehnten bleibt die effiziente Individualisierung von Software weiterhin ein wichtiges Forschungsthema. So müssen zum Beispiel wegen der hohen Variabilität vieler SPL Nutzern durch einfacher und verständlicher Konfigurationsprozess unterstützt werden. Zwar wurden in der Vergangenheit bereits einige *interaktive*, *teil-automatisierte* und auch *automatisierte* Ansätze für den Konfigurationsprozess vorgestellt, allerdings erreichen diese immer noch nicht das volle Potential bei der Unterstützung der Nutzer.

[1] Englischer Titel: "Personalized Recommender Systems for Software Product Line Configurations"

[2] Otto-von-Guericke-Universität Magdeburg, juliana.alves-pereira@ovgu.de

Verstärkt wird dies noch, sobald die Komplexität des Konfigurationsprozess weiter erhöht wird, wie zum Beispiel durch die Einführung von Abhängigkeiten zwischen Features und nicht-funktionalen Eigenschaften.

Typischerweise konfigurieren Nutzer ihr individuellen Produkte über ein Konfigurationswerkzeug, in dem sie nacheinander die gewünschten Features auswählen. In diesem Kontext gibt es viele Ansätze, die Nutzer zu einer validen Konfiguration führen und sicherstellen, das keine Abhängigkeiten der SPL verletzt werden [PCF15]. Allerdings müssen Nutzer bei diesen Ansätzen auch immer wieder Features beachten, die für sie nicht relevant sind [Pe16a]. Dies kann dazu führen, dass die Effizienz des Konfigurationsprozess sinkt, da Nutzer mehr Entscheidungen treffen müssen. Gerade bei hoch konfigurierbaren SPL mit komplexen Abhängigkeiten wird der Konfigurationsprozess langwierig und fehleranfällig. Somit wird zusätzlicher Support notwendig, um Nutzer durch den Konfigurationsprozess zu führen und ihren Fokus auch die für sie relevanten Features zu lenken.

Für einen *teil-automatisierten* Konfigurationsprozess haben Galindo et al. [Ga15] ein dynamisches Entscheidungsmodell mit einer Menge von Fragen und möglichen Antworten vorgeschlagen. Andere Autoren [As14, BE14, TLL14] haben einen Ansatz mit paarweisen Entscheidungen vorgeschlagen, in dem Nutzer jeweils zwei Feature miteinander im Bezug auf ihrer Relevanz für die Erfüllung von nicht-funktionalen Eigenschaften vergleichen. Allerdings können Fragen-basierte Entscheidungsysteme zu vagen und irreführenden Beschreibungen in Fragebögen führen. Hingegen können paarweise Entscheidungen Inkonsistenzen in der Rangfolge der Features verursachen. Wenn Features oder Antworten in den Fragebögen für den Benutzer gleichwertig oder von geringer Relevanz sind, kann der Benutzer bei der Konfiguration nicht weiter unterstützt werden. Da zudem ein einzelnes Feature viele nicht-funktionalen Eigenschaften beeinflussen kann, ist es möglich, dass die Menge und Komplexität an Informationen Benutzer überwältigt und die Konfiguration erschwert.

Für einen *automatisierten* SPL-Konfigurationsprozess existieren Ansätze [Oc17, Oc18], die Constraint Programming und evolutionäre Algorithmen verwenden, um eine Konfiguration, die den Produktanforderungen eines Benutzers entspricht, in einem einzigen Schritt zu generieren. Zwar garantieren exakte Ansätze, wie Constraint Programming eine optimale Konfiguration, aufgrund der rechnerischen Komplexität des Problems haben diese jedoch eine ineffiziente Laufzeit für eine hohen Zahl von Konfigurationsoptionen. Daher wurden heuristische Verfahren, wie unter anderem evolutionäre Algorithmen, eingehender untersucht, um auch für große Konfigurationsräume nahezu optimale Konfigurationen effizient generieren zu können. Ein Nachteil bei der Verwendung von automatisierten Ansätzen ist, dass die Spezifikation mehrerer Anforderungen zu sehr unterschiedlichen Konfigurationen führen kann. So lassen sich beispielsweise bei einem Mobiltelefon nur schwer die beiden Eigenschaften Sicherheit und Kosten gleichzeitig in der selben Konfiguration optimieren. Oft erzeugen automatisierte Ansätze in solchen Situationen mehrere Konfigurationen und überlassen den Benutzern die endgültige Entscheidung. Dabei leiten aktuellen Ansätze

die Benutzer weder bei der Auswahl einer geeigneten Konfiguration, noch bieten sie Unterstützung bei der Festlegung der Benutzerpräferenzen.

Ausgehend von den identifizierten Problemen bisheriger Ansätze besteht der Beitrag dieser Doktorarbeit darin, einen effizienterer Konfigurationsprozess für hoch konfigurierbare SPL vorzuschlagen. Um dieses Ziel zu erreichen, schlagen wir zum ersten Mal ein Verfahren vor, das ein kollaboratives Recommender-System verwendet, das auf Konfigurationen vorheriger Benutzer basiert, um personalisierte Empfehlungen für einen aktuellen Benutzer zu generieren. Darüber hinaus bietet unser System visuelle Unterstützungen, die es Benutzern ermöglicht, sich auf wenige, relevante Informationen aus Teilen des Konfigurationsraums zu konzentrieren. Insgesamt beantworten wir die folgenden Forschungsfragen:

- *RQ1.* Welche Konzepte gibt es zur Unterstützung des SPL-Konfigurationsprozess in der Literatur?
- *RQ2.* Wie lassen sich kollaborative Recommender-Systeme nutzen, um relevante Features anhand von expliziten Nutzerinformationen zu bestimmen?
- *RQ3.* Wie lässt sich der Nutzerkontext in eine personalisiertes kollaborative Recommender-System integrieren?
- *RQ4.* Wie können die Selbstkonfiguration dynamischer SPL durch kollaborative Recommender-Systeme unterstützen?
- *RQ5.* Wie lassen sich die vorgeschlagenen Techniken in ein state-of-the-art Konfigurator integrieren?

Um *RQ1* zu beantworten, führen wir eine *Systematischen Literaturrecherche* (SLR) zum SPL-Konfigurationsprozess durch und klassifizieren einen Korpus von 157 Veröffentlichungen. Basierend auf unseren Erkenntnissen definieren wir eine Reihe offener Probleme in diesem Bereich. Darauf aufbauend untersuchen wir für *RQ2*, wie sich sechs verschiedene, aktuelle kollaborativen Recommender-Algorithmen in den SPL-Konfigurationsprozess integrieren lassen. Zur Untersuchung von *RQ3* erweitern wir diese Algorithmen, so dass sie nicht-funktionale Anforderungen berücksichtigen können. Um *RQ4* zu beantworten schlagen wir einen Tensor-basierten Recommender-Algorithmus vor, der die dynamische Konfiguration von SPL zur Laufzeit durch Modellierung eines N-dimensionalen Tensors *User-Feature-Context* ermöglicht. Bei dem vorgeschlagenen Ansatz werden verschiedene Arten von Kontextdaten als zusätzliche Dimensionen betrachtet. Um die Genauigkeit der vorgeschlagenen Algorithmen abzuschätzen, verwenden wir in unserer empirischen Evaluation mehrere große, reale Konfigurationsdatensätze aus mittleren und großen Produktlinien unterschiedlicher Domänen. Zur Beantwortung von *RQ5* entwickeln wir schließlich einen Open-Source-Konfigurator namens PROFilE. PROFilE ist ein Eclipse-Plug-In, in welchem alle vorgestellten Konzepte und Visualisierungsmechanismen zur Vereinfachung des Konfigurationsprozesses implementiert sind.

## 2 Aktuelle Forschung zur Konfiguration von Software-Produktlinien

Laut Benavides et al. [Be13a] ist der SPL-Konfigurationsprozess ein aktives Forschungsfeld und wurde in den letzten Jahren sowohl von Praktikern als auch Forschern aufgegriffen. Seit der Einführung von Feature-Modellen in den 90er Jahren durch Kang et al. [Ka90] wurden viele neue Techniken, Werkzeuge und Algorithmen bereitgestellt, um Entscheidungsträger im SPL-Konfigurationsprozess zu unterstützen. Diese Ansätze konzentrieren sich jedoch auf sehr spezifische Bereiche, sodass in der Literatur immer noch keine zusammenhängende Übersicht aller Mechanismen des Konfigurationsprozesses und deren Tool-Unterstützung existieren. Daher führen wir eine systematische Literaturrecherche nach den Richtlinien von Kitchenham und Charters [KC07] durch. Wir identifizieren, klassifizieren und bewerten vorhandene Veröffentlichungen zum SPL-Konfigurationsprozess und geben einen vollständigen Überblick über die in diesem Bereich erzielten Fortschritte. Unsere Literaturübersicht stellt dabei eine Ergänzung zu bereits durchgeführten Studien [BSRC10, Be13a] dar.

Insgesamt vergleichen und klassifizieren wir 157 Primärstudien. Wir betrachten verschiedene Details zum SPL-Konfigurationsprozess, wie z.B. die Beschreibung von Konfigurationsbeschränkungen und die Effektivität und Effizienz vorhandene Konfigurationsansätze. Basierend auf dieser Klassifizierung definieren wir einen Satz von 5 Konfigurationsaktivitäten und 17 Konfigurationsmechanismen [Pe17]. Wir geben einen detaillierten Einblick in die Mechanismen, die in jeder Arbeit verwendet werden, wie diese Mechanismen empirisch bewertet werden, sowie ihre Hauptnachteile. Die Ergebnisse unserer Recherche bestätigen, dass die Forschung in der SPL-Konfiguration immer noch fragmentiert und facettenreich ist. So mussten wir feststellen, dass viele Techniken unabhängig voneinander entwickelt wurden und die gleichen Mechanismen adressieren. Unser Übersicht zeigt, dass *(i)* die Qualität der empirischen Bewertung vorhandener Ansätze verbessert werden muss; *(ii)* ganzheitliche Lösungen zur Unterstützung mehrerer Konfigurationsbeschränkungen fehlen; und *(iii)* die Skalierbarkeit und Effektivität existierender Ansätze verbessert werden muss. Angesichts dieser Erkenntnisse und der zunehmenden Bedeutung dieses Forschungsfeldes ergeben sich viele interessante Forschungsmöglichkeiten. Im folgenden Abschnitt greifen wir diese Erkenntnisse auf, indem wir Recommender-Algorithmen verwenden, um Entscheidungsträger bei der Konfigurationsaufgabe einfach und präzise zu unterstützen.

## 3 Personalisierte Konfiguration von Software-Produktlinien

Zur Umsetzung eines personalisierten Konfigurationsprozesses schlagen wir die Verwendung von Recommender-Algorithmen vor. Recommender-Algorithmen können Vorschläge enthalten, die einen großen Konfigurationsraum effektiv beschneiden, sodass Benutzer auf Features hingewiesen werden, die ihren Bedürfnissen und Präferenzen am besten entsprechen. Unser Ansatz zielt auf die folgenden Herausforderungen, abgeleitet aus der bestehenden Literatur und aktuellen Industriebedürfnissen:

- Die Zahl der Features und deren Abhängigkeiten beeinflusst sowohl die Erstellungszeit als auch die Qualität einer Konfiguration. Variabilitätsmodelle reale SPL haben oft

viele und komplexe Abhängigkeiten und Beschränkungen von nicht-funktionalen Eigenschaften (z.B. der Linux-Kernel [Si10]). Obwohl einige Konfiguratoren die Einhaltungen von Abhängigkeiten garantieren können, indem sie implizierte Entscheidungen automatisch treffen, kann dieses Verhalten für Benutzer zu unachvollziehbaren Ergebnissen führen. Folglich kann es für einen Benutzer schwierig sein, eine gültige Konfiguration zu erstellen, die alle ihre (nicht-funktionalen) Anforderungen erfüllt.

- Mithilfe der Auswahl von funktionalen Features lässt sich einfach eine gültige Konfiguration bestimmen, die auf die funktionalen Anforderungen eines Produkts abgestimmt ist. Dabei werden jedoch keine nicht-funktionelle Anforderungen berücksichtigt (z.B. langsame Leistung oder zu hohe Kosten). Obwohl es automatische Ansätze für das Generieren von Konfigurationen gibt, die nicht-funktionale Anforderungen optimieren [Oc17, Oc18], geben diese potentiell eine große Anzahl gültiger Konfigurationen zurück. Eine Auswahl zwischen diesen zu treffen, ist für Entscheidungsträger oft schwierig, da sie viele detaillierte technische Informationen über die Features und deren Kontext berücksichtigen müssen.
- Psychologische Studien haben gezeigt, dass Entscheidungsträger bei einer großen Auswahl an Entscheidungen in Bezug auf die Bedürfnisse der Nutzer in der Regel unsicher sind. Laut einer Industrieumfrage [Be13b], geben etwa 60% der Teilnehmer das Verstehen und die Visualisierung von Variabilitätsmodellen als ein Problem an. Eine weitere Umfrage [Ba17] zeigt, dass 17% der Konfiguratoren explizit angeben, dass sie Einschränkungen bei ihrer Visualisierung haben.

Um diese Herausforderungen zu bewältigen, haben wir einen personalisierten Recommender-Algorithmus eingeführt, der relevante Features aus früheren Konfigurationen bestimmt. Der Hauptbeitrag unserer Arbeit besteht darin, den Konfigurationsprozess zu vereinfachen, indem Entscheidungsträger effizient durch den Konfigurationsprozess geleitet werden. Der in Abb. 1 dargestellte Workflow bietet einen Überblick über den vorgeschlagenen Konfigurationsprozess. Der Konfigurationsprozess wird in fünf Hauptaktivitäten untergliedert: *Configure*, *Propagate Decisions*, *Check Validity*, *Claculate Recommendations* und *Visualize*. Als Eingabe wird das Variabilitätsmodell und die (nicht-funktionalen) Produktanforderungen von Benutzern, Kunden und anderen Interessengruppen verwendet. Dasselbe Variabilitätsmodell wird verwendet, um andere Produkte an die Anwendungsszenarien neuer Kunden anzupassen.

Der Konfigurationsprozess beginnt, indem ein Benutzer einzelne Features aus dem Variabilitätsmodell auswählt. Dabei beschränkt eine fokussierte Ansicht die direkte Auswahlmöglichkeit des Benutzers, da diese zunächst nur abstraktere Features anzeigt. Sobald der Benutzer ein Feature aus- oder abwählt, wird *Decision Propagation* angewendet, sodass alle, durch diese Entscheidung und die Abhängigkeiten im Variabilitätsmodell implizierten, Features automatisch aus- oder abgewählt werden [Ja08]. Damit wird verhindert, dass ein Benutzer zwei widersprüchliche Features auswählen kann [Pe16a]. Als nächstes wird geprüft, ob die gegebene Teilkonfiguration alle Abhängigkeiten des Variabilitätsmodells erfüllt sind oder ob noch weitere Features gewählt werden müssen. Ist die Teilkonfiguration bereits

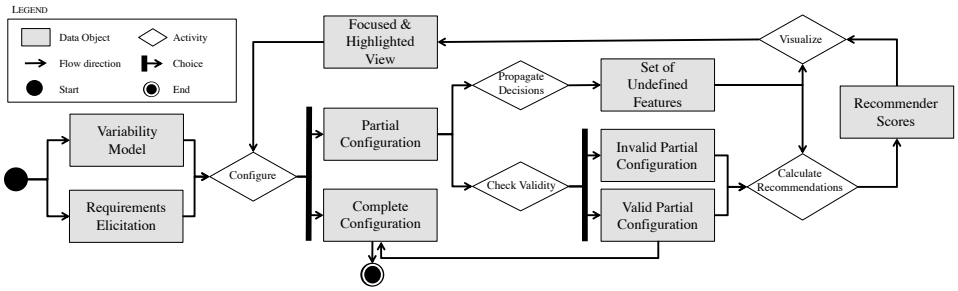



Abb. 1: Graphische Darstellung des Ablaufs des Konfigurationsprozesses.

gültig, so kann der Benutzer den Konfigurationsvorgang abschließen, indem alle bisher unkonfigurierten Features automatisch abwählt werden. Andernfalls muss der Benutzer weitere Features wählen. Dabei zeigt eine hervorgehobene Ansicht dem Benutzer, welche Entscheidungen zu einer gültige Konfiguration führen. Der Prozess endet, sobald eine vollständige Konfiguration erstellt ist. Durch die Anwendung von *Decision Propagation* ist diese immer gültig.

Vor jeder Auswahl des Benutzers bewertet das Recommender-System die Relevanz jedes unkonfigurierten Features (siehe Abschnitt 3.1). Im interaktiven Konfigurationsprozess wird dies als 5-Sterne-Klassifizierung im Konfigurator angezeigt. Im automatischen Konfigurationsprozess wird jeweils das Feature mit der höchsten Relevanz (unter Berücksichtigung der Abhängigkeiten) automatisch ausgewählt. Die Automatisierung unseres Ansatzes erfordert als Eingabe eine Teilkonfiguration mit einem oder mehr ausgewählten Features und mindestens einer vorherigen Konfiguration als Datenbasis. Allerdings steigt die Effektivität unseres Ansatz mit einer umfangreicheren Datenbasis deutlich.

### 3.1 Vorgeschlagene Recommender-Ansätze

Um Empfehlungen zu berechnen, schlagen wir drei Ansätze vor: *Binary-Based Recommender*, *Context-Based Recommender* und *Runtime-Based Recommender*.

**Binary-Based Recommender.** Dieser Ansatz beruht auf der einmaligen Verwendung von Daten aus vorherigen Konfigurationen, um personalisierte Empfehlungen für einen aktuellen Benutzer zu generieren. Wir haben sechs *Collaborative Filtering* (CF) Algorithmen für dieses SPL-Konfigurationsszenario angepasst: *Benutzerbasierte CF*, *CF mit Signifikanzgewichtung*, *CF mit Shrinkage*, *CF with Hoeffding Bound*, *Average Similarity (AS)* und *Matrix Factorization (MF)* [Pe16b, Pe18b].

**Context-Based Recommender.** Unser bisheriger Ansatz weist eine Einschränkung auf: *Die Ergebnisse sind stark abhängig von der Anzahl der bisher konfigurierten Features*. Folglich wird die Empfehlung neuer nützlicher Informationen für Benutzer erschwert. Um diese Einschränkung zu überwinden, kombinieren wir vier Empfehlungstechniken: *context-aware*, *knowledge-based*, *CF-based* und *rule-based*. Ein *context-aware Recommender*

versucht Empfehlungen basierend auf Schlussfolgerungen zu Präferenzen des Benutzers vorzuschlagen. Er verfügt über Kontextwissen über Produktanforderungen, z.B. über den finanziellen Kontext eines Benutzers. Der *knowledge-based Recommender* erstellt eine vollständige Nutzenfunktionen aus vorherigen Konfigurationen. Die Nutzenfunktionen berücksichtigt verschiedene Faktoren, die zur Bewertung einer Konfiguration beitragen, indem die Bedeutung jedes Features für jeden Benutzer gewichtet wird. Der *CF-based recommender* erfasst zusätzlich Ähnlichkeiten zwischen Benutzern auf Grundlage der vorherigen Konfigurationen. Insgesamt haben wir fünf CF-Algorithmen angepasst: *User-Based CF*, *Feature-Based CF*, *User-Based AS*, *Feature-Based AS* und *MF* [Pe18d].

**Runtime-Based Recommender.** Der bisherige Ansatz bietet keine einfache Möglichkeit, Kontextdaten (d.h. nicht-funktionale Eigenschaften [BSRC10]) in das Empfehlungsmodell zu integrieren [KBV09]. Es wird ein reduktionsbasierter Ansatz verwendet, um das Problem der N-dimensionalen ($User \times Feature \times Contexts$)-Empfehlung auf zweidimensionale ($User \times Feature$)-Empfehlung zu reduzieren. Dies führt zwar zu relevanteren Daten für die Berechnung von Empfehlungen, führt jedoch auch dazu, dass Daten verwendet werden, die nur auf Konfigurationen mit dem gleichen oder einem ähnlichen Kontext basieren. Darüber hinaus sind Reduktionsansätze rechenintensiv, da für jede Kombination von Kontextanforderungen ein Empfehlungsmodell trainiert und getestet werden muss. Bei dynamischen SPL, in dem ein System zur Laufzeit neu konfiguriert wird, empfiehlt sich daher die Verwendung eines mehrdimensionalen Ansatzes. Daher verwenden wir statt einer zweidimensionale Matrix einen mehrdimensionalen Tensor und treffen Featureempfehlungen mithilfe von *Tensor Factorization* (TF), einer N-dimensionalen Erweiterung von MF [Pe18c]. TF basiert auf einem einzelnen wenig rechenintensivem Modell und skaliert auf eine beliebige Menge von Kontextinformationen.

### 3.2 Analyse der Ergebnisqualität

Zur Bewertung der Qualität der Empfehlung unseres Recommender-Systems, führen wir eine empirische Evaluation durch. Wir vergleichen in unsere Experimenten zehn Recommender-Algorithmen anhand mehrerer realer Konfigurationsdatensätze. Wir präsentieren im Folgenden nur die wichtigsten Ergebnisse unserer Untersuchungen. Alle Details zum Experimentdesign, die empirischen Bewertung der Algorithmen und zusätzliches Material (Variabilitätsmodell, Konfigurationsdatensätze, Ergebnisse) stellen wir auf unserer Webseite[5] zur Verfügung.

*1. Können Recommender-Algorithmen den SPL-Konfigurationsprozess in realistischen Konfigurationsszenarien unterstützen?* Ja, wir haben gezeigt, dass die implementierten Recommender-Algorithmen relevante Features effektiv identifizieren. Sie liefern sogar bessere Ergebnisse als von Domänexperten in einem interaktiven Konfigurationsprozess. Zudem reduzieren sie den Zeitaufwand und die Fehleranfälligkeit des Prozesses.

[5] http://wwwiti.cs.uni-magdeburg.de/~jualves/PROFilE/

*2. Ab welcher Phase der Produktkonfiguration kann ein Recommender-Algorithmus gute Empfehlungen geben?* Die von uns untersuchten Algorithmen liefern bereits in der Anfangsphase (d.h. mit 10 % der ausgewählten Features) bessere Empfehlungen als unsere Baseline.

*3. Welchen Einfluss haben die implementierten Algorithmen jeweils auf die Qualität der Empfehlungen?* Die Wahl des Algorithmus hat einen großen Einfluss auf die Qualität der Empfehlungen. Für Szenarien, in denen lediglich bisherige Konfigurationen als Datenbasis verfügbar sind, empfehlen wir die Verwendung eines der drei Algorithmen: CF-shrinkage, CF-significance weighting oder MF. Sind zusätzlich Kontextinformationen verfügbar, empfehlen wir die Verwendung von kontextbasierten MF und TF. Unsere Daten zeigen, dass der TF-Algorithmus bei kleinen Datensätzen besser funktioniert, da der kontextbasierte MF-Algorithmus die Datenbasis für seine lokale Vorhersagemodelle noch weiter aufteilt. Folglich basieren bei MF diese Empfehlungen auf einer kleinen Anzahl von Konfigurationen, die auf denselben oder einen ähnlichen Kontext beschränkt sind. Bei dynamischen SPL, in dem sich ein System ständig neu konfiguriert, erscheint ebenfalls die Verwendung eines TF-Algorithmus sinnvoll (siehe Abschnitt 3.1).

*4. Was sind die Vorteile eines kontextabhängigen gegenüber einem kontextunabhängigen Recommender-Algorithmus bezogen auf die Ergebnisqualität?* In unseren Experimenten übertraf die Leistung eines kontextabhängigen Recommender-Algorithmus die eines kontextunabhängigen Algorithmus in allen Phasen des Konfigurationsprozesses deutlich, wobei die Vollständigkeit der Konfiguration außer Acht gelassen wurde. Daher glauben wir, dass kontextabhängige Recommender-Algorithmen für die meisten Anwendungen mit verfügbaren Kontextinformationen kontextunabhängige Algorithmen übertreffen sollten. Welcher dieser beiden Ansätze dominiert, kann jedoch von vielen verschiedenen Faktoren abhängen, wie z.B. der Anwendungsdomäne und den Eigenschaften der verfügbaren Kontextdaten.

*5. Wie lange dauert im Durchschnitt die automatische, vollständige Konfiguration durch einen Recommender-Algorithmus*? Alle Algorithmen hatten in unseren Experimenten eine praktikable Laufzeit (bis zu 112,5 ms für einen 7-dimensionalen Tensor mit über hundert historischen Konfigurationen).

### 3.3 PROFilE: Tool-Unterstützung

Im Rahmen unserer Forschung entwickelten wir den open-source SPL-Konfigurator PROFilE. Dieser ist als Erweiterung des SPL-Tools FeatureIDE [Th14] öffentlich verfügbar. PROFilE bietet eine Funktion zum Ausblenden von Teilen des Variabilitätsmodells mit dem Ziel, relevante Informationen in fokussierten Ansichten darzustellen. [Pe16a]. Weiterhin bietet es eine Reihe von miteinander verknüpften Konfigurationsansichten, die dem Benutzer für eine Menge von Produktanforderungen die Auswirkungen seiner Entscheidungen auf andere Features und auf nicht-funktionale Eigenschaften visualisiert [Pe18a]. Darüber hinaus werden relevante Features bewertet (5-Sterne-Klassifizierung) und nach ihrer Wichtigkeit für den Benutzer kategorisiert (Top-10-Merkmale). Wir bewerten die Leistung von PROFilE

in Bezug auf *Effektivität*, *Skalierbarkeit* und *Effizienz* für elf SPL [Pe18a]. Die Ergebnisse unserer Experimente zeigen, dass die Verwendung eines Recommender-Systems: *(i)* den Konfigurationsprozess vereinfacht, *(ii)* die Erfüllung von Anforderungen an das Endprodukts erhöht und *(iii)* die mentale Belastung für die Entscheidungsträger erheblich reduziert.

## 4 Fazit

Diese Arbeit besteht aus vier Phasen. Erstens, führen wir ein SLR zum SPL-Konfigurationsprozess durch, um die grundlegenden Herausforderungen in diesem Forschungsfeld zu erfassen. Zweitens, stellen wir personalisierte Recomender-Algorithmen für den Konfigurationsprozess vor. Drittens, evaluieren wir die Algorithmen anhand mehrerer realer SPL. Viertens, basierend auf Erkenntnissen aus früheren empirischen Studien ([Co16, Pe13]), erweitern wir einen etablierten Konfigurator mit unseren Recommender-Algorithmen. Darüber hinaus schlagen wir eine Reihe interaktiver und automatisierter visueller Mechanismen vor, die Benutzer bei der Featureauswahl unterstützen und unnötige, sowie ungültige Entscheidungen verhindern. Dadurch können Benutzer: *(a)* sich auf für sie relevante Features fokussieren; *(b)* implizite und explizite Abhängigkeiten zwischen funktionalen und nicht-funktionalen Eigenschaften erkennen; *(c)* auf eine Liste von am meisten empfohlenen Features zurückgreifen; und *(d)* an jedem Punkt im Konfigurationsprozesses die Konfiguration automatisiert vervollständigen lassen. In zukünftige Arbeiten wollen wir erforschen, wie sich die Anzahl der vorherigen Konfigurationen auf die Effektivität der Algorithmen und Visualisierungsmechanismen auswirkt.

## Literaturverzeichnis


[As14] Asadi, Mohsen; Soltani, Samaneh; Gasevic, Dragan; Hatala, Marek; Bagheri, Ebrahim: Toward automated feature model configuration with optimizing non-functional requirements. Information and Software Technology (IST), 56(9):1144–1165, 2014.

[Ba17] Bashroush, Rabih; Garba, Muhammad; Rabiser, Rick; Groher, Iris; Botterweck, Goetz: CASE tool support for variability management in software product lines. ACM Computing Surveys (CSUR), 50(1):14:1–14:45, 2017.

[Be13a] Benavides, David; Felfernig, Alexander; Galindo, José A; Reinfrank, Florian: Automated analysis in feature modelling and product configuration. In: Safe and Secure Software Reuse, S. 160–175. Springer, 2013.

[Be13b] Berger, Thorsten; Rublack, Ralf; Nair, Divya; Atlee, Joanne M.; Becker, Martin; Czarnecki, Krzysztof; Wasowski, Andrzej: A survey of variability modeling in industrial practice. In: Proceedings of the Workshop on Variability Modelling of Software-intensive Systems (VaMoS). ACM, S. 7:1–7:8, 2013.

[BE14] Bagheri, Ebrahim; Ensan, Faezeh: Dynamic decision models for staged software product line configuration. Requirements Engineering Journal (REJ), 19(2):187–212, 2014.

[BSRC10] Benavides, David; Segura, Sergio; Ruiz-Cortés, Antonio: Automated analysis of feature models 20 years later: a literature review. Information Systems, 35(6):615–708, 2010.

[Co16] Constantino, Kattiana; Pereira, Juliana Alves; Padilha, Juliana; Vasconcelos, Priscilla; Figueiredo, Eduardo: An empirical study of two software product line tools. In: International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE). Springer, 2016.

[Ga15] Galindo, José A; Dhungana, Deepak; Rabiser, Rick; Benavides, David; Botterweck, Goetz; Grünbacher, Paul: Supporting distributed product configuration by integrating heterogeneous variability modeling approaches. Information and Software Technology (IST), 62:78–100, 2015.

[Ja08] Janota, Mikolás: Do SAT solvers make good configurators? In: International Systems and Software Product Line Conference (SPLC). Springer, S. 191–195, 2008.

[Ka90] Kang, Kyo C.; Cohen, Sholom G.; Hess, James A.; Novak, William E.; Peterson, A. Spencer: Feature-oriented domain analysis (FODA) feasibility study. Bericht CMU/SEI-90-TR-21, Software Engineering Institute, 1990.

[KBV09] Koren, Y.; Bell, R.; Volinsky, C.: Matrix factorization techniques for recommender systems. Computer, 42:30–37, August 2009.

[KC07] Kitchenham, B.; Charters, S: Guidelines for performing Systematic Literature Reviews in Software Engineering. Citeseer, 2007.

[Oc17] Ochoa, Lina; Pereira, Juliana Alves; González-Rojas, Oscar; Castro, Harold; Saake, Gunter: A survey on scalability and performance concerns in extended product lines configuration. In: Proceedings of the Workshop on Variability Modelling of Software-intensive Systems (VaMoS). ACM, S. 5–12, 2017.

[Oc18] Ochoa, Lina; Gonzalez-Rojasa, Oscar; Pereira, Juliana Alves; Castro, Harold; Saake, Gunter: A Systematic Literature Review on the Semi-Automatic Configuration of Extended Product Lines. Journal of Systems and Software, 2018. Accepted.

[PCF15] Pereira, Juliana Alves; Constantino, Kattiana; Figueiredo, Eduardo: A systematic literature review of software product line management tools. In: International Conference on Software Reuse (ICSR). Springer, S. 73–89, 2015.

[Pe13] Pereira, Juliana Alves; Souza, Carlos; Figueiredo, Eduardo; Abilio, Ramon; Vale, Gustavo; Costa, Heitor Augustus Xavier: Software variability management: an exploratory study with two feature modeling tools. In: Brazilian Symposium on Software Components, Architectures and Reuse (SBCARS). IEEE, S. 20–29, 2013.

[Pe16a] Pereira, Juliana Alves; Krieter, Sebastian; Meinicke, Jens; Schröter, Reimar; Saake, Gunter; Leich, Thomas: FeatureIDE: scalable product configuration of variable systems. In: International Conference on Software Reuse (ICSR), S. 397–401. Springer, 2016.

[Pe16b] Pereira, Juliana Alves; Matuszyk, Pawel; Krieter, Sebastian; Spiliopoulou, Myra; Saake, Gunter: A feature-based personalized recommender system for product-line configuration. In: ACM SIGPLAN International Conference on Generative Programming: Concepts and Experiences (GPCE). ACM, S. 120–131, 2016.

[Pe17] Pereira, J. A.: Personalized Recommender Systems for Software Product Line Configurations. Dissertation, University of Magdeburg, Germany, 2017.

[Pe18a] Pereira, Juliana Alves; Martinez, Jabier; Gurudu, Hari Kumar; Krieter, Sebastian; Saake, Gunter: Visual Guidance for Product Line Configuration Using Recommendations and Non-Functional Properties. In: ACM Symposium on Applied Computing (SAC). ACM, 2018.

[Pe18b] Pereira, Juliana Alves; Matuszyk, Pawel; Krieter, Sebastian; Spiliopoulou, Myra; Saake, Gunter: Personalized Recommender Systems for Product-line Configuration Processes. Computer Languages, Systems & Structures (COMLAN), 2018.

[Pe18c] Pereira, Juliana Alves; Schulze, Sandro; Figueiredo, Eduardo; Saake, Gunter: N-dimensional Tensor Factorization for Self-Configuration of Software Product Lines at Runtime. In: International Systems and Software Product Line Conference (SPLC). ACM, 2018. to appear.

[Pe18d] Pereira, Juliana Alves; Schulze, Sandro; Krieter, Sebastian; Ribeiro, Márcio; Saake, Gunter: A Context-Aware Recommender System for Extended Software Product Line Configurations. In: Proceedings of the Workshop on Variability Modelling of Software-intensive Systems (VaMoS). ACM, S. 1–8, 2018.

[Si10] Sincero, Julio; Tartler, Reinhard; Egger, Christoph; Schröder-Preikschat, Wolfgang; Lohmann, Daniel: Facing the linux 8000 feature nightmare. In: European Conference on Computer Systems (EuroSys). 2010.

[Th14] Thüm, Thomas; Kästner, Christian; Benduhn, Fabian; Meinicke, Jens; Saake, Gunter; Leich, Thomas: FeatureIDE: an extensible framework for feature-oriented software development. Science of Computer Programming (SCP), 79(0):70–85, 2014.

[TLL14] Tan, Lei; Lin, Yuqing; Liu, Li: Quality ranking of features in software product line engineering. In: Asia-Pacific Software Engineering Conference (APSEC). Jgg. 2. IEEE, S. 57–62, 2014.

**Juliana Arriel** wurde am 22. August 1989 in Brasilien geboren. Sie hat ihr Informatik Studium 2012 mit einem Bachelor an der Universität Lavras (Brasilien) und 2014 mit einem Master an der Universität Minas Gerais (Brasilien) abgeschlossen. Im Jahr 2018 promovierte sie an der Otto-von-Guericke Universität Magdeburg mit Auszeichnung. Ihre Doktorarbeit wurde mit dem Fakultätspreis für die beste Dissertation ausgezeichnet. Zum jetzigen Zeitpunkt ist sie Postdoc an der Universität Rennes 1 (Frankreich). Der Fokus ihrer Forschung ist automatisierte Software Entwicklung unter Anwendung und Kombination von Methoden aus der Software Analyse, dem Maschinenlernen und der metaheuristischen Optimierung. Sie ist regelmäßig an der Organisation von bedeutenden internationalen Konferenzen beteiligt. Außerdem publiziert sie in und begutachtet für führende Zeitschriften und Tagungen im Bereich Software Entwicklung. Sie ist Mitglied in der Association for Computing Machinery (ACM) und der Arbeitsgruppe DiverSe (Diversity-Centric Software Engineering) an der Universität Rennes 1.

# Quantifizierung und Schutz der Privatsphäre in der (Epi-)genetik[1]

Pascal Berrang[2]

**Abstract:** Die stetig sinkenden Kosten für molekulares Profiling haben der Biomedizin zahlreiche neue Arten von Daten geliefert und den Durchbruch für eine präzisere und personalisierte Medizin ermöglicht. Die Veröffentlichung dieser inhärent hochsensiblen Daten stellt jedoch eine neue Bedrohung für unsere Privatsphäre dar. Während die IT-Sicherheitsforschung sich bisher hauptsächlich auf die Auswirkungen *genetischer* Daten auf die Privatsphäre konzentriert hat, wurden die vielfältigen Risiken durch andere Arten biomedizinischer Daten größtenteils außer Acht gelassen.

Wir stellen Verfahren zur Messung und Abwehr solcher Privatsphärerisiken vor. Neben dem Genom konzentrieren wir uns auf zwei der wichtigsten gesundheitsrelevanten epigenetischen Elemente: microRNAs und DNA-Methylierung. Wir quantifizieren die Privatsphäre für mehrere realistische Angriffsszenarien. Unsere Resultate bekräftigen, dass die Privatsphärerisiken solcher Daten ernst zu nehmen sind. Zudem präsentieren und evaluieren wir Lösungen zum Schutz der Privatsphäre. Sie reichen von der Anwendung von Differential Privacy bis zu kryptographischen Protokollen.

## 1 Einführung

Seit der ersten vollständigen Genomsequenzierung im Jahr 2001 sind die Kosten für molekulares Profiling stetig gesunken und haben somit den Durchbruch für eine präzisere und personalisierte Medizin ermöglicht. Während dieser Durchbruch durch den enormen Zuwachs an verfügbaren biomedizinischen Daten begleitet und bestärkt wird, zeigt dies auch gleichzeitig die Schattenseite der Entwicklung: neue Privatsphärerisiken im Gesundheitsbereich. Unser Genom ist in dieser Hinsicht besonders betroffen, da es uns nicht nur eindeutig identifiziert, sondern sich auch während des gesamten Lebens kaum verändert. Außerdem lassen sich aufgrund der Vererbung des Genoms sogar Rückschlüsse auf Verwandte ziehen [Hu13]. Diese Besonderheiten erklären vermutlich, wieso sich die IT-Sicherheitsgemeinde bisher hauptsächlich auf die Implikationen *genetischer* Daten fokussiert hat. Verschiedene Angriffsvektoren und Schutzmaßnahmen in diesem Bereich wurden bereits 2014 weitreichend untersucht und kategorisiert [EN14].

Das Genom ist jedoch nicht der einzige Bestandteil des menschlichen Körpers, welcher für unsere Gesundheit von Bedeutung ist. Umgebungseinflüsse wie Umweltverschmutzung, unsere Ernährung und unser Lebensstil sind oft ausschlaggebend für die Entwicklung der meisten Erkrankungen. Multi-omik Ansätze wie Epigenetik, Transkriptomik und Proteomik versprechen genau diese Lücke zwischen dem Genom und unserem Gesundheitszustand zu schließen. Während unser Genom kodiert wie sich Zellen potentiell verhalten können, geben das Epigenom und Transkriptom Aufschluss darüber wie sich eine

[1] Englischer Titel der Dissertation: “Quantifying and Mitigating Privacy Risks in Biomedical Data”

[2] CISPA, Universität des Saarlandes, contact@paberr.net

einzelne Zelle zum aktuellen Zeitpunkt wirklich verhält. In einer Computer Analogie zusammengefasst: wenn das Genom unsere Hardware ist, dann entspricht das Epigenom unserer Software [Cl10]. Genau wie unser Gesundheitszustand variieren epigenetische Daten daher über die Zeit und werden von der Umgebung beeinflusst.

Obwohl die Epigenetik in der Biomedizin ständig an Bedeutung dazu gewinnt, wurden damit verbundene Privatsphärerisiken bisher größtenteils außer Acht gelassen. Mit dem wachsenden Verständnis für epigenetische Daten wird allerdings klar, welche Sensibilität die darin enthaltenen Informationen haben. Es wurden nicht nur Verbindungen zu einer Reihe schwerer Erkrankungen (wie Krebs, Diabetes oder Alzheimer [Wo07, JB07, QM11, FF15]), sondern auch zu der sexuellen Orientierung einer Person [Ne15] hergestellt.

Während genetischen Daten in den meisten Gesetzgebungen ein besonderer Schutz zugesprochen wird, trifft dies jedoch nicht unbedingt auch auf epigenetische Daten zu. Der US Genetic Information Nondiscrimination Act (GINA) beispielsweise bezieht sich explizit auf *genetische* Daten und im wissenschaftlichen Sinn sind diese nicht gleichzusetzen mit *epigenetischen* Daten [RCM09, Dy15].

Unsere Datenschutzbedenken werden außerdem durch die Existenz von Datenbanken verstärkt, in denen verschiedenste Arten von biomedizinischen Daten – z.B. für Forschungszwecke – gesammelt und bereitgestellt werden. Oftmals veröffentlichen Forscher in solchen Datenbanken ihre biomedizinischen Studien. Auf diese Art sind eine Vielzahl epigenetischer Datensätze (in pseudonymisierter Form) bereits heute online für jedermann frei zugänglich. In Anbetracht des Milliarden-Dollar-Geschäfts mit dem Verkauf von privaten Gesundheitsdaten und Brokern, die solche pseudonymisierten Patientendaten aus verschiedenen Quellen untereinander verbinden [Yo, Th], bedarf es zweifelsohne der Möglichkeit, verbundene Privatsphärerisiken zu quantifizieren und abzuwehren.

Meine Dissertation [Be18a] betrachtet Verfahren zur Quantifizierung und Abwehr solcher Privatsphärerisiken. Neben dem Genom konzentrieren wir uns dabei auf zwei der wichtigsten gesundheitsrelevanten epigenetischen Elemente: microRNAs und DNA-Methylierung. Für die Quantifizierung der Privatsphärerisiken betrachten wir mehrere realistische Angriffsszenarien: (1) Verknüpfung von Profilen über die Zeit, Verknüpfung verschiedener Datentypen und verwandter Personen, (2) Feststellung der Studienteilnahme und (3) Inferenz von Attributen. Unsere Resultate bekräftigen, dass die Privatsphärerisiken solcher Daten ernst zu nehmen sind. Außerdem präsentieren und evaluieren wir Lösungen zum Schutz der Privatsphäre, sowohl auf Basis von Veränderung der Daten, als auch auf Basis von Kryptographie. Sie reichen von der Anwendung von Differential Privacy unter Berücksichtigung des Nutzwertes bis zu kryptographischen Protokollen zur Auswertung eines Random Forests. Während Differential Privacy besonders geeignet ist für das Veröffentlichen von Datensätzen und Statistiken, erlauben kryptographische Anwendungen die sichere Speicherung und Analyse von Daten ohne deren Qualität zu beeinflussen.

Die Dissertation basiert auf unseren vorangegangenen Arbeiten, welche allesamt auf Top-Tier-Konferenzen im Bereich der IT-Sicherheit präsentiert wurden [Ba16a, Ba16b, Ba17, Be18b] und welche sich als Leitmotiv durch die Dissertation ziehen.

# 2 Hintergrund

Bevor wir beginnen die Privatsphärerisiken, die mit epigenetischen Daten einhergehen, zu analysieren, geben wir zuerst eine kurze Einführung in die Thematik von Differential Privacy und erklären die Relevanz der Privatsphäre-Nutzen-Abwägung im Hinblick auf biomedizinische Daten und Anwendungsfälle. Im Anschluss fassen wir die wichtigsten Eigenschaften der drei von uns betrachteten Elemente zusammen: dem Genom, microRNAs und DNA-Methylierung.

## 2.1 Differential Privacy

Differential Privacy ist eine Technik, die darauf abzielt die Privatsphäre von Individuen zu schützen, die Teil einer Datenbank sind. Zu diesem Zweck werden die Inhalte der Datenbank unter Hinzufügen von Rauschen so verändert, dass die Zugehörigkeit einzelner Individuen zur Datenbank von einem Angreifer nicht mehr zweifelsfrei festgestellt werden kann. Gleichzeitig soll jedoch die statistische Aussagekraft der Datenbank und damit die Genauigkeit von statistischen Anfragen an die Datenbank maximiert werden. Da diese Technik die Daten verändert, entsteht allerdings ein Konflikt zwischen der Privatsphäre der Teilnehmer und dem Nutzwert der Daten. Der Nutzen der Daten ist dabei stark anwendungsabhängig. In vielen medizinischen Szenarien, wie beispielsweise einer Diagnose, ist der Nutzwert besonders kritisch. Er kann beispielsweise als die Genauigkeit der Diagnose definiert und gemessen werden.

## 2.2 Genetik und Epigenetik

Das Genom enthält das sogenannte Erbgut eines Organismus und verändert sich mit Ausnahme geringster Mutationen über die gesamte Lebensdauer nicht. Es ist Informationsträger und bestimmt die potentiellen Verhaltensmöglichkeiten einer Zelle. Die Genexpression, also die Teile des Genoms welche in einer Zelle aktiv sind, bestimmt das eigentliche Verhalten unserer Zellen. Beim Menschen wird das Genom durch eine Rekombination von den Eltern an ihre Kinder vererbt.

Die Epigenetik ist ein Forschungsgebiet, das sich mit solchen Faktoren befasst, welche die Aktivität und Entwicklung einer Zelle beeinflussen, jedoch nicht auf Veränderungen des Genoms zurückzuführen sind. Solche externen Faktoren sind beispielsweise Chemikalien in der Umgebung, Alterung oder auch Ernährung. Epigentik bezieht sich zudem auf die direkten Veränderungen in der Zelle (z.B. DNA-Methylierung), welche Einfluss auf die Genexpression nehmen, ohne das Genom selbst zu verändern. Im Gegensatz zum Genom bildet die Epigenetik einen Schnappschuss unseres aktuellen Zustands ab und verändert sich somit über die Zeit.

MicroRNAs (abgekürzt auch miRNAs) sind epigenetische Elemente, die ein wichtiger Bestandteil der Genregulation (der Regulation von Genexpression) sind. Studien haben gezeigt, dass die miRNA Expression – also das Vorhandensein bestimmter miRNAs –

in direktem Zusammenhang mit neurodegenerativen Erkrankungen, Herzerkrankungen, Diabetes und einer Vielzahl von Krebsarten steht [Lu05, Wo07, JB07, QM11, FF15].

DNA-Methylierung ist eines der am besten verstandenen epigenetischen Elemente. Es ist ein essentieller Regulator der Gentranskription (d.h. der Synthese von RNA anhand der DNA). Von der Norm abweichende DNA-Methylierungsmuster wurden so mit verschiedensten Krebsarten in Verbindung gebracht [EH02, DS04, Vo13]. Ein DNA-Methylierungsprofil beschreibt an welchen Stellen des Genoms eine Methylierung vorliegt.

## 3 Verknüpfbarkeit von miRNA Expressionsprofilen

Insbesondere aufgrund der zeitlichen Veränderbarkeit epigenetischer Daten untersuchen wir ein Angriffsszenario, in dem epigenetische Profile über die Zeit rückverfolgt und verknüpft werden. Im Speziellen betrachten wir die Verknüpfbarkeit von miRNA Expressionsprofilen. Wir zeigen auf, dass die Variabilität von miRNA Expressionsprofilen keineswegs ein Garant für natürlichen Datenschutz ist und legen somit, als eine der ersten Arbeiten im Bereich epigenetischer Privatsphärerisiken überhaupt, den Grundstein für weitere Forschung.

Wir analysieren die sogenannte *zeitliche Verknüpfbarkeit* von miRNA Expressionsprofilen, indem wir zwei Arten von Angriffen präsentieren und diese daraufhin sorgfältig evaluieren. Der erste Angriff ist ein Identifikationsangriff. Das bedeutet, dass der Angreifer ein miRNA Expressionsprofil seines Ziels gegeben hat und damit das korrespondierende Profil in einer Datenbank sucht, welche solche Profile von einem anderen Zeitpunkt beinhaltet. Der zweite Angriff ist ein Abgleichungsangriff – eine Verallgemeinerung des Identifikationsangriffs. In diesem Fall besitzt der Angreifer bereits Zugriff auf eine Datenbank von miRNA Expressionsprofilen und versucht korrespondierende Profile in einer anderen Datenbank (von einem anderen Zeitpunkt) zu finden. Wie bereits zuvor dargestellt sind solche Datenbanken heute schon Realität. Sowohl durch Forscher, die diese Datenbanken zur Veröffentlichung von Studiendaten verwenden, als auch durch das Hacken oder den Verkauf von Patientendaten.

Unsere Angriffe evaluieren wir auf öffentlich verfügbaren, pseudonymisierten Datensätzen von Verlaufsstudien aus dem Gene Expression Omnibus (GEO) [Ge]. In unseren Experimenten zeigen wir die Effektivität dieser Angriffe. So konnten wir beim Abgleichen von blut-basierten miRNA Expressionsprofilen, die in einem Abstand von einer Woche erfasst wurden, eine Erfolgsrate von bis zu 90% verzeichnen. Außerdem zeigen wir, dass eine Variation des Zeitraums zwischen den Profilen von einer Woche bis zu einem Jahr kaum Einfluss auf die Erfolgsrate des Angriffs hat.

Im Anschluss widmen wir uns dem Schutz vor solchen Angriffen und präsentieren zwei mögliche Abwehrmechanismen, welche die Daten verändern. Unser erster Abwehrmechanismus arbeitet durch gezieltes Verbergen einer Teilmenge der miRNA Expressionen (z.B. solche, die für einen gegebenen medizinischen Anwendungsfall irrelevant sind). Unser zweiter Abwehrmechanismus hingegen nutzt Differential Privacy, um die miRNA Expres-

sionsprofile zu verrauschen. Hierbei kann jeder Teilnehmer selbst den Grad der Verrauschung wählen – beispielsweise bei der Teilnahme an einer Studie.

Wir evaluieren unsere Abwehrmechanismen auf den gleichen Datensätzen, die wir bereits zur Evaluation der Angriffe genutzt haben. Dabei betrachten wir nicht nur die Auswirkungen der Abwehrmechanismen auf die Privatsphäre, sondern auch auf den biomedizinischen Nutzen der Daten. Die Privatsphäre messen wir invers proportional zum Angriffserfolg. Den Nutzwert der Daten schätzen wir anhand einer Klassifizierung der Profile in gesund oder erkrankt mit Hilfe von krankheitsspezifischen Studien. Dabei steht die Genauigkeit der Krankheitsprognose in einem Zielkonflikt mit dem Datenschutz. Unsere Experimente ergeben, dass die Methode auf Basis des Verrauschens diesen Zielkonflikt hier geeignet abwägen kann. Unter Anwendung von Differential Privacy ist es möglich die Verknüpfbarkeit um mindestens 50% zu senken, während die Genauigkeit der Prognose kaum beeinträchtigt ist ($< 1\%$).

## 4 Gruppenzugehörigkeit von miRNA Expressionsprofilen

Eine weitere Gruppe von Angriffen betrifft sogenannte *Gruppenzugehörigkeits-Angriffe*. Diese beziehen sich meist auf biomedizinische Studien, welche zusammengefasste Statistiken (wie z.B. den Durchschnitt) über die genutzten Daten veröffentlichen. Während diese Art von Angriffen bereits für genetische Daten untersucht wurde, sind wir die Ersten, die diese für epigenetische Daten erforschen. Bei einem Gruppenzugehörigkeits-Angriff auf miRNA Expressionsprofilen hat der Angreifer Zugang zu einem solchen Profil und versucht nun die Zugehörigkeit des Profils zu einer Gruppe von Personen anhand zusammengefasster Statistiken zu ermitteln. In anderen Worten: ein Angreifer kann so die Studienteilnahme seiner Zielperson feststellen und somit beispielsweise herausfinden, ob diese Person Teil einer Gruppe erkrankter Personen einer Krebsstudie ist.

Wir präsentieren zwei solche Angriffe: einen Angriff basierend auf der $L_1$ Distanz und einen Angriff basierend auf dem Likelihood-Quotienten-Test. Die Angriffe evaluieren wir auf Statistiken sowohl über krankheitsspezifischen als auch zufällig zusammengesetzten Gruppen. Unsere Ergebnisse demonstrieren, dass krankheitsspezifische Gruppen besonders anfällig für diese Angriffe sein können, mit einer Richtig-positiv-Rate von bis zu 77% bei einer Falsch-negativ-Rate von weniger als 1%. Außerdem zeigen wir, dass der auf dem Likelihood-Quotienten-Test basierende Angriff den größten Angriffserfolg zu verzeichnen hat und leiten eine theoretische Obergrenze für diesen Angriffserfolg her.

Wir stellen zwei Techniken vor, um epigenetische Daten vor solchen Angriffen zu schützen. Ähnlich zu der Verknüpfung von epigenetischen Daten, setzen wir hier wieder auf das Verbergen von Teildaten, sowie Differential Privacy als Alternative. Im Falle von Differential Privacy, betrachten wir außerdem zwei Angreifermodelle mit unterschiedlichen Annahmen über das Wissen des Angreifers. Wir evaluieren unsere Techniken wieder sowohl mit Hinblick auf die Privatsphäre als auch den Nutzwert der Daten. Hierbei kann der auf Differential Privacy basierende Abwehrmechanismus die Privatsphäre in miRNA basierten Studien ohne große Verluste beim Nutzen der Daten schützen, solange die Datensätze verhältnismäßig groß sind. Es zeigt sich, dass der Einfluss des Rauschens auf

den Nutzwert der Daten in solchen Statistiken wesentlich höher ist als bei anderen Arten von Datensätzen (wie beispielsweise solche mit einzelnen Expressionsprofilen). Unsere theoretische Herleitung zeigt, dass der Angriffserfolg sich linear in steigender Anzahl an Gruppenmitgliedern verschlechtert. In Kombination mit der aktuell bekannten Anzahl an miRNAs empfehlen wir daher zusammengesetzte Statistiken über miRNA Expressionsprofile lediglich für Datensätze von einigen hundert Individuen zu veröffentlichen.

## 5 Genotyp-Inferenz aus DNA-Methylierungsprofilen

Während sich unsere bisherigen Untersuchungen auf einzelne Typen biomedizinischer Daten beschränkten, ist es von ebenso immenser Bedeutung Abhängigkeiten zwischen mehreren Arten von Daten zu betrachten. Solche Abhängigkeiten ermöglichen es beispielsweise verschiedene Arten biomedizinischer Daten untereinander zu verknüpfen. Wir untersuchen daher die Abhängigkeiten zwischen dem Genom und dem epigenetischen Element der DNA-Methylierung. Dabei zeigen wir, dass die Veröffentlichung eines DNA-Methylierungsprofils einer Veröffentlichung des eigentlichen Genoms gleicht.

Im Speziellen zeigen wir, dass eine geringe Menge an vom Genom beeinflussten DNA-Methylierungsregionen ausreicht, um die entsprechenden Genotypen abzuleiten. Dies kann daraufhin ausgenutzt werden, um ein DNA-Methylierungsprofil auf das entsprechende Genom abzubilden. Wir formalisieren diese Art eines *Wiedererkennungsangriffs* und stellen einen statistischen Test bereit, der falsche Verknüpfungen erkennen und eliminieren kann.

Wir evaluieren diesen Wiedererkennungsangriff anhand eines großen Datensatzes von Genom- und DNA-Methylierungsdaten, welche von Mutter-Kind-Paaren erfasst wurden. Damit ist es uns auch möglich den Angriff für solche Fälle zu evaluieren, in denen sehr ähnliche Genome von verwandten Personen beteiligt sind. Unsere Ergebnisse zeigen, dass ein Wiedererkennungsangriff selbst dann noch mit 97,5% Genauigkeit möglich ist, wenn ein DNA-Methylierungsprofil mit dem entsprechenden Genom in einer großen Genomdatenbank von über 2500 Genomen verknüpft werden soll. Unser statistischer Test erlaubt es uns weiterhin die wenigen falsch verknüpften Profile zu eliminieren.

Da unsere Evaluation ein erhebliches Privatsphärerisiko in der Veröffentlichung von DNA-Methylierungsprofilen aufzeigt, untersuchen wir potentielle Schutzmaßnahmen für DNA-Methylierungsprofile. Hierzu betrachten wir ein Szenario aus der medizinischen Praxis, in dem Patientendaten erhoben und zur Krankheitsdiagnose analysiert werden. Während der eigene Arzt die Patientendaten erhebt, wird die eigentliche Analyse oftmals von Drittanbietern durchgeführt. In der medizinischen Praxis können DNA-Methylierungsprofile beispielsweise genutzt werden, um den genauen Typ eines Tumors festzustellen. Dies kann mit Hilfe eines Random Forest Klassifizierers geschehen, wie 2015 von Danielsson et al. demonstriert wurde [Da15].

Während bisherige Forschung bereits Protokolle zur kryptographischen, privaten Auswertung von Entscheidungsbäumen entwickelt hat [Bo15], gehen wir ein Stück weiter und präsentieren ein System zur privaten Auswertung von Random Forests, welches in dem oben beschriebenen, typischen Szenario eingesetzt werden kann. Unser Protokoll basiert

auf homomorpher Verschlüsselung und wir beweisen dessen Sicherheit unter Annahme eines ehrlichen, aber neugierigen Angreifers – eine häufige Annahme in diesem Bereich. Das vorgestellte Protokoll ist prinzipell nicht anwendungsspezifisch für epigenetische Daten, sondern kann für jede Klassifikation eines Random Forests verwendet werden. Wir evaluieren unser System anhand echter DNA-Methylierungsprofile und der Random Forest Instanz die von Danielsson et al. vorgestellt wurde. Wir zeigen, dass sich der Kosten-Overhead in einem für unser Szenario akzeptablen Bereich hält.

## 6 Privatsphärerisiken durch Interdependenz biomedizinischer Daten

Unsere bisherigen Resultate haben die Notwendigkeit zur Quantifizierung und zum Schutz der Privatsphäre im Umgang mit biomedizinischen Daten dargelegt. Außerdem haben wir die damit einhergehenden Risiken in Bezug auf spezialisierte Angriffe untersucht. In einer logischen Konsequenz generalisieren wir daher die betrachteten Szenarien und präsentieren eine Methodik, die ein allgemeines Framework zur Quantifizierung von Privatsphärerisiken durch die Interdependenz biomedizinischer Daten darstellt. Dies ist der erste Schritt in Richtung einer umfassenden Betrachtung und Quantifizierbarkeit solcher Privatsphärerisiken im Umgang mit biomedizinischen Daten im Allgemeinen.

Dazu schlagen wir ein Modell basierend auf Bayesschen Netzen vor, welches wir im konkreten Fall sowohl mit genetischen als auch epigenetischen Daten instantiieren. Außerdem umfasst unser Modell die Verwandtheitsbeziehung zwischen Müttern und ihren Kindern, sowie die zeitliche Variabilität von epigenetischen Daten. Wir führen einen generischen Algorithmus zum Lernen der Struktur eines Bayesschen Netzes ein, der vorhandenes Vorwissen mit Trainingsdaten kombiniert, und beweisen dessen Korrektheit. Dieser Algorithmus ist nicht auf unser Modell beschränkt, sondern kann für Bayessche Netze im Allgemeinen verwendet werden. Dieser Algorithmus erlaubt es uns beispielsweise bekannte Relationen in unseren Bayesschen Netzen festzusetzen – wie die Vererbungsregeln des Genoms zwischen Mutter und Kind – und lernt unbekannte Relationen aus vorhandenen Trainingsdaten.

Wir nutzen die konkrete Instantiierung unserer Methodik dann, um eine umfassende Quantifizierung der Privatsphäre vorzunehmen. Hierzu legen wir dem Modell den Datensatz zu Grunde, der bereits für die Genotyp-Inferenz genutzt wurde, nutzen jedoch auch den longitudinalen Aspekt dieses Datensatzes aus. Wir quantifizieren die Privatsphäre anhand etablierter Privatsphäre-Metriken wie der erwartete Abweichung vom korrekten Ergebnis oder der Entropie. Unser Modell erlaubt es uns beliebige Angriffsszenarien durchzuspielen, die Ausgangs-Wissenslage des Angreifers zu variieren und ebenso festzulegen welche Attribute der Angreifer voraussagen möchte. In unseren Experimenten zeigen wir die Flexibilität dieser Methodik auf und bestätigen die inhärenten Privatsphärerisiken durch die Interdepenz der Daten. Wir können spezielle Angriffe reproduzieren und demonstrieren, dass die Privatsphärerisiken ernst zu nehmen sind.

Neben der Inferenz von Attributen dient unser Modell außerdem auch als Grundbaustein für weitere Anwendungen zur Quantifizierung der Privatsphäre. So zeigen wir, wie un-

ser Modell genutzt werden kann, um einen Verknüpfungsangriff zu simulieren. Wir messen auf diese Weise den Erfolg eines Angreifers, der eine Art Mutterschaftstest auf Basis von DNA-Methylierungsprofilen durchführt. Gegeben die DNA-Methylierungsprofile von Müttern versucht der Angreifer die DNA-Methylierungsprofile der entsprechenden Kinder in einer Datenbank auszusondern. Dabei kann der Angreifer aufgrund unseres Bayesschen Netzwerk Modells indirekt auch Wissen über Abhängigkeiten zwischen DNA-Methylierung und Genom ausnutzen, ohne jemals Genomdaten seiner Ziele zu besitzen. Die Evaluierung dieses Verknüpfungsangriffs unterstreicht unsere Privatsphärebedenken mit einer Erfolgsrate von 95%.

## 7 Zusammenfassung

Während sich bisherige Arbeiten im Bereich der Privatsphäre biomedizinischer Daten hauptsächlich auf genetische Daten fokussiert haben, demonstrieren wir in unserer Arbeit die Notwendigkeit für Methoden zur Quantifizierung und zum Schutz der Privatsphäre in Bezug auf weitere biomedizinische Daten – wie z.B. epigenetischer Daten. Indem wir mehrere Angriffsszenarien beleuchten und geeignete Abwehrmechanismen vorschlagen, schaffen wir das Bewusstsein für die Bedeutung dieser Art von Forschung. Wir betrachten unter anderem Verknüpfungs-, Identifikations- und Inferenzangriffe auf Patientendaten. Zudem gehen wir den ersten Schritt in Richtung einer umfassenden Sicht auf solche Angriffe und präsentieren eine generalisierte Methodik zur Quantifizierung der Privatsphäre für untereinander abhängige Daten.

Während die Evaluation dieser Angriffe auch die Werkzeuge liefert, die damit einhergehenden Privatsphärerisiken besser einschätzen zu können, schlagen wir ebenso geeignete Gegenmaßnahmen vor, um das Risiko dieser Angriffe zu mindern. Damit lassen sich die Ergebnisse unserer Arbeiten in drei Kategorien einteilen: (1) Tools zur Quantifizierung der Privatsphärerisiken, (2) Gegenmaßnahmen, welche die Datensätze verändern und somit Einfluss auf den Nutzwert derer haben können, und (3) Gegenmaßnahmen basierend auf Kryptographie. Wir heben dabei hervor, dass es entscheidend ist solche Gegenmaßnahmen in enger Zusammenarbeit mit biomedizinischen Experten zu entwerfen. Nur so können deren Belange beachtet werden und damit Lösungen entwickelt werden, welche in praxisrelevanten Szenarien wirklich verwendet werden können. Insbesondere für Gegenmaßnahmen, welche die Datensätze verändern oder verrauschen, bedeutet dies die kritische Abwägung des Konflikts zwischen Privatsphäre, dem Nutzen der Daten und einer einfach nutzbaren Lösung. Lässt man diesen Konflikt außer Acht, wird man keine Gegenmaßnahme entwickeln können, welche in der Realität Anwendung findet. Im schlimmsten Fall kann das Missachten des Nutzwertes beispielsweise dazu führen, dass die Gesundheit von Patienten durch ungenaue Diagnosen gefährdet wird [Fr14]. So sind Maßnahmen basierend auf dem Verrauschen von Daten hauptsächlich für die Veröffentlichung von Datensätzen geeignet, während Diagnosemaßnahmen eher durch kryptographische Methoden abgesichert werden sollten. Im Falle der von uns vorgestellten Gegenmaßnahmen gehen wir jeweils auf diesen vorgestellten Trade-Off ein und evaluieren eine Vielzahl von Parametern, um eine geeignete Abwägung zu ermöglichen.

## Literaturverzeichnis


[Ba16a] Backes, Michael; Berrang, Pascal; Hecksteden, Anne; Humbert, Mathias; Keller, Andreas; Meyer, Tim: Privacy in Epigenetics: Temporal Linkability of MicroRNA Expression Profiles. In: Proceedings of the 25th USENIX Security Symposium (Security). USENIX Association, S. 1223–1240, 2016.

[Ba16b] Backes, Michael; Berrang, Pascal; Humbert, Mathias; Manoharan, Praveen: Membership Privacy in MicroRNA-based Studies. In: Proceedings of the 23rd ACM Conference on Computer and Communication Security (CCS). ACM, S. 319–330, 2016.

[Ba17] Backes, Michael; Berrang, Pascal; Bieg, Matthias; Eils, Roland; Herrmann, Carl; Humbert, Mathias; Lehmann, Irina: Identifying Personal DNA Methylation Profiles by Genotype Inference. In: Proceedings of the 38th IEEE Symposium on Security and Privacy (S&P). IEEE, S. 957–976, 2017.

[Be18a] Berrang, Pascal: Quantifying and Mitigating Privacy Risks in Biomedical Data. Dissertation, Saarland University, 2018.

[Be18b] Berrang, Pascal; Humbert, Mathias; Zhang, Yang; Lehmann, Irina; Eils, Roland; Backes, Michael: Dissecting Privacy Risks in Biomedical Data. In: Proceedings of the 2018 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE, 2018.

[Bo15] Bost, Raphael; Popa, Raluca Ada; Tu, Stephen; Goldwasser, Shafi: Machine Learning Classification over Encrypted Data. In: Proceedings of the 22nd Annual Network and Distributed System Security Symposium (NDSS). The Internet Society, 2015.

[Cl10] Cloud, John: Why Your DNA Isn't Your Destiny. Time Magazine, 6, 2010.

[Da15] Danielsson, Anna; Nemes, Szilárd; Tisell, Magnus; Lannering, Birgitta; Nordborg, Claes; Sabel, Magnus; Carén, Helena: MethPed: a DNA methylation classifier tool for the identification of pediatric brain tumor subtypes. Clinical Epigenetics, 7(1):1, 2015.

[DS04] Das, Partha M; Singal, Rakesh: DNA methylation and cancer. Journal of clinical oncology, 22(22):4632–4642, 2004.

[Dy15] Dyke, Stephanie OM; Cheung, Warren A; Joly, Yann; Ammerpohl, Ole; Lutsik, Pavlo; Rothstein, Mark A; Caron, Maxime; Busche, Stephan; Bourque, Guillaume; Rönnblom, Lars et al.: Epigenome data release: a participant-centered approach to privacy protection. Genome Biology, 16(1):1–12, 2015.

[EH02] Esteller, Manel; Herman, James G.: Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours. The Journal of Pathology, 196(1):1–7, 2002.

[EN14] Erlich, Yaniv; Narayanan, Arvind: Routes for breaching and protecting genetic privacy. Nature Reviews Genetics, 15(6):409–421, 2014.

[FF15] Feinberg, Andrew P; Fallin, M Daniele: Epigenetics at the Crossroads of Genes and the Environment. JAMA, 314(11):1129–1130, 2015.

[Fr14] Fredrikson, Matthew; Lantz, Eric; Jha, Somesh; Lin, Simon; Page, David; Ristenpart, Thomas: Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In: Proceedings of the 23rd USENIX Security Symposium (Security). USENIX Association, S. 17–32, 2014.

[Ge] Gene Expression Omnibus. `http://www.ncbi.nlm.nih.gov/geo`. Accessed: 29/01/2019.

[Hu13] Humbert, Mathias; Ayday, Erman; Hubaux, Jean-Pierre; Telenti, Amalio: Addressing the concerns of the Lacks family: quantification of kin genomic privacy. In: Proceedings of the 20th ACM Conference on Computer and Communication Security (CCS). ACM, S. 1141–1152, 2013.

[JB07] Jones, Peter A; Baylin, Stephen B: The epigenomics of cancer. Cell, 128(4):683–692, 2007.

[Lu05] Lu, Jun; Getz, Gad; Miska, Eric A; Alvarez-Saavedra, Ezequiel; Lamb, Justin; Peck, David; Sweet-Cordero, Alejandro; Ebert, Benjamin L; Mak, Raymond H; Ferrando, Adolfo A et al.: MicroRNA expression profiles classify human cancers. Nature, 435(7043):834–838, 2005.

[Ne15] Ngun et al., Tuck: Abstract: A novel predictive model of sexual orientation using epigenetic markers. In: American Society of Human Genetics 2015 Annual Meeting. 2015.

[QM11] Qureshi, Irfan A; Mehler, Mark F: Advances in epigenetics and epigenomics for neurodegenerative diseases. Current neurology and neuroscience reports, 11(5):464–473, 2011.

[RCM09] Rothstein, Mark A; Cai, Yu; Marchant, Gary E: The ghost in our genes: legal and ethical implications of epigenetics. Health matrix (Cleveland, Ohio: 1991), 19(1):1, 2009.

[Th] The Black Market For Stolen Health Care Data. `http://www.npr.org/sections/alltechconsidered/2015/02/13/385901377/the-black-market-for-stolen-health-care-data`. Accessed: 29/01/2019.

[Vo13] Vogelstein, Bert; Papadopoulos, Nickolas; Velculescu, Victor E; Zhou, Shibin; Diaz, Luis A; Kinzler, Kenneth W: Cancer genome landscapes. Science, 339(6127):1546–1558, 2013.

[Wo07] Wood, Laura D; Parsons, D Williams; Jones, Siân; Lin, Jimmy; Sjöblom, Tobias; Leary, Rebecca J; Shen, Dong; Boca, Simina M; Barber, Thomas; Ptak, Janine et al.: The genomic landscapes of human breast and colorectal cancers. Science, 318(5853):1108–1113, 2007.

[Yo] Your private medical data is for sale – and it's driving a business worth billions. `https://www.theguardian.com/technology/2017/jan/10/medical-data-multibillion-dollar-business-report-warns`. Accessed: 29/01/2019.

**Pascal Berrang** wurde am 6. Dezember 1991 in Saarbrücken, Deutschland geboren. Er begann sein Informatikstudium 2010 an der Universität des Saarlandes und trat 2013 – direkt nach Abschluss seines Bachelorstudiums – der Graduiertenschule an der Universität des Saarlandes bei. Er promovierte dort am IT-Sicherheitsinstitut CISPA (*Center for IT Security, Privacy and Accountability*) in der Gruppe von Prof. Michael Backes mit einem Fokus auf die IT-Sicherheits und Datenschutz Aspekte von biomedizinischen Daten. Seine interdisziplinäre Forschung erfolgte unter anderem in Kollaboration mit Biomedizinern des DKFZ (*Deutsches Krebsforschungszentrum*) und resultierte in mehreren Top-Tier-Publikationen. Seine Promotion schloss er 2018 mit Auszeichnung ab.

# Verhaltensverifizierung für Geschäftsprozesse basierend auf Testverfahren und Anomalieerkennung[1]

Kristof Böhmer[2]

**Abstract:** Obwohl Geschäftsprozesse (kurz Prozesse) in Organisationen maßgebliche Aufgaben übernehmen, wurden diese bisher nur unzureichend abgesichert. Dies führte dazu, dass Fehler in parallelen Prozessausführungen oder sicherheitskritische Vorfälle unerkannt blieben. Eine Limitierung, welche die Dissertation mittels neuer Verfahren begegnet. Hervorzuheben ist hierbei der Fokus auf ganzheitliche und vollautomatisierte Verfahren, um, unter anderem, das Zusammenspiel aller Prozesse in einer Organisation als Ganzes zu analysieren oder auch komplexe sicherheitskritische Vorfälle, wie kollektive Anomalien, zu erkennen. Durch den breiten Einsatz von Prozessen sind diese Verfahren nicht nur für Prozessexperten relevant, sondern auch für die Gesellschaft als Ganzes.

## 1 Einführung

Prozesse sind das *Herzblut* moderner Organisationen; sie bilden grundlegende Vorgänge ab, verarbeiten *personenbezogene* Daten, *verhindern* Gesetzesverstöße und *koordinieren* Maschinen, Menschen und Unternehmen. Kurz, heutzutage hängen selbst die Lebensmittel- und Energieversorgung direkt oder indirekt von korrekt arbeitenden Prozessen und deren Definitionen ab. Hierbei lassen sich diese als eine Art (grafische) Programmiersprache verstehen, vgl. Abb., 1, welche einen Prozess zur Kreditantragsbearbeitung zeigt.

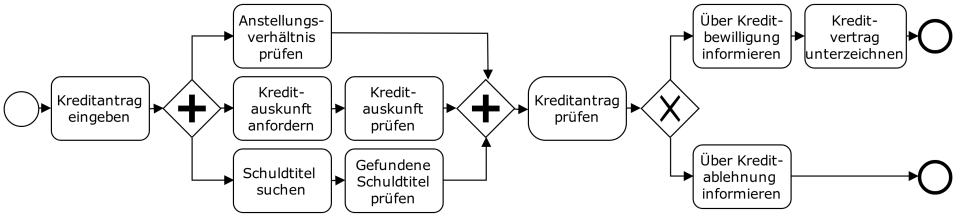



Abb. 1: Beispielhafter Prozess (Kreditantrag) in Business Process Model and Notation

Prozesse durchlaufen einen komplexen Lebenszyklus, welcher sich grob in deren *Definition* und *Ausführung* unterteilen lässt. Während der Definition wird festgelegt, wie die Ziele eines Prozesses erreicht werden können. Hierbei können selbst kleine Fehler eine große Auswirkung haben. Beispielsweise mussten 2015 60% der nordamerikanischen Starbucks-Franchisenehmer Kaffee *kostenlos* „verteilen“, da dieser aufgrund fehlerhafter Zahlungsprozesse von den Kunden nicht bezahlt werden konnte. Dieser und andere prozessbezogene Fehler führten bereits zu Milliardenkosten und haben das Vertrauen zwischen Kunden und betroffenen Organisationen, wie dem US National Grid, Bridgestone

[1] Englischer Titel der Dissertation: “Behavior Verification for Business Processes based on Testing and Anomaly Detection”

[2] Institut für Informatik, Universität Wien, Wien, Österreich, kristof.boehmer@univie.ac.at

oder Starbucks, belastet, vgl. [Zi14]. Daher stellt sich die Dissertation [B8] die Frage, wie solche Fehler frühzeitig erkannt werden können – vor allem deshalb, weil sich diese meist in komplexen, parallelisierten Definitionen „*verstecken*“, für welche sich die Fehlersuche, ohne Unterstützung durch automatische Verfahren, fast aussichtslos gestaltet [B8, S. 2].

Während der Ausführung wird das „grafische Programm“ Prozess instanziiert und durchlaufen. „IT-Sicherheit“ ist hierbei besonders relevant, da Prozesse oft *tief* in bestehende „IT-Landschaften“ integriert sind. Greifen diese doch während ihrer Ausführung auf unterschiedlichste Datenquellen zu und verarbeiten dabei *kritische* geschäftliche und personenbezogene Informationen. Hierdurch werden Prozesse auch zu einem interessantes Ziel für Angreifer und Betrüger, sodass prozessgetriebene Systeme oft auch bei sogenannten „Data Breaches“ beteiligt sind. Bei letzterem werden private und geschäftliche Daten (z.B. medizinische Analysen) aus einer Organisation unberechtigterweise ausgelesen. Untersuchungen aus dem Jahr 2016 zufolge sind hiervon jährlich 4,2 Milliarden Datensätze und Millionen von Personen betroffen, vgl. [Cy16]. Solche Vorfälle können, neben einem Vertrauensverlust auf Seiten der Kunden und Partner, auch in Geldstrafen resultieren. Die Dissertation stellt Verfahren bereit, um Angriffe auf Prozesse und deren missbräuchliche Verwendung *automatisch* zu erkennen – was durch die steigende Komplexität, Flexibilität und Dynamik der Prozessausführungen erschwert wird. Gilt es doch, beispielsweise echte Angriffe von ungewöhnlichen, aber *harmlosen Verhalten* und Fehlbedienung möglichst gut zu unterscheiden – obwohl beide einander auf den ersten Blick ähneln [B8, S. 120].

Bisher fehlte es jedoch an gründlichen Verfahren, um Prozesse *a*) automatisch auf *Fehler* zu prüfen und *b*) vor unerwünschter oder *missbräuchlicher Verwendung* zu *schützen*. Diese Forschungslücken werden von der Dissertation mittels *systematischer Literaturstudien* konkretisiert und passende Lösungsverfahren identifiziert, implementiert und evaluiert. Unter anderem wird hierzu *Machine Learning* eingesetzt, um für jeden Prozess zu bestimmen, wie in diesem am schnellsten Fehler gefunden werden können. Weiters werden neuartige *wahrscheinlichkeitsbasierte Algorithmen* entworfen, um unwahrscheinliches Ausführungsverhalten (Anomalien) – und damit potentielle Sicherheitsprobleme – zu identifizieren. Die hierbei entstandenen Forschungsbeiträge sind *nicht nur* für Sicherheits- oder Prozessexperten relevant, da die Gesellschaft und die in ihr operierenden Organisationen mehr und mehr von fehlerfreien sicheren Prozessen abhängig werden [B8, S. 12].

## 2 Hauptinhalt der Dissertation

In den letzten Jahren sind *Millionen von Prozessdefinitionen* für Bereiche wie Forschung, Entwicklung oder Bildung entstanden; allein ein großes Bergbauunternehmen wie BHP Billiton nutzt mehr als 100.000 davon. Durch die starke Verbreitung und häufig unkontrollierte (voll-)*automatische Ausführung* der Prozesse wird die Auswirkung von Fehlern in ihren Definitionen und Anomalien in deren Ausführungen stark erhöht und kann sich auch zu einer direkten Gefahr für Leib und Leben auswachsen, indem z.B. ein Fehler zur Zusammenstellung eines unverträglichen Medikamenten-Cocktails führt. Diese Dissertation schlägt daher Verfahren vor, um Prozesse auf Fehler (Abschn. 2.1) und Anomalien (Abschn. 2.2) möglichst schnell, gründlich und vollautomatisch zu überprüfen [B8, S. 4].

## 2.1 Erster Hauptteil: Fehlererkennung während der Definition eines Prozesses

Eine Standardlösung, um Fehler zu identifizieren, ist der Einsatz sogenannter Tests; einer Zusammenstellung von Daten, um *Prozesse auszuführen* und daraufhin zu überprüfen, ob diese das *erwartete Verhalten* zeigen. Diese Idee wird in unterschiedlichen Ausprägungen angewandt, z.B. um Prozessverhalten während Hochlastszenarien zu prüfen oder sicherzustellen, dass einzelne Prozessbestandteile korrekt zusammenarbeiten. Aufgrund dieser Diversität geht diese Dissertation zunächst der Frage nach, welche Testverfahren derzeit im Prozessbereich eingesetzt werden und welche Schwächen/Stärken diese aufweisen.

### 2.1.1 Systematische Literaturanalyse und Forschungslücken

Hierzu wurden, im Rahmen einer *systematischen Literaturanalyse*, Forschungsdatenbanken (wie SCOPUS, IEEE Xplore und DBLP) und 30 Journals/Konferenzbände nach prozessfokussierten Testverfahren durchsucht. 6638 Arbeiten wurden hierbei als potentiell relevant erkannt, von denen nach *mehreren Auswahlrunden* 159 detailliert analysiert wurden. Es zeigte sich, dass die derzeitige Forschung aus dem Bereich „Prozess-Tests“ stark von verwandten Gebieten wie den „Software-Tests“ beeinflusst wird. Dabei wird eine Vielzahl von Themen abgedeckt, wie Testgenerierung, Integrationstests, Regressionstests oder die Überprüfung vorgegebener Dienstgüteanforderungen [B8, S. 28 ff.].

Darüber hinaus stellen viele existierende Verfahren überraschend *umfangreiche Anforderungen* an deren Anwender (z.B. indem seltene formale Sprachen eingesetzt werden). Dies erschwert unserer Annahme nach deren Einsatz, weshalb in dieser Dissertation darauf geachtet wurde, einen hohen Automatisierungsgrad zu erreichen, sodass die vorgestellten Verfahren weitestgehend ohne Schulungen breit eingesetzt werden können. Abschließend zeigte sich auch ein *durchwachsenes Bild* hinsichtlich der *durchgeführten Evaluierungen* – wiesen doch viele Arbeiten keine oder nur eine unzureichende Evaluierung mit kleinen, selten frei verfügbaren, selbst generierten Datensätzen auf, siehe Abb. 2. Um diesem Trend zu begegnen, wurden die im Rahmen der Dissertation vorgestellten Verfahren *prototypisch implementiert* und mit öffentlichen *realen und realistischen Daten* evaluiert [B8, S. 49 ff.].

### 2.1.2 Testauswahl mittels Machine Learning

Bezüglich der von Software-Test-Verfahren inspirierten Arbeiten zeigte sich, dass diese kaum auf die *Unterschiede* zwischen Quellcode und Prozessdefinitionen, wie der unterschiedlichen Verhaltensgranularität, eingehen [B8, S. 27]. Dies führt dazu, dass die derzeit angewendeten Testverfahren Prozesse nur mit einer *unzureichender Gründlichkeit* analysieren und entsprechend Fehler übersehen. Um diese Lücke zu schließen, wurde im Rahmen der Dissertation ein neues, auf Machine Learning basierendes Verfahren vorgestellt. Dieses erlaubt es, Prozessdefinitionen und deren Bestandteile *a*) nach *Anwenderwunsch* (z.B. einer maximal möglichen Testausführungsdauer) und *b*) einer automatisch errechneten notwendigen *Testintensität* nach Fehlern zu durchsuchen [B8, S. 57 ff.].

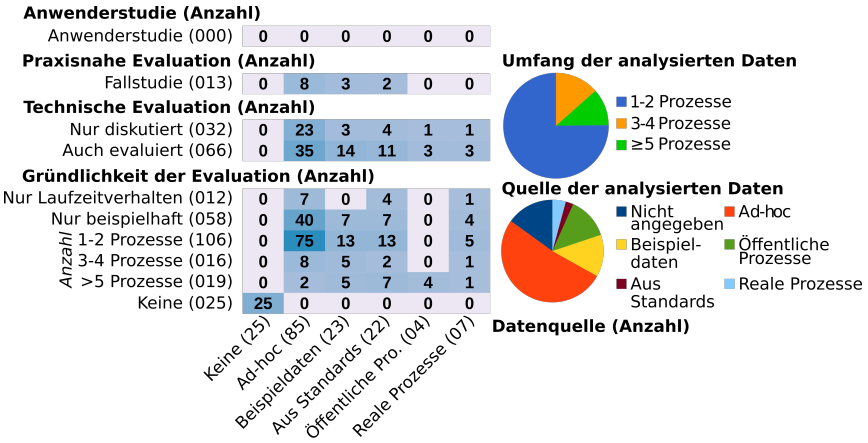



Abb. 2: Zusammenfassung der Evaluierungsqualität der analysierten Publikationen

In der Dissertation wird das vorgeschlagene Verfahren mit fünf bereits etablierten Verfahren – welche überwiegend aus dem Forschungsbereich der Softwareentwicklung in den Bereich Prozessfehlersuche übertragen wurden – verglichen. Hierbei konnte das vorgeschlagene Verfahren je nach Datensatz und Aufgabenstellung zwischen 3,5 und 10,3 Prozent mehr Fehler und Probleme identifizieren als die etablierten Vergleichsverfahren. Zusätzlich wurde noch erhoben, wie das vorgestellte Verfahren im Fall von Änderungen reagiert, beispielsweise weil nach einem erkannten Fehler die analysierte Prozessdefinition geändert wurde. Solche Änderungen passieren häufig und werden oft auch durch *externe Faktoren* wie Gesetzesänderungen erzwungen – eine hohe Performance in solchen Fällen ist daher maßgeblich für die praktische Anwendbarkeit eines Testverfahrens. Hier zeigte sich im Vergleich zur vollständigen Neuberechnung eine *Reduktion* der zu investierenden Rechenleistung zwischen 21,8 und 58513,5 Prozent. Gerade in Zeiten großer und umfangreicher Testsammlungen mit tausenden Einzeltests ist dies wichtig, um Tests auch bei immer knapper werdenden Entwicklungszyklen noch einbinden zu können [B8, S. 69].

### 2.1.3 Verifizierung von (versteckten) hoch parallelen Verhalten

Eine besondere Herausforderung bei der Fehlersuche stellt die den Prozessen inhärente Parallelität während deren Ausführungen dar. Insgesamt konnten zwei „Arten“ von Parallelität identifiziert werden. Einerseits die *explizite* Parallelität, welche während der Definition der Prozesse durch parallele Ausführungspfade (explizit) definiert wird. Selbst diese Art führt oft zu Fehlern aufgrund ungenügend berücksichtigter Überlappungen und damit einhergehender Nebeneffekte. Zusätzlich konnte die Dissertation die sogenannte *implizite* Parallelität identifizieren, die *bis jetzt* noch keine Berücksichtigung erfuhr [B8, S. 83 ff.].

Implizite Parallelitäten können ebenfalls die Ursache für zahlreiche Fehler sein, sind je doch, im Vergleich zu expliziten Parallelitäten, deutlich schwieriger zu *erkennen und einzuplanen*. Hervorgerufen wird dies durch den Umstand, dass diese, im Gegensatz zu ex-

pliziten Parallelitäten, nicht absichtlich eingebracht werden, sondern durch die häufige Parallelausführung mehrerer Prozessinstanzen „nebenbei“ (implizit) entstehen. Implizite Parallelitäten umfassen entsprechend unvorhersehbare, wechselnde und potentiell riskante Überlappungen und parallele Datenzugriffe, die aus dem normalen Geschäftsalltag und den dabei stattfindenden parallelen Ausführungen mehrerer Prozessinstanzen erwachsen.

Um implizite Parallelitäten zu identifizieren und effizient auf Fehler zu überprüfen, unterteilt die Dissertation Prozessdefinitionen in mehrere Teilbereiche, je nachdem, ob sich diese mit anderen Prozessdefinitionen während deren Ausführung nicht, teilweise, oder vollständig überlappen. Hierzu werden alle *historischen*, vollautomatisch aufgezeichneten Prozessausführungen einer Organisation analysiert, deren Ausführungsverhalten extrahiert und ermittelt, ob und wie es zwischen mehreren Prozessdefinitionen zu impliziten Parallelitäten kommt und wie *wahrscheinlich* diese in Fehlern resultieren. Dies erlaubt es, anschließend automatisch eine Sammlung von Tests zusammenzustellen, welche mit möglichst *geringem Zeitaufwand* besonders fehlerträchtige Teile von Prozessdefinitionen intensiv auf durch implizite Parallelitäten verursachte Fehler prüfen können [B8, S. 90 f.].

Insgesamt konnten während der Evaluierung in sechs realen Datensammlungen mehrere *hunderttausend implizite Parallelitäten* identifiziert werden [B8, S. 93]. Anschließend wurde das neu vorgeschlagene Verfahren mit zwei aus dem Software Engineering Bereich abstammenden Standardverfahren zur Zusammenstellung von prozessfokussierten Tests verglichen. Es zeigte sich, dass das vorgestellte Verfahren es erlaubt, mit impliziten Parallelitäten in Zusammenhang stehende Fehler *effizienter zu identifizieren* und so die für Testausführungen aufzuwendende Zeit von Tagen auf Stunden reduziert werden konnte, vgl. Abb. 3 und [B8, S. 94]. Die Datensätze stammen aus Callcentern (TeleClaim) und niederländischen Baubehörden (BPIC15).

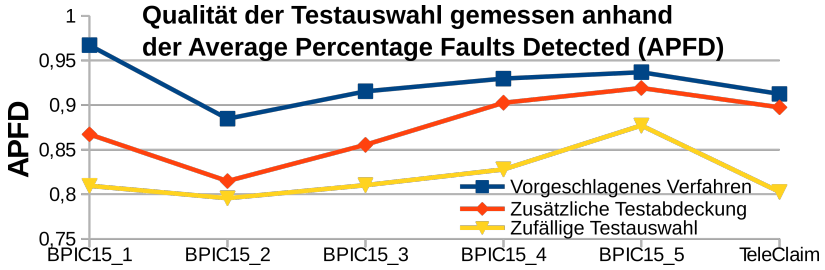



Abb. 3: Vergleich des vorgestellten Verfahrens mit zwei Alternativen

### 2.1.4 Demonstration der entwickelten Verfahren in den Energienetzen der Zukunft

Die zuvor vorgestellten Verfahren wurden im Rahmen des geförderten Forschungsprojektes PROMISE[2] in der *Praxis erprobt*. Hierbei konzentriert sich die Dissertation auf Prozesse zur Abrechnung und Steuerung von intelligenten Energienetzen und sogenannten Smart Metern. Letztere stellen eine *kritische Infrastrukturkomponente* dar, da Fehler in den zugehörigen Prozessen in landesweiten Blackouts resultieren können [B8, S. 97 ff.].

[2] Österreichische Forschungsförderungsgesellschaft (FFG), Projektnummer 849914

Eine besondere Herausforderung erwächst hierbei daraus, dass die verwendeten Prozesse äußerst komplex und verschachtelt sind und zahlreiche externe Datenquellen und Hardwarekomponenten zur Steuerung und Analyse des Netzzustands miteinbeziehen bzw. ansprechen. Gerade Letzteres zu berücksichtigen ist äußerst aufwändig, sodass dieser Aspekt von bestehenden Verfahren zumeist ignoriert oder sehr stark abstrahiert wird, was die Aussagekraft der Fehlersuchergebnisse beschränkt. Um dieser Einschränkung zu begegnen, wurden die zuvor beschriebenen Verfahren und Erkenntnisse mit Verfahren aus dem Bereich des Process-Mining kombiniert. Es konnte gezeigt werden, dass hierdurch reale Prozesse, Daten und Ausführungsumgebungen während der Fehlersuche flexibel miteinbezogen werden können. Dies ermöglicht es, alle Phasen einer Prozessdefinition (von den ersten Entwürfen bis hin zum Echtbetrieb) *nahtlos mit Fehlersuchmaßnahmen* zu begleiten. Hierdurch können Fehler frühzeitig und kostengünstiger behoben werden, als wenn erst nach Fertigstellung einer Prozessdefinition mit der Fehlersuche begonnen würde.

## 2.2 Zweiter Hauptteil: Anomalieerkennung während der Prozessausführung

Nach der Definition eines Prozesses wird das durch diesen beschriebene Verhalten zumeist voll- oder teilautomatisch von prozessgesteuerten IT-Systemen ausgeführt. Unter anderem werden hierdurch Energienetze gesteuert, die Produktionsabläufe in Fabriken koordiniert und medizinische Analysen standardkonform realisiert. Die hierzu eingesetzten Prozesse sind aus IT-Sicherheitssicht als *äußerst schützenswert* anzusehen. Prozesse benötigen und erhalten doch oft einen direkten Zugriff auf zahlreiche Systeme, Datenspeicher und IT-Landschaften innerhalb von (Partner-) Organisationen und bieten so einen vielversprechenden Einstiegspunkt für Angriffe und Betrügereien. In diesem Abschnitt wird daher die Frage geklärt, inwiefern die Ausführung von Prozessen überwacht werden kann, um Ausführungsverhalten zu identifizieren, welches auf Betrug, sicherheitskritische Ereignisse oder (un-)absichtliche fehlerhafte Verwendung hindeutet (sogenannte Anomalien).

### 2.2.1 Systematische Literaturanalyse und Forschungslücken

Um auch mit diesem Teil der Dissertation gezielt Forschungslücken zu identifizieren und zu schließen, wird auch dieser mit einer systematischen Literaturanalyse eingeleitet. Diese ist unseres Wissens nach die Erste in diesem Forschungsbereich, welche es erlaubt, einen vollständigen Überblick über den stark zersplitterten Bereich der Prozessverhaltensanalyse zu erlangen. Die erstellte Literaturanalyse identifizierte, basierend auf mehreren Forschungsdatenbanken (unter anderem Google Scholar, DBPL und IEEE Xplore), mehrere hundert potenziell relevante Publikationen, von denen 35 schlussendlich als relevant erkannt und näher analysiert wurden. Hierbei zeigte sich, dass das Interesse an der Thematik gestiegen ist bzw. die relevanten Publikationszahlen in den letzten Jahren zugenommen haben (3 im Jahr 2014 im Vergleich zu 10 im Jahr 2018). Derzeit scheint noch kein dominierendes Verfahren gefunden worden zu sein, da ein bunter *Mix an Technologien* und Konzepten mit unterschiedlichen Stärken, Schwächen und Zielen eingesetzt wird (z.B. statistische Analysen, neuronale Netze oder auch regelbasierte Verfahren), siehe Abb. 4.

Es zeigt sich, dass die existierenden Verfahren überwiegend die Analyse eindimensionaler Daten und die Erkennung einfacher „Point Anomalies“ (Punkt Anomalien) unterstützen. Unter Letzteren werden Anomalien verstanden, welche anhand eines punktuell stark herausstechenden Verhaltens identifiziert werden können (z.B. eine einzelne extrem hohe Überweisungssumme). Da Angreifer aber zumeist versuchen, ihre Aktivitäten zu verstecken, erachtet die Dissertation beispielsweise „Collective Anomalies“ (Kollektive Anomalien) als *deutlich realitätsnäher*. Bei diesen werden Angriffe in mehrere einzelne, für sich gesehen unauffällige, Einzelaktionen aufgespalten [B8, S. 111 ff.]. Nur Verfahren, die in der Lage sind, diese zu korrelieren und als Einheit zu analysieren, können diese sicher erkennen und eine umfassende Schutzwirkung gewährleisten. Ergo hat die Dissertation eine Reihe von neuartigen Verfahren vorgestellt, um solche und andere Beschränkungen aufzuheben – auf diese wird im Folgenden eingegangen [B8, S. 118-121].

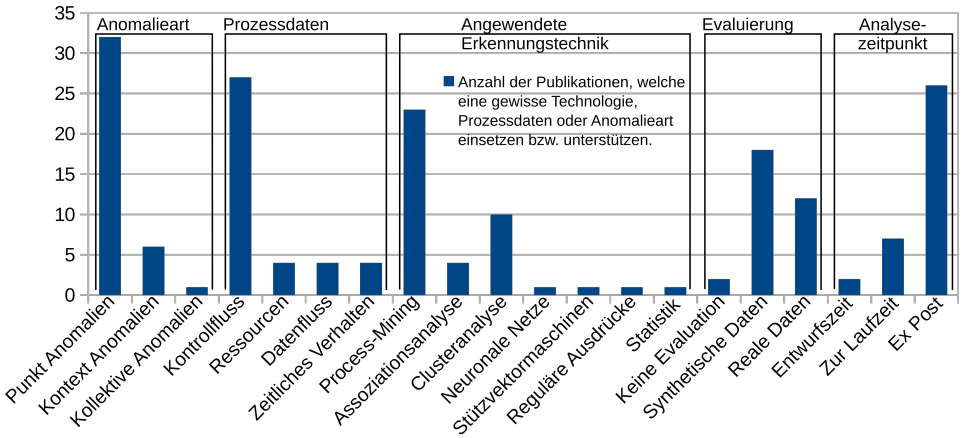



Abb. 4: Überblick über bestehende Anomalieerkennungsarbeiten für Prozesse

### 2.2.2 Anomalieerkennung in den ausgetauschten und verarbeiteten Daten

Prozesse verarbeiten und tauschen während ihrer Ausführung typischerweise eine Vielzahl von Datensätzen mit zahlreichen Systemen, Diensten und Partnerorganisationen aus. Jeder dieser Datensätze stellt eine potenzielle Bedrohung dar, da er dazu genutzt werden könnte, sicherheitskritisches Verhalten oder die Ausführung von Schadcode zu forcieren. Gegen solche Bedrohungen sind Prozessausführungen zumeist *nicht abgesichert*, nimmt doch die flexible, schnelle und unkomplizierte Einbindung unterschiedlicher Datenquellen einen höheren Stellenwert als IT-Sicherheit ein. Auch fehlen oft eine entsprechende Dokumentation und Standardisierung, um beispielsweise händisch Regeln aufstellen zu können, welche es erlauben, zwischen validen und bedrohlichen Datensätzen zu unterscheiden.

Die Dissertation löst diese Herausforderung mittels eines neuartigen Verfahrens, welches ausnützt, dass Prozessausführungen zumeist lückenlos aufgezeichnet werden. Anhand solcher Aufzeichnungen werden anschließend *automatisch Regeln* (in der Form von regulären Ausdrücken) abgeleitet, welche es erlauben, automatisch festzustellen, ob Datensätze hin-

sichtlich Struktur und Inhalt dem zuvor aufgezeichneten bzw. gewohnten/erwarteten Verhalten folgen. Durch den Einsatz von *regulären Ausdrücken* können die hierbei in zwei frei wählbaren Komplexitätsleveln anfallenden Regeln mit geringem Aufwand von Menschen gelesen und adaptiert werden um individuelle Anpassungen durchzuführen [B8, S. 126].

Das zuvor beschriebene Verfahren wurde in Abstimmung mit den Sicherheitsexperten von SBA Research evaluiert, um sicherzustellen, dass die angenommenen Bedrohungsszenarien als *relevant und realistisch* einzuschätzen sind. Insgesamt wurden hierbei 240.000 verschiedene Datensätze in drei unterschiedlichen Formaten (EDIFACT, XML, JSON) und verschiedenen Bedrohungsszenarien evaluiert. Als sicherheitskritisch einzustufende Datensätze konnten hierbei mit 59%-100% Genauigkeit (je nach Format und Bedrohungsszenario) korrekt erkannt werden, siehe Abb. 5. Das Verfahren ermöglicht es, hierbei auch automatisch zu ermitteln, wie stark ein als bedrohlich eingestufter Datensatz vom Erwarteten abweicht, um entsprechend auf *wechselnde Bedrohungen* zu reagieren [B8, S. 139].

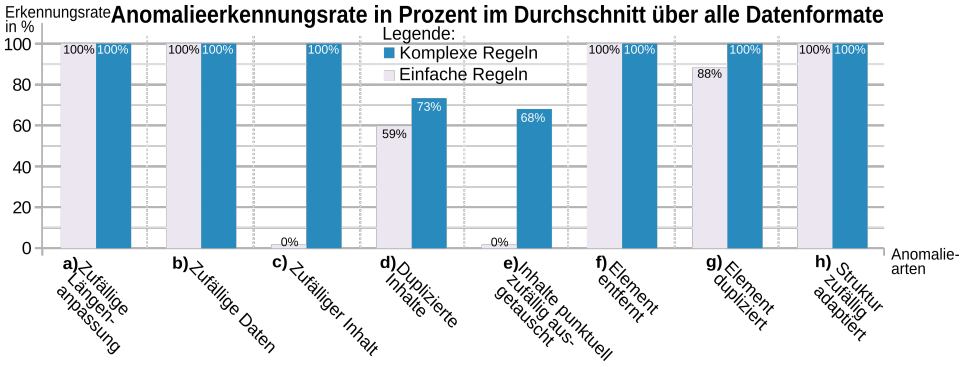



Abb. 5: Evaluierungsergebnisse zur Anomalieerkennung in Prozessdaten

### 2.2.3 Berücksichtigung aller Aspekte eines Prozesses

Das im vorangegangenen Abschnitt beschriebene Verfahren konzentriert sich vor allem auf die während einer Prozessausführung anfallenden und ausgetauschten Datensätze. Damit allein ist jedoch noch keine vollständige Absicherung möglich. Sind doch auch Aspekte wie z.B. die in einen Prozess eingebundenen (menschlichen) Ressourcen oder der während einer Ausführung durchlaufene Kontrollfluss sicherheitsrelevant – beispielsweise, um Verletzungen des Vier-Augen-Prinzips zu erkennen. Um eine vollständige Absicherung zu erreichen, müssen diese und weitere Aspekte *ganzheitlich Berücksichtigung* erfahren.

Die Dissertation bietet hierzu ein Verfahren, welches erlaubt, basierend auf aufgezeichneten Prozessausführungen die *Wahrscheinlichkeit* jedes möglichen Ausführungsverhaltens zu berechnen. Basierend auf der gängigen Annahme, dass Anomalien mit unwahrscheinlichem Verhalten gleichzusetzen sind, lassen sich diese hierüber identifizieren. Dieses neuartige Verfahren lässt sich nicht nur flexibel erweitern (je nachdem, welche Informationsquellen/Aspekte gerade zur Verfügung stehen), sondern erlaubt es auch, einzelne Ereignisse zu korrelieren. Hierdurch würde beispielsweise eine Kombination von leicht un-

wahrscheinlichen Ereignissen genauso als Sicherheitsproblem identifiziert werden wie ein einzelnes, sehr unwahrscheinliches Ereignis. Hierdurch können nicht nur versteckte Angriffe („Collective Anomalies“) identifiziert werden, sondern es ist auch möglich, Analysen während einer noch laufenden Prozessausführung durchzuführen (Ad-hoc) [B8, S. 163]. Vergleichbare Verfahren begannen bis zu diesem Zeitpunkt immer erst mit der Analyse, nachdem eine Prozessausführung vollständig abgeschlossen worden ist (Ex Post). Jedoch wurde zu diesem Zeitpunkt das sicherheitskritische Verhalten des Prozesses bereits vollständig durchlaufen und so z.B. ein System bereits erfolgreich attackiert. Im Gegensatz hierzu ermöglicht das im Rahmen der Dissertation vorgestellte Verfahren automatisch, während der Laufzeit des Prozesses jederzeit dessen Status einzuschätzen, sodass Angriffe frühzeitig, bereits in ihrer *Anfangsphase, gestoppt* werden können [B8, S. 112 ff.].

Um die Anwendbarkeit und Fähigkeiten des neuen Verfahrens zu evaluieren, wurden mehrere reale Datensätze herangezogen und auf Anomalien hin analysiert. Hierbei konnten diese mit einer Präzision von 82% gefunden werden. Allerdings werden auch teilweise Anomalien übersehen, was in einem Recall von 65% und Accuracy von 78% resultiert. Letzteres wurde hierbei von den befragten Sicherheitsexperten aber als wenig relevant beurteilt, da es ihnen für die angedachten Einsatzszenarien als wichtiger erschien, dass gemeldete Anomalien wirklich Anomalien sind, als dass alle Anomalien gefunden wurden, um den derzeit durch *Fehlalarme entstehenden Aufwand* zu minimieren [B8, S. 163].

### 2.2.4 Vollständige Absicherung einer Organisation

Bestehende Prozessanomalieerkennungsverfahren analysieren jeden Prozess und dessen Ausführungen individuell, unabhängig von allen anderen in einer Organisation anfallenden Prozessausführungen. Dies vereinfacht und beschleunigt den Analysevorgang, erlaubt es jedoch *Angriffe zu verstecken*, indem diese in mehrere „harmlose“ Bestandteile aufgeteilt und durch eine Kombination mehrerer Prozesse umgesetzt werden. Bis jetzt wurden solche Angriffe und die damit in Zusammenhang stehenden Anomalien nicht erkannt.

Die Dissertation überwindet diese Beschränkung durch eine Erweiterung des in Abschnitt 2.2.3 beschriebenen Verfahrens mit Algorithmen zur *Zeitreihenanalyse* [B8, S. 169]. Während hierdurch einerseits das Zusammenspiel mehrerer Prozesse (beispielsweise deren Reihenfolgen und Überlappungen) analysiert wird, kann andererseits die individuelle Wahrscheinlichkeit jeder Prozessausführung bestimmt werden. Beide Analysen zu kombinieren ermöglicht es nun, trotz großer heterogener Datenmengen alle Prozessausführungen in einer Organisation – samt deren Zusammenspiel – in ihrer Gesamtheit zu analysieren.

Das zuvor beschriebene Verfahren wurde anhand mehrerer realer, frei verfügbarer Datensätze evaluiert. Hierbei wurden Anomalien mit einer Genauigkeit von 78% erkannt. Hervorzuheben ist hierbei, dass dieses Ergebnis *trotz starker Fluktuationen* im Ausführungsverhalten erreicht werden konnte. Letzteres schlug sich unter anderem in wechselnden Ausführungszeiten und unterschiedlichen parallelen Ausführungen nieder welche die Unterscheidung zwischen (un-)wahrscheinlichem Verhalten erschwerte [B8, S. 176].

## 3 Zusammenfassung und Ausblick

Es zeigte sich, dass aufgrund der Komplexität und Vielschichtigkeit von Prozessdefinitionen und Ausführungen ein einzelnes isoliertes Verfahren alleine kaum ausreichend ist, um „alle“ Fehler oder „alle“ Anomalien zu identifizieren. Stattdessen ist es notwendig, *mehrere Verfahren* zu kombinieren. Dies wird mit den Verfahren in Abschnitt 2.2.4 und 2.1.4 für den Bereich Fehlererkennung bzw. der Erkennung von Sicherheitsproblemen demonstriert. Wir sehen es als wichtig an, weitere Verfahren zu entwickeln, welche sich, ähnlich der hier vorgestellten, à la *Plug und Play* einfach kombinieren lassen [B8, S. 187].

Die Ergebnisse dieser Dissertation fokussieren sich auf Prozesse und die im Rahmen von Prozessausführungen anfallenden Daten. Jedoch sind diese auch für *andere Datenquellen* und Forschungsbereiche relevant. Prozessähnliches Verhalten bzw. ähnliche Herausforderungen und Fragestellungen treten auch während der Entwicklung und Ausführung von Software, der Auswertung von Logfiles und innerhalb von Systemen zur Steuerung von Maschinen und von Telekommunikations- und Energienetzen auf. Letztere sind hierbei besonders interessant, da deren Verhalten mittels prozessähnlicher Methoden modelliert wird und auch bereits zentrale Datensammel- und Ausführungsplattformen existieren, die überwacht werden können. Erste Gespräche mit Sicherheitsunternehmen wie TendMicro über potentielle Interessenten aus dem *kanadischen Telekommunikationsbereich* verliefen vielversprechend und zielen darauf ab, Betrugsfälle und Fehler in den für die Anruf- und Netzverwaltung relevanten Systemen und Prozessen zu identifizieren [B8, S. 187 f.].

## Literaturverzeichnis

[BRM18] Böhmer, Kristof; Rinderle-Ma, Stefanie: Association Rules for Anomaly Detection and Root Cause Analysis in Process Executions. In: International Conference on Advanced Information Systems Engineering. Springer, S. 3–18, 2018.

[B8] Böhmer, Kristof: Behavior Verification for Business Processes based on Testing and Anomaly Detection. Dissertation, Universität Wien, 2018.

[Cy16] CyberScout: ITRC Data Breach Reports. Technischer Report, 2016.

[Zi14] Zibelman, Audrey: PSC completes audit of national grid gas companies. Technischer Report, 2014.

**Kristof Böhmer** studierte Informatik und Wirtschaftsinformatik an der FH Technikum Wien und an der Universität Wien. Während des Studiums arbeitete er als Softwareentwickler und anschließend als wissenschaftlicher Mitarbeiter. Für [BRM18] wurde ihm der *Best Paper Award* von der “Conference on Advanced Information Systems Engineering 2018“ verliehen. Seine Beteiligung an der Lehre und Konzeptionierung neuer Lehrveranstaltungen führte dazu, dass er seine Forschungstätigkeit nun als Senior Lecturer an der Universität Wien fortsetzt, um eine Habilitation zu erlangen.

# Inkrementelles Lernen latenter linguistischer Strukturen durch approximative Suche

Anders Björkelund[1]

**Abstract:** Diese Dissertation setzt maschinelles Lernen ein, um aus linguistisch annotierten Textsammlungen Modelle für die Analyse natürlicher Sprache zu trainieren (Natural Language Processing, NLP), beispielsweise für die Ermittlung der syntaktischen Struktur von Sätzen in einem Textdokument. Die Suchräume, die sich in diesem Bereich ergeben, sind in der Regel so umfangreich, dass exakte Suchverfahren nicht in Frage kommen. In der Praxis muss jeder Ansatz einen Kompromiss finden zwischen der Ausdrucksstärke der verwendbaren Features einerseits und der Effizienz des Suchverfahrens andererseits.

Die Arbeit präsentiert ein Meta-Framework, das sich für unterschiedliche Aufgaben aus der maschinellen Sprachverarbeitung instantiieren lässt und das es jeweils erlaubt, mit alternativen Strategien beim maschinellen Lernen zu experimentieren. So wird eine systematische Untersuchung von Verfahren des inkrementellen Lernens mit latenten linguistischen Strukturen und approximativen Suchverfahren ermöglicht. Es zeigt sich, dass etablierte Methoden aus der aktuellen NLP-Forschung sehr empfindlich sind, was die Wahl von Updatemethoden betrifft. Wir schlagen neue Updatemethoden vor und zeigen, dass diese zu mindestens gleichwertigen Ergebnissen führen, in einigen Fällen jedoch zu erheblichen Qualitiätsverbesserungen. Die Resultate tragen zu einem vertieften Verständnis der Charakteristika unterschiedlicher Analyseaufgaben bei.

## 1 Einführung

Bei der automatisierten Analyse natürlicher Sprache werden in der Regel maschinelle Lernverfahren eingesetzt, um verschiedenste linguistische Information wie beispielsweise syntaktische Strukturen vorherzusagen. **Structured Prediction** (dt. etwa Strukturvorhersage; [Sm11]), also der Zweig des maschinellen Lernens, der sich mit der Vorhersage komplexer Strukturen wie formalen Bäumen oder Graphen beschäftigt, hat deshalb erhebliche Beachtung in der Forschung zur automatischen Sprachverarbeitung gefunden.

In manchen Fällen ist es vorteilhaft, die gesuchte **linguistische Struktur** nicht direkt zu modellieren und stattdessen **interne Repräsentationen** zu lernen, aus denen dann die gewünschte linguistische Information abgeleitet werden kann. Da die internen Repräsentationen allerdings selten direkt in Trainingsdaten verfügbar sind, sondern erst aus der linguistischen Annotation inferiert werden müssen, kann es vorkommen, dass dabei mehrere äquivalente Strukturen in Frage kommen. Anstatt nun vor dem Lernen eine Struktur beliebig auszuwählen, kann man diese Entscheidung dem Lernverfahren selbst überlassen, welches dann selbständig die für das Modell am besten passende auszuwählen lernt. Unter diesen Umständen bezeichnet man die interne, nicht a priori bekannte Re-

[1] Department of Astronomy and Theoretical Physics, Lund University, anders.bjorkelund@thep.lu.se

präsentation für eine gesuchte Zielstruktur als **latent** (vgl. das Beispiel in Abb. 1 für die Aufgabe der Koreferenz-Resolution).

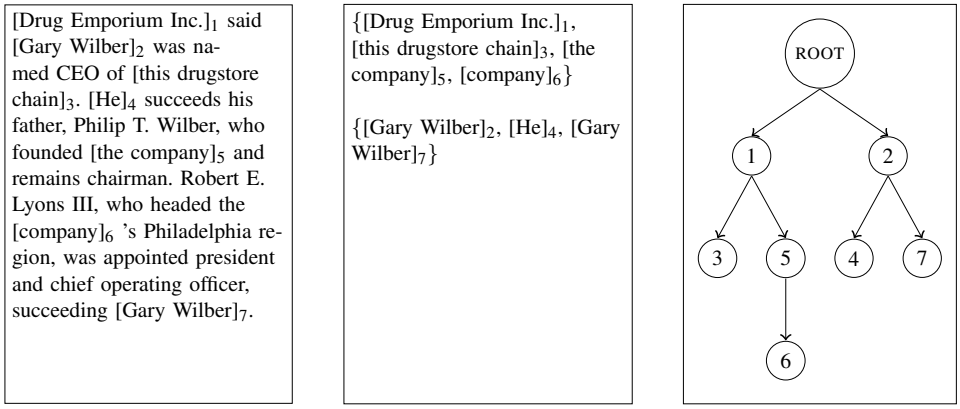



Abb. 1: *Illustration der Eingabe-, Ausgabe- und Vorhersageräume am Beispiel der Aufgabenstellung der Koreferenz-Resolution: links ist als Eingabebeispiel ein Textdokument mit markierten referentiellen Phrasen dargestellt, in der Mitte die Ausgabestruktur der Koreferenz-Resolution: eine Partition sämtlicher referentieller Phrasen in Mengen mit dem gleichen Referenten. Rechts ist eine (mögliche) latente Baumstruktur dargestellt, die sich im Zuge des Lernprozesses als effektiv erwiesen haben mag (die Entscheidung, eine gegebene Phrase der einen oder einer anderen existierenden Partition zuzuschlagen kann so auf möglichst zuverlässiger Basis erfolgen).*

Diese Dissertation stellt ein **Structured Prediction Framework** vor, mit dem man den Vorteil latenter Repräsentationen nutzen kann und welches gleichzeitig von konkreten Anwendungsfällen abstrahiert. Diese Modularisierung ermöglicht die Wiederverwendbarkeit und den Vergleich über mehrere Aufgaben und Aufgabenklassen hinweg. Um das Framework auf ein reales Problem anzuwenden, müssen nur einige Hyperparameter definiert und einige problemspezifische Funktionen implementiert werden.

Das vorgestellte Framework basiert auf dem **Structured Perceptron** [Ro58, Co02]. Der Perceptron-Algorithmus ist ein inkrementelles Lernverfahren (engl. online learning), bei dem während des Trainings einzelne Trainingsinstanzen nacheinander betrachtet werden. In jedem Schritt wird mit dem aktuellen Modell eine Vorhersage gemacht. Stimmt die Vorhersage nicht mit dem vorgegebenen Ergebnis überein, wird das Modell durch ein entsprechendes Update angepasst und mit der nächsten Trainingsinstanz fortgefahren. Das Structured Perceptron wird im vorgestellten Framework mit **Beam Search** kombiniert. Beam Search ist ein approximatives Suchverfahren, welches auch in sehr großen Suchräumen effizientes Suchen erlaubt (s. Abb. 2; hier wird *Beam Search* im Vergleich zu einem radikaleren approximativem Verfahren, *Greedy Search* illustriert). Es kann aus diesem Grund aber keine Garantie dafür bieten, dass das gefundene Ergebnis auch das optimale ist. Das Training eines Perceptrons mit Beam Search erfordert deshalb besondere Update-Methoden, z.B. *Early-* [CR04] oder *Max-Violation-Updates* [HFG12], um mögliche Vorhersagefehler, die auf den Suchalgorithmus zurückgehen, auszugleichen.

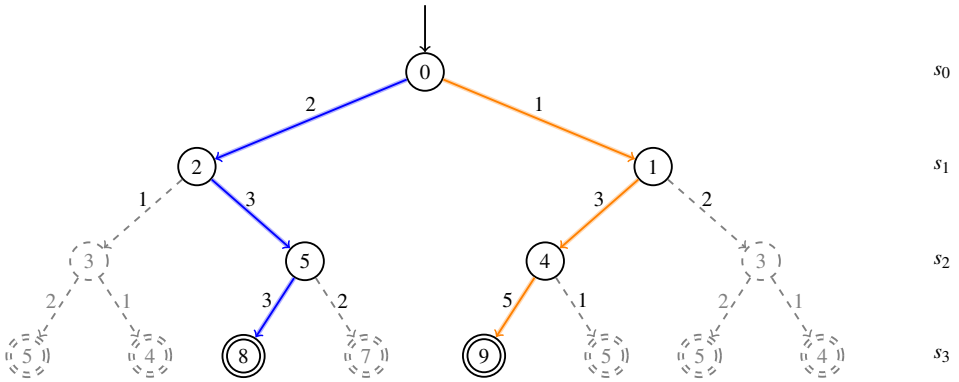



Abb. 2: *Beispiel eines Suchbaums (mit einer positiven* Score*-Wertung an den Kanten; gesucht ist der Pfad mit der größten Summe); in diesem Beispiel führt eine* Greedy Search*-Strategie zu einem suboptimalen Ergebnis (blauer Pfad), da der optimale (orange-farbene) Pfad erst spät zu einer höheren Wertung führt. Ein* Beam Search*-Verfahren mit* Beam*-Weite 2 würde die beiden farbig hervorgehobenen Pfade berücksichtigen (und die gestrichelten Pfade verwerfen).*

Das Framework ist angelegt auf einen systematischen Vergleich des Effekts verschiedener *Machine Learning*-Strategien: approximative Suchverfahren (mit **Beam Search** oder **Greedy Search**) werden gegenübergestellt mit exakten Suchverfahren (mit eingeschränkter Ausdrucksstärke der einsetzbaren Features); für approximative Suchverfahren können unterschiedliche Update-Methoden verglichen werden. Über die in der Literatur vorgeschlagenen Methoden hinaus werden weitere Techniken entwickelt, insbesondere die Erweiterung der **LaSO**-Methode ("Learning as Search Optimization") von [DIM05] als "Delayed LaSO" (**DLaSO**). Abb. 3 illustriert den Effekt unterschiedlicher Update-Methoden anhand eines abstrakten Suchbaums.

## 2 Behandelte NLP-Strukturvorhersageaufgaben

Das Framework wird in der Dissertation auf drei NLP-Aufgaben angewandt, für die ein überwachtes Lernen jeweils mit einer latenten Vorhersagerepräsentation umgesetzt werden kann: Koreferenzresolution [BK14], Dependenzparsing [BN15] und Dependenzparsing mit gleichzeitiger Satzsegmentierung [Bj16].

Das vorgestellte Modell zur **Koreferenzresolution** ist eine Erweiterung eines existierenden Modells [FdSM12], welches Koreferenz mit Hilfe latenter Baumstrukturen repräsentiert (wie oben bereits in Abb. 1 illustriert). Dieses Modell wird um Features erweitert, mit denen nicht-lokale Abhängigkeiten innerhalb eines größeren strukturellen Kontexts modelliert werden. Die Modellierung nicht-lokaler Abhängigkeiten macht durch die kombinatorische Explosion der Features die Verwendung eines approximativen Suchverfahrens notwendig. Es zeigt sich aber, dass das so entstandene Koreferenzmodell trotz der approximativen Suche dem Modell ohne nicht-lokale Features überlegen ist, sofern hinreichend gute Update-Verfahren beim Lernen verwendet werden. Für das **Dependenzparsing**

verwenden wir ein transitionsbasiertes Verfahren, bei dem Dependenzbäume (Abb. 4) inkrementell durch Transitionen zwischen definierten Zuständen konstruiert werden [Ni08]. Im ersten Schritt erarbeiten wir eine umfassende Analyse des latenten Strukturraums eines bekannten Transitionssystems, nämlich ArcStandard mit Swap [Ni09]. Diese Analyse erlaubt es uns, die Rolle der latenten Strukturen in einem transitionsbasierten Dependenzparser zu evaluieren. Wir zeigen dann empirisch, dass die Nützlichkeit latenter Strukturen von der Wahl des Suchverfahrens abhängt – in Kombination mit Greedy-Search verbessern sich die Ergebnisse, in Kombination mit Beam-Search bleiben sie gleich oder verbessern sich leicht gegenüber vergleichbaren Modellen. Für die dritte Aufgabe wird der Parser noch einmal erweitert: wir entwickeln das Transitionssystem so weiter, dass es neben **syntaktischer Struktur auch Satzgrenzen vorhersagt** (vgl. Abb. 5) und testen das System auf verrauschten und unredigierten Textdaten. Mit Hilfe sorgfältig ausgewählter Baselinemodelle und Testdaten messen wir den Einfluss syntaktischer Information auf die Vorhersagequalität von Satzgrenzen und zeigen, dass sich in Abwesenheit orthographischer Information wie Interpunktion und Groß- und Kleinschreibung das Ergebnis durch syntaktische Information verbessert.

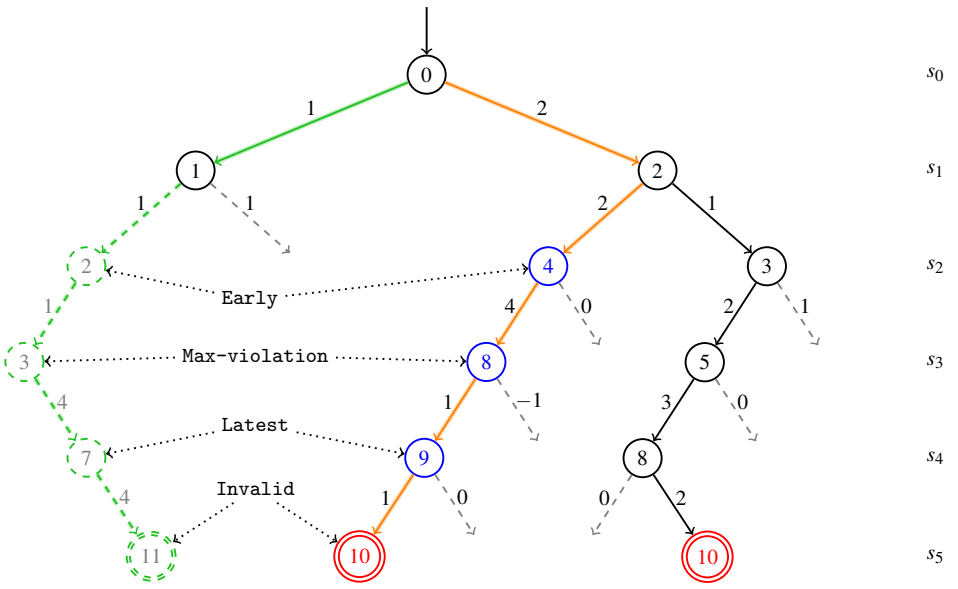



Abb. 3: *Beispiel-Illustration unterschiedlicher Update-Verfahren anhand eines Suchbaums für* Beam Search *mit* Beam*-Weite 2 (die Zustände innerhalb des* Beam*s sind mit durchgezogenen Linien dargestellt). Der korrekte Pfad ist grün hervorgehoben. Die blau hervorgehobenen Zustände sind jeweils die Basis für einen möglichen validen Update (gepaart mit den korrespondierenden grünen Zuständen außerhalb des* Beam*s). Rot hervorgehoben sind Zustände, anhand derer kein valider Update mehr möglich wäre (bei Schritt $s_5$ ist der Score des (verlorenen) korrekten Pfads größer als des besten* Beam*-Pfads.*

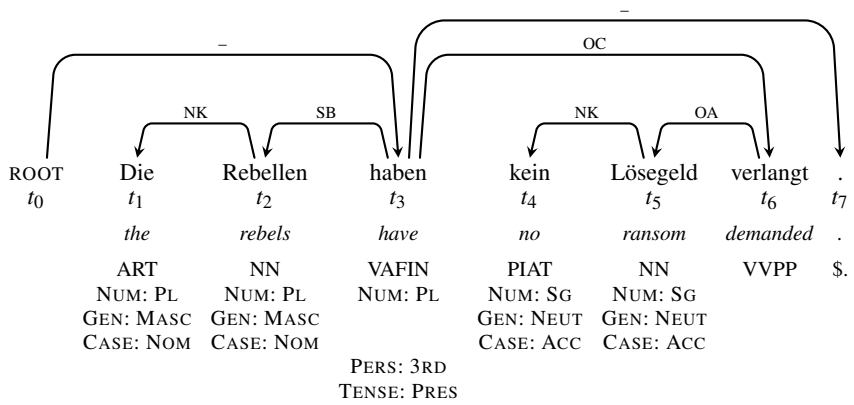
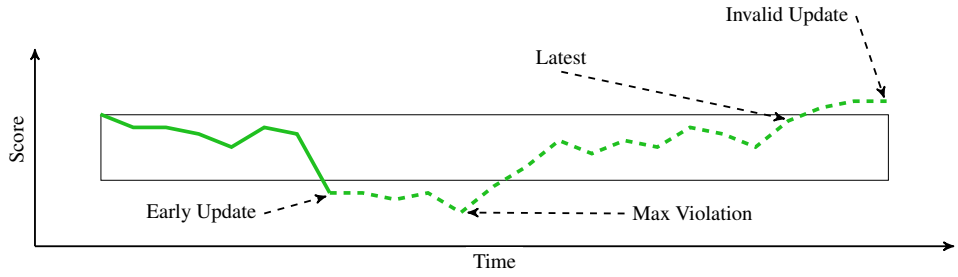



Abb. 4: *Dependenzstrukturbaum zur Illustration der Zielstruktur für die Aufgabe des Dependenzparsings. Die Baumstruktur oben entspricht der gesuchten Ausgabestruktur* **y***, die Eingabe* **x** *besteht in der Tokenfolge unten, einschl. atomaren Features auf den Tokens wie etwa deren Part-of-Speech-Tags.*

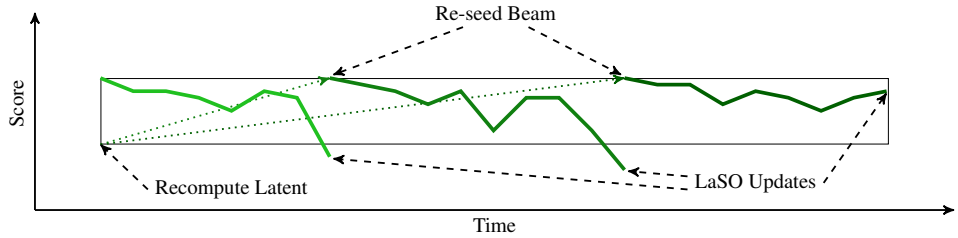
[1]cum ergo natus esset Iesus in Bethleem Iudaeae in diebus Herodis regis ecce magi ab oriente venerunt Hierosolymam [2]dicentes ubi est qui natus est rex Iudaeorum vidimus enim stellam eius in oriente et venimus adorare eum [3]audiens autem Herodes rex turbatus est et omnis Hierosolyma cum illo [4]et congregans omnes principes sacerdotum et scribas populi sciscitabatur ab eis ubi Christus nasceretur

"[1]Now when Jesus was born in Bethlehem of Judaea in the days of Herod the king, behold, there came wise men from the east to Jerusalem, [2]Saying, Where is he that is born King of the Jews? for we have seen his star in the east, and are come to worship him. [3]When Herod the king had heard these things, he was troubled, and all Jerusalem with him. [4]And when he had gathered all the chief priests and scribes of the people together, he demanded of them where Christ should be born."

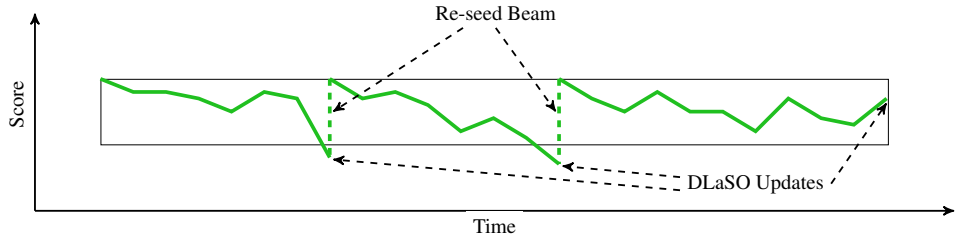
Abb. 5: *Beispiel zur Illustration der Herausforderungen bei der NLP-Aufgabe der gemeinsamen Satzgrenzenbestimmung und Satzstrukturanalyse: oben findet sich der Anfang von Matthäus 2 aus der lateinischen Vulgate-Bibel, unten die entsprechende englische Übersetzung in der King-James-Fassung. Vers-Zahlen sind als Superskripte dargestellt; satzinitiale Tokens sind unterstrichen dargestellt.*




(a) Early, Max-violation, Latest.

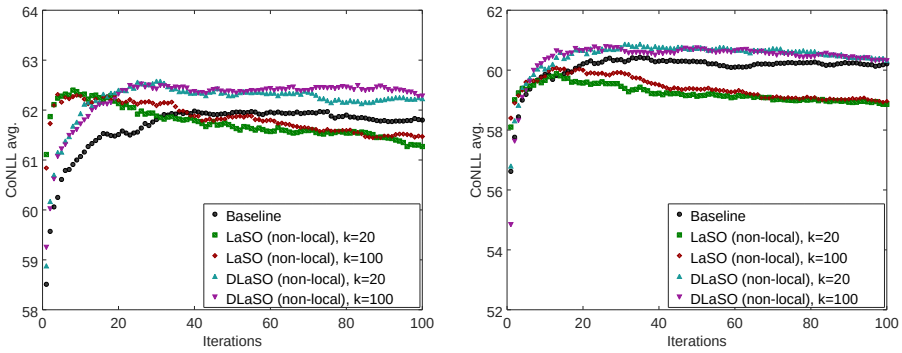



(b) LaSO.




(c) DLaSO.

Abb. 6: *Graphische Darstellung des Effekts unterschiedlicher Update-Strategien.*




(a) English (b) Chinese

Abb. 7: *Koreferenz-Resolution: Lernkurven für den Vergleich einer Baseline (exakte Suche mit ausdrucksschwächeren Features) und **LaSO** ("Learning as Search Optimization") [DIM05] und **DLaSO** ("Delayed Learning as Search Optimization"); (a) Kurve für Englisch, (b) Kurve für Chinesisch (Evaluation nach kombiniertem Maß aus CoNLL-Shared-Task zu Koreferenz-Resolution)*

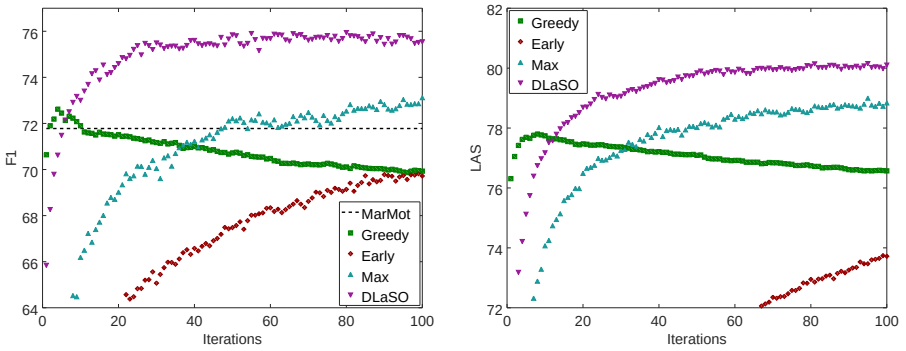
## 3 Verbesserte Update-Methoden bei der approximativen Suche

Das abstrakte Framework für die Modellierung von Strukturvorhersageaufgaben als Suchproblem mit einer latenten Vorhersagerepräsentation ermöglicht systematische Vergleichsexperimente zur Effektivität von unterschiedlichen *Machine Learning*-Strategien.

In der Dissertation wird das Framework für jede der drei erwähnten NLP-Analyseaufgaben instantiiert und führt einerseits zu aufgabenspezifischen empirischen Ergebnissen, andererseits zu systematischen Einsichten aus dem Vergleich. So erweist sich z.B. in mehreren Experimenten, dass die etablierten Update-Methoden, also Early- oder Max-Violation-Update, nicht mehr gut funktionieren, sobald die vorhergesagte Struktur eine gewisse Größe überschreitet.

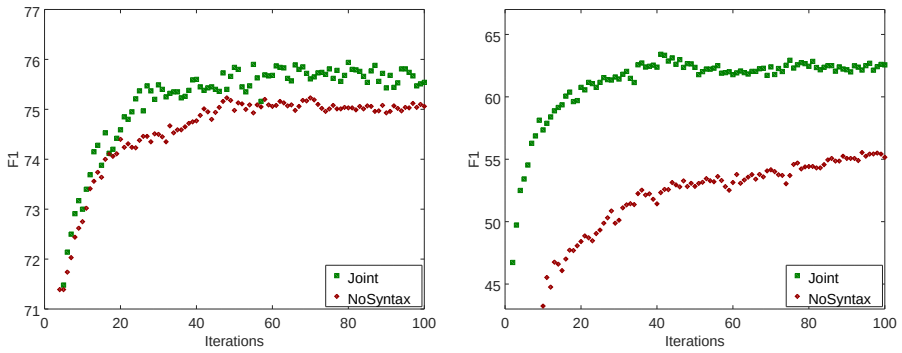
Es zeigt sich, dass das Hauptproblem dieser Methoden das **Auslassen von Trainingsdaten** ist, und dass sie desto mehr Daten auslassen, je größer die vorhergesagte Struktur wird. Dieses Problem kann durch bessere Update-Methoden vermieden werden, bei denen stets alle Trainingsdaten verwendet werden. Wir stellen eine neue Methode vor, **DLaSO** ("Delayed Learning as Search Optimization") [BK14], und zeigen, dass diese Methode konsequent bessere Ergebnisse liefert als alle Vergleichsmethoden. In der schematischen Darstellung in Abb. 6 wird deutlich, wie das Zurücksetzen der *Beam*-Kandidaten dazu führt, dass das Auslassen von Trainingsdaten vermieden wird; Abb. 7-9 zeigen Ausschnitte aus den experimentellen Ergebnissen.

Überdies zeigen die Experimente, dass eine erhöhte Beamgröße beim Suchen das Problem der ausgelassenen Trainingsdaten nicht kompensieren kann und daher keine Alternative zu besseren Update-Methoden darstellt.




(a) sentence boundary detection (b) parsing

Abb. 8: *Satzgrenzen- und Dependenzstruktur-Vorhersage mit approximativen Suchverfahren: Vergleich unterschiedlicher Update-Strategien auf den Switchboard-Korpus-Entwicklungsdaten. (a) zeigt den $F_1$ für Satzgrenzenerkennung; (b) Labeled Attachment Score für die Dependenzparsing-Aufgabe. DLaSO führt für beide Teilaufgabe zu einer erheblich höheren Vorhersage-Qualität.*




(a) Switchboard (b) WSJ*

Abb. 9: *Ergebnisse zur gemeinsamen Satzgrenzen- und Dependenzstruktur-Vorhersage: Effekt der Verfügbarkeit von syntaktischer Information für die Satzgrenzenvorhersage für zwei unterschiedliche Korpora: grün mit syntaktischer Information, rot ohne. (a) Switchboard; (b) WSJ* (Zeitungstext ohne Satzzeichen und ohne Großschreibung am Satzanfang).*

## Literaturverzeichnis

[Bj16] Björkelund, Anders; Faleńska, Agnieszka; Seeker, Wolfgang; Kuhn, Jonas: How to Train Dependency Parsers with Inexact Search for Joint Sentence Boundary Detection and Parsing of Entire Documents. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Berlin, Germany, S. 1924–1934, August 2016.

[BK14] Björkelund, Anders; Kuhn, Jonas: Learning Structured Perceptrons for Coreference Resolution with Latent Antecedents and Non-local Features. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Pa-

pers). Association for Computational Linguistics, Baltimore, Maryland, USA, S. 47–57, June 2014.

[BN15] Björkelund, Anders; Nivre, Joakim: Non-Deterministic Oracles for Unrestricted Non-Projective Transition-Based Dependency Parsing. In: Proceedings of the 14th International Conference on Parsing Technologies. Association for Computational Linguistics, Bilbao, Spain, S. 76–86, July 2015.

[Co02] Collins, Michael: Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, S. 1–8, July 2002.

[CR04] Collins, Michael; Roark, Brian: Incremental Parsing with the Perceptron Algorithm. In: Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume. Barcelona, Spain, S. 111–118, July 2004.

[DIM05] Daumé III, Hal; Marcu, Daniel: Learning as search optimization: approximate large margin methods for structured prediction. In: ICML. S. 169–176, 2005.

[FdSM12] Fernandes, Eraldo; dos Santos, Cícero; Milidiú, Ruy: Latent Structure Perceptron with Feature Induction for Unrestricted Coreference Resolution. In: Joint Conference on EMNLP and CoNLL - Shared Task. Association for Computational Linguistics, Jeju Island, Korea, S. 41–48, July 2012.

[HFG12] Huang, Liang; Fayong, Suphan; Guo, Yang: Structured Perceptron with Inexact Search. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Montréal, Quebec, Canada, S. 142–151, June 2012.

[Ni08] Nivre, Joakim: Algorithms for Deterministic Incremental Dependency Parsing. Computational Linguistics, 34(4):513–553, 2008.

[Ni09] Nivre, Joakim: Non-Projective Dependency Parsing in Expected Linear Time. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Association for Computational Linguistics, Suntec, Singapore, S. 351–359, August 2009.

[Ro58] Rosenblatt, Frank: The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. Psychological Review, 65(6):386–408, 1958.

[Sm11] Smith, Noah A.: Linguistic Structure Prediction. Synthesis Lectures on Human Language Technologies. Morgan and Claypool, May 2011.

Anders Björkelund, geboren 1985, hat in Lund, Schweden Informatik studiert und einen Master-Abschluss erworben. In Stuttgart promovierte er im Rahmen des Sonderforschungsbereichs 732 „Inkrementelle Spezifikation im Kontext“ am Institut für Maschinelle Sprachverarbeitung mit der Dissertation Online Learning of Latent Linguistic Structure with Approximate Search. Seit 2019 ist er zurück in Schweden und arbeitet am Institut für Astrophysik und Theoretische Physik der Universität Lund.

# Eine Experimentierumgebung zur Studie der physikalischen Schicht einer WLAN-Variante für die Anwendung in Fahrzeugnetzen[1]

Bastian Bloessl[2]

**Abstract:** Zukünftig werden Automobile mit Kommunikationsmodulen ausgestattet, die einen direkten Datenaustausch zwischen Fahrzeugen ermöglichen. Auf diese Weise können sich Verkehrsteilnehmer koordinieren, um so den Straßenverkehr sicherer, effizienter und komfortabler zu gestalten. Eine der Technologien, die dafür in Betracht gezogen wird, ist IEEE 802.11p, eine an Fahrzeugnetze angepasste Version von Wireless LAN (WLAN). Um die Eignung des Standards in diesem von der normalen WLAN-Nutzung sehr unterschiedlichen Umfeld zu untersuchen, haben wir prototypisch einen Software Defined Radio (SDR)-basierten IEEE 802.11p Transceiver implementiert. SDRs, programmierbare Funksende- und -empfangseinheiten, erlauben vollen Zugriff auf alle Aspekte drahtloser Kommunikation, bis hin zur elektromagnetischen Wellenform, und sind damit prädestiniert, die physikalische Schicht zu untersuchen. Eine Besonderheit unserer Implementierung besteht darin, dass wir durch abbilden des Standards in Software, dieselbe Implementierung für Simulationen und Messungen nutzen können, was eine detaillierte und umfassende Untersuchung erlaubt. Um die Vorteile und den flexiblen Einsatz unseres Transceivers herauszustellen, gehen wir auf zwei Studien ein, die ohne eine SDR-Implementierung nicht möglich gewesen wären. Zum einen untersuchen wir den Einfluss von Interferenz auf IEEE 802.11p und validieren so ein in vielen Studien genutztes Simulationsmodell. Zum anderen stellen wir einen neuen Angriff auf die Privatsphäre in Fahrzeugnetzen vor, der erst durch die Möglichkeit auf alle Daten des Empfangsprozesses zuzugreifen realisiert werden kann.

## 1 Einführung

Schon heute erleben wir, wie autonome Fahrzeuge dabei sind den Verkehr, und damit große Teile unserer Gesellschaft, zu revolutionieren. Viele der grundlegenden Fragestellungen sind gelöst und die ersten Prototypen fahren auf Deutschlands Straßen. Zukünftig werden wir noch einen Schritt weiter gehen und Fahrzeuge mit Funkmodulen ausstatten, um durch Kommunikation untereinander und optional auch mit fest installierten Infrastrukturknoten ein Fahrzeugnetz aufzubauen. Damit ermöglichen wir die Weiterentwicklung vom autonomen zum kooperativem Fahren. Wenn wir uns ins Gedächtnis rufen wie revolutionär die Möglichkeit der Vernetzung bei Computern, und später dem Smartphone, war, können wir abschätzen, was diese Entwicklung für den Verkehr bedeuten kann. So schaffen Fahrzeugnetze die Grundlage für eine Vielzahl von Anwendungen, die häufig unter Cooperative

[1] Eine Kurzfassung der Dissertation: B. Bloessl: A Physical Layer Experimentation Framework for Automotive WLAN, PhD Thesis (Dissertation), Paderborn, Germany: Department of Computer Science, Juni 2018.

[2] Connect Center, Trinity College Dublin, Ireland, E-Mail: mail@bastibl.net.

Intelligent Transportation Systems (C-ITS) zusammengefasst werden. Heute beschäftigen sich Forscher zum Beispiel mit Systemen zur Warnungen vor Querverkehr, verteilen Verkehrsinformationssystemen und automatisiertem Fahren in geringem Abstand [SD14]. Durch Anwendungen wie diesen werden Fahrzeugnetze den Verkehr in Zukunft sicherer, effizienter und komfortabler gestalten. Allein anhand der Vielfalt der möglichen Applikationen gibt es keine Technologie, die allen Anforderungen gerecht werden kann. Fahrzeugnetze werden deshalb mit großer Wahrscheinlichkeit heterogen aufgebaut sein und einen Mix aus Technologien, wie etwa LTE, WLAN, Millimeter Wave (mmW) und Visible Light Communication (VLC), verwenden.

## 1.1 Grundlagen

Eine zentrale Rolle könnte dabei IEEE 802.11p spielen. IEEE 802.11p ist eine an die Anforderungen in Fahrzeugnetzen angepasste Version von WLAN, die auf einem dedizierten Frequenzband um 5.9 GHz operiert. Mit diesem Standard können Fahrzeuge auch direkt miteinander kommunizieren und so ein dezentrales Netz, ein Vehicular Ad Hoc Network (VANET), bilden. Der IEEE 802.11p-Standard ist vor Allem relevant, weil er Vorteile bietet, die ihn besonders für Anwendungen aus dem Sicherheits- und Effizienzbereich geeignet erscheinen lassen:

**Geringe Kosten:** Es nutzt günstige, in großen Mengen produzierte Chips, die auf einem dedizierten, frei zugänglichen Frequenzband operieren.

**Niedrige Latenz:** Es unterstützt direkte Kommunikation zwischen Fahrzeugen und ist nicht auf eine Basisstation oder einen anderen Netzzugangsknoten angewiesen.

**Hohe Reichweite:** Es ermöglicht Reichweiten von über 800 m [GAS06] und ist nicht auf eine direkte Sichtverbindung angewiesen oder wird von Regen geblockt.

**Keine Direktionalität:** Im Gegensatz zu VLC und mmW hat WLAN keine inhärente Direktionalität und ist deswegen gut für broadcast-basierte Kommunikation zu allen Fahrzeugen in der Umgebung geeignet.

Angesichts dieser Eigenschaften und der Tatsache, dass Sicherheits- und Effizienzanwendungen als eine der wichtigsten Argumente für die Einführung von C-ITS angeführt werden, ist es wahrscheinlich, dass die Technologie eine zentrale Rolle in zukünftigen Fahrzeugnetzen spielen wird. Ihr derzeit größter Konkurrent ist Cellular Vehicle-to-Everything (C-V2X), eine im GPP-Umfeld entwickelter Kommunikationsstandard, der in LTE Release 14 aufgenommen wurde [WMG17]. C-V2X bietet ähnliche Vorteile und entwickelt sich rasch weiter, ist aber im Vergleich zu IEEE 802.11p weniger gut erforscht. Welche der beiden Technologien sich durchsetzt wird die Zukunft zeigen.

Im Gegensatz zu C-V2X wurde die physikalische Schicht von IEEE 802.11p initial nicht für die Anwendung in Fahrzeugnetzen konzipiert. Verglichen mit einem IEEE 802.11a/g-Signal wurde lediglich die Bandbreite von 20 MHz auf 10 MHz reduziert. Diese Reduzierung der Bandbreite führt zu einer Dehnung im Zeitbereich, was das Signal robuster in Kanälen mit starker Mehrwegeausbreitung macht. Die Frage, die sich hier anschließt, ist jedoch, ob diese Änderung ausreicht, um zuverlässige Kommunikation in den im Vergleich zu normalem WLAN viel dynamischeren Fahrzeugnetzen zu ermöglichen. In der Literatur wurde das Thema häufig aufgegriffen [AHG07; Fe12; Me11; NBS14]. Und auch heute ist die Eignung der Technologie und die Implementierung von leistungsfähigen Empfängern das Thema vieler Studien.

Ein Problem in diesem Kontext ist die Methodik. Simulationen der physikalischen Schicht basieren beispielsweise häufig auf vielen Annahmen und können deswegen unter Umständen nicht vollkommen überzeugen. Ein Versuch mit echter Hardware hingegen ist aufwendig und unflexibel. Zudem sind die in den Chips verwendeten Algorithmen meist nicht bekannt und können nicht verändert werden. Die Eignung dieser Prototypen als Experimentier- und Forschungsplattform ist deswegen stark eingeschränkt.

### 1.2 Wissenschaftlicher Beitrag

Um die Nachteile bestehenden Tools zu überwinden, haben wir einen SDR-basierten Prototypen des IEEE 802.11p-Standards implementiert. SDRs sind programmierbare Funksende- und Empfangseinheiten, die beliebige elektromagnetische Wellenformen senden und empfangen können. Sie sind damit das perfekte Werkzeug, um neue Technologien zu entwickeln, prototypisch Umzusetzen und experimentell erproben zu können. Unsere Transceiver nutzt GNU Radio, eine Echtzeit-Signalverarbeitungs-Umgebung die es erlaubt Kommunikationsstandards in Software auf einem normalen PC zu implementieren. So können Kommunikationssyteme schnell und mit relativ wenig Aufwand in Hochsprachen wie C++ und Python programmiert werden [Sk16]. Durch Implementierung des Standards in Software sind wir nicht an eine Hardware oder ein Betriebssystem gebunden. Unser Transceiver funktioniert mit allen SDRs, die von GNU Radio unterstützt werden, und kann zum Beispiel auch auf ARM-Plattformen genutzt werden. Er ist außerdem modular aufgebaut, das heißt Algorithmen können auf einfache Weise ausgetauscht und so verschiedene Empfängerarchitekturen miteinander verglichen werden. Die Implementierung wurde sowohl simulativ als auch mit kommerziellen WLAN-Karten und IEEE 802.11p-Prototypen validiert. Um zu zeigen, dass die Komplexität nicht die Rechenleistung eines PCs übersteigt, haben wir darüber hinaus Tests mit einem voll belegten Kanal durchgeführt. Hier wurde keine Überlast festgestellt.

Der größte Nachteil des Implementierens der physikalische Schicht auf dem PC ist die Latenz. Zum einen müssen die Daten zwischen dem SDR und dem PC übertragen werden, zum anderen werden sie auf einem Betriebssystem verarbeitet, das nicht auf Echtzeitanwendungen optimiert ist. Zeitkritische Funktionen sind deswegen nicht ohne Weiteres umsetzbar. Im

Rahmen der Arbeit haben wir gezeigt, dass durch Auslagern von Funktionalität auf den FPGA des SDR, Kanalzugriff und automatisiertes Anpassen der Empfangsverstärkung realisiert werden können ohne die Vorteile einer PC-Implementierung aufzugeben.

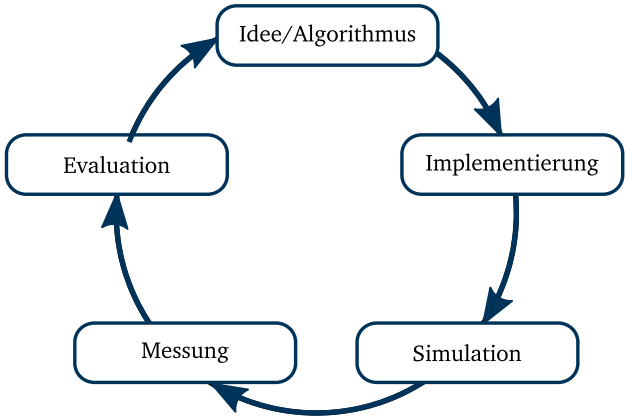



Abb. 1: Unser SDR-basierter Transceiver kann sowohl in Simulationen als auch in Messungen genutzt werden und ermöglicht einen einzigartigen Forschungsprozess, der so mit anderen Werkzeugen nicht möglich ist.

Der größte Vorteil unserer Implementierung ist, dass sie den in Abb. 1 skizzierten Forschungsprozess ermöglicht. Er beginnt mit einer Idee, wie etwa einer Variation des Standards oder eines neuen Signalverarbeitungs-Algorithmus. Diese Idee wird basierend auf unserer Implementierung realisiert und danach zunächst in Simulationen mit verschiedenen Kanalmodellen getestet. Anschließend wird ein und die selbe Implementierung mit SDRs für Messungen im Labor oder in einem Feldtest genutzt. Sollten die Ergebnisse nicht den Erwartungen entsprechen, kann das Design überdacht und die Forschungsfrage iterativ bearbeitet werden. Dieser Prozess ist so mit anderen Werkzeugen nicht möglich. Simulationen sind auf ihre Domäne beschränkt und können nicht für Messungen verwendet werden. Prototypen hingegen sind wenig flexibel und auf Messungen beschränkt. Unsere Implementierung erlaubt einen nahtlosen Übergang zwischen Simulation und Messung, und schafft es so die beiden Domäne zu verbinden. Damit ermöglichen wir ein tiefgreifenderes Verständnis und genauere Leistungsbewertung der WLAN-Technologie in Fahrzeugnetzen.

Insgesamt können die wissenschaftlichen Beiträge der Arbeit wie folgt zusammen gefasst werden:

- Wir entwickeln einen Open-Source SDR-basierten WLAN Transceiver, untersuchen seine Komplexität durch Laufzeittests und validieren ihn in Interoperabilitätstests mit anderen IEEE 802.11p prototypen.
- Wir zeigen, dass es möglich ist, auch zeitkritische Funktionen wie Kanalzugriff oder eine automatische Anpassung der Empfangsverstärkung zu verwirklichen, ohne die Vorteile einer PC-Implementierung aufgeben zu müssen.

Mit Hilfe des Transceivers bearbeiten wir offene Forschungsfragen:

- Wir führen zwei Feldtests durch, in denen wir zum einen die Leistung unserer Implementierung mit anderen Prototypen vergleichen und zum anderen die Tauglichkeit verschiedener Empfangsalgorithmen zu analysieren.
- Wir charakterisieren die Einfluss von Noise und Interferenz auf IEEE 802.11p und validieren so ein häufig verwendetes Simulationsmodell.
- Wir zeigen wie Informationen, die nur mit einer SDR-Implementierung zugänglich sind, genutzt werden können, um ein Bewegungsprofil von Fahrzeugen zu erstellen und quantifizieren die Auswirkungen dieses neuartigen Angriffs auf die Privatsphäre durch Simulationen.

Im Folgenden gehen wir kurz auf zwei Arbeiten ein, die durch diesen SDR-basierten Transceiver ermöglicht wurden.

## 2 Einfluss von Noise und Interferenz auf IEEE 802.11p

Für makroskopische Studien von Fahrzeugnetzen werden häufig Netzwerksimulatoren eingesetzt. Diese Simulatoren sind gut geeignet, um VANET-Anwendungen zu entwickeln und zu testen, da sie es erlauben auch große Szenarien einfach und reproduzierbar zu betrachten. Wie realistisch und aussagekräftig die Ergebnisse solcher Simulationen sind, hängt maßgeblich von der Qualität der verwendeten Simulationsmodelle ab. Besonders wichtig ist hier das Modell der physikalischen Schicht, das entscheidet, ob eine Übertragung bei gegebenem Signal-, Interferenz- und Noiselevel erfolgreich war. Populäre Simulatoren, wie ns-3 und Veins, nutzen hierfür Fehlerkurven basierend auf dem NIST-Model, das Framelänge, Kodierung und Signal to Interference and Noise Ratio (SINR) auf eine Fehlerwahrscheinlichkeit abbildet. Das Modell ist analytisch abgeleitet und empirisch mit kommerzieller Hardware verifiziert.

Die Krux des Modells ist die Verwendung der SINR. In ihm steckt die implizite Annahme, dass sich Noise und Interferenz in gleicher Weise auf die Fehlerwahrscheinlichkeit auswirken. Im Hinblick darauf, dass das Modell zunächst für das 2.4 GHz-Band mit vielen verschiedenen Interferenzquellen genutzt wurde, mag diese Annahme sinnvoll erscheinen. Da IEEE 802.11p auf einem dedizierten Band arbeitet, treten hier jedoch nur Interferenzen mit anderen IEEE 802.11p-Übertragungen auf. Deswegen und auch aufgrund der Tatsache, dass unabhängige Studien nahelegen, dass sich der Einfluss von Noise und Interferenz unterscheiden [FR15], wollen wir diese Annahme genauer untersuchen.

Mit Hilfe unseres SDR-Transceivers haben wir Simulationen aufgesetzt, in denen ein 546 Byte, QPSK-½-kodierter Frame einmal durch weißes Rauschen und ein anderes mal von einem interferierenden IEEE 802.11p-Frame gestört wurde. Die Störung setzte in

beiden Fällen etwas verzögert ein (nach 122 µs), so bleibt die Synchronisierungssequenz intakt und wir isolieren den Effekt auf die Datensymbole. Um Fehlerkurven zu erzeugen, wurde die relative Energie der beiden Übertragungen variiert und so unterschiedliche SINRs konfiguriert. Zu unserer Überraschung konnten wir zwischen beiden Szenarien keine wesentlichen Unterschiede feststellen. Um dieses Ergebnis zu verifizieren, haben wir dasselbe Szenario zusätzlich mit realer Hardware getestet. Wir senden mit dem SDR das gemixte Signal aus Frame und Noise beziehungsweise Interferenz und nutzen eine kommerzielle WLAN-Karte als Empfänger.

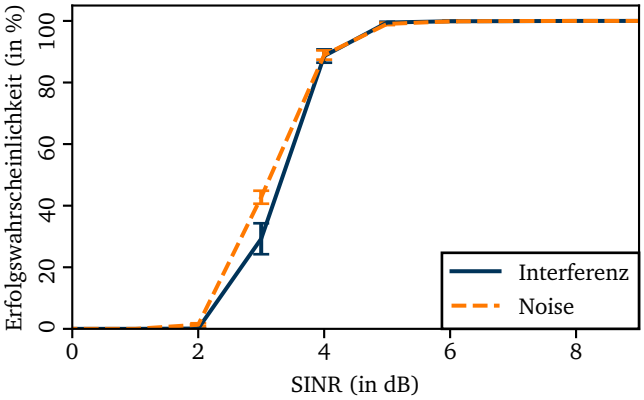



Abb. 2: Erfolgswahrscheinlichkeit für die Übertragung eines Frames, der durch einen anderen Frame, bzw. durch entsprechende Noise, gestört wird. (Reproduziert von [Bl17], © 2017 IEEE.)

Die Ergebnisse dieses Versuchs sind in Abb. 2 zu sehen. Die Fehlerbalken in dieser und der folgenden Abbildung entsprechend Konfidenzintervallen mit einem Signifikanzniveau von 95 %. Auch hier stimmen beide Szenarios sehr gut überein und unterstützen damit die häufig verwendete Annahme, dass Noise und Interferenz einen gleichen, beziehungsweise sehr ähnlichen, Einfluss auf die Leistung der physikalischen Schicht haben. In der Arbeit finden sich weitere Experimente und eine detaillierte Diskussion der Ergebnisse.

Um die Flexibilität unseres SDR-Transceivers zu zeigen, gehen wir im Folgenden über die bloße Leistungsbewertung der physikalischen Schicht hinaus und stellen einen neuartigen Angriff auf die Privatsphäre vor, der durch vollen Zugriff auf den Dekodierprozess ermöglicht wird.

## 3 Auswirkungen deterministischer Scrambler-Seeds

Bei der Entwicklung von Kommunikationsprotokollen für VANETs, also beispielsweise ETSI ITS-G5 in Europa und IEEE 1609 Wave in den USA, wurde die Privatsphäre von Grund auf mitgedacht. Durch Vermeidung von permanenten Identifikationsmerkmalen wie MAC-Adressen soll verhindert werden, dass auf einfache Weise ein Bewegungprofil von Fahrzeugen erstellt werden kann. Um temporäre Identitäten zu erlauben, werden Pseudonyme

verwendet werden. Diese Pseudonyme sind einmalig und über eine Zertifikatsinfrastruktur zentral verwaltet. Ob dieses Verfahren ausreicht, um die Privatsphäre von Nutzern zu schützen, wird aktuell diskutiert. Unabhängig davon unterstreicht die Debatte aber das Ziel Massenüberwachung zu vermeiden oder zumindest zu erschweren.

Zur Identifikation von Geräten aufgrund spezifischer Charakteristika ihrer Aussendungen gibt es viele Arbeiten. Hier werden jedoch oft kleine Abweichungen der analogen Radiokomponenten genutzt. Verglichen damit ist unser Angriff wesentlicher robuster, da er eine Initialisierungssequenz der physikalischen Schicht nutzt, die von jedem Empfänger am Anfang der Übertragung dekodiert werden muss. Konkret nutzen wir Schwächen der Scrambler-Implementierung. In digitalen Kommunikationssystemen werden die Daten vor der Modulation häufig gescramblt, also mit Hilfe einer pseudo-zufälligen Bitfolge randomisiert. Unabhängig von eventuell vorhandenen Strukturen in den Nutzdaten, also etwa langen Folgen von Nullen oder Einsen, hat die so entstehende gescramblte Bitfolge eine unkorrelierte Gleichverteilung. Diese Randomisierung ist kein Sicherheitsmerkmal, sie wird ausschließlich wegen Vorteilen bei der Signalverarbeitung durchgeführt.

Beim WLAN wird diese Zufallsbitfolge mit einem rückgekoppelten Schieberegister erzeugt, das laut Standard für jeden Frame mit einem pseudozufälligen Seed initialisiert werden soll. Um dem Empfänger das Dekodieren zu ermöglichen, wird der Seed vor den Nutzdaten gesendet. Da der Scrambling-Prozess tief in der physikalischen Schicht verankert ist, kann er mit normalem Netzwerkmonitoring, zum Beispiel mit Wireshark, nicht nachvollzogen werden. Mit unserer SDR-Implementierung ist es hingegen leicht möglich, da der Seed sowieso im Zuge des Dekodier-Prozesses ermittelt werden muss. Um zu sehen wie, die vom Standard geforderte Zufällige Initialisierung Implementiert ist, haben wir mit Hilfe des SDR die Sequenzen für zwei populäre IEEE 802.11p-Prototypen geloggt.

Zum einen untersuchen wir eine, auf einem Atheros AR5413-Chip basierende, Unex DCMA-86P2-Karte, die in vielen Feldtests genutzt wurde. Zum anderen untersuchen wir mit dem Cohda Wireless MK2 einen kommerziellen IEEE 802.11p-Prototypen, der vor allem zum Testen von VANET-Anwendungen weite Verbreitung findet. Bei der Betrachtung der Scrambler-Seeds konnten wir feststellen, dass beide Karten einfache und, was noch schlimmer ist, deterministische Seeds verwenden.

Bei der Atheros-Karte werden die Seeds einfach inkrementiert, es werden also Frames mit Seed 1, 2, 3, ... versendet. Das MK2 hingegen reinitialisert den Scrambler nicht. Auch hier kann ein Angreifer, der einen Frame empfangen hat, den nächsten Seed vorhersagen, da er den Zustand des Schieberegisters kennt. Ist außerdem die Framelänge bekannt, so kann der Angreifer zukünftige Seeds vorhersagen. Bei Atheros-Karten ist das unabhängig von der Framelänge immer möglich. Dieses deterministische Verhalten ist im Hinblick auf die Leistung des Transceivers völlig unkritisch. Mit Blick auf die Privatsphäre ergibt sich jedoch ein anderes Bild, da die Schwächen der Implementierungen eine einfache Zuordnung von Frames zu einem Sender erlauben. Werden etwa Frames mit Scrambler-Seeds von 10, 11, 12, 13, ... empfangen, so ist klar, dass diese mit großer Wahrscheinlichkeit von einem

Fahrzeug mit einer Atheros-Karte stammen. Diese Möglichkeit, Frames unabhängigen von MAC-Adressen oder Pseudonymen einem Empfänger zuzuordnen, hebelt den Schutz der Privatsphäre zum großen Teil aus.

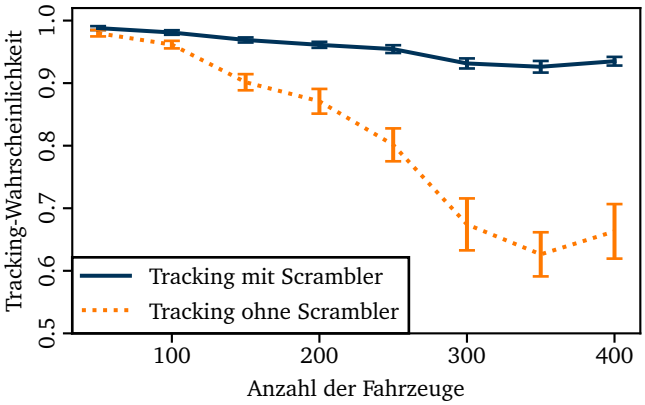



Abb. 3: Einfluss determinstischer Scrambler-Seeds auf die Tracking-Wahrscheinlichkeit. (Reproduziert von [Bl15], © 2015 IEEE.)

Um die Auswirkungen dieser Schwächen zu quantifizieren, haben wir verschiedene Szenarien mit Veins, einem Netzwerksimulator für Fahzeugnetze untersucht. In Abb. 3 werden exemplarisch die Ergebnisse für ein Autobahnszenario gezeigt. Hier werden an einer dreispurigen Autobahn statische Knoten platziert, zwischen denen ein 800 m langer blinder Fleck ist, an dem keiner der Knoten Daten empfangen kann. Die Fragestellung ist hier, ob die Knoten Fahrzeuge zuordnen und damit tracken können. Wir nehmen an, die Fahrzeuge nutzen jeweils zu ⅓, wirklich zufällige, inkrementelle und uninitialisierte Scrambler-Seeds. Die Abbildung vergleicht die Erfolgswahrscheinlichkeit abhängig vom Verkehrsaufkommen. Für die Zuordnung von Fahrzeugen nutzen wir einen dem aktuellen Stand der Forschung entsprechenden Tracking-Algorithmus und vergleichen die Standardversion (orange) mit einer erweiterten Version, die zusätzlich deterministische Seeds ausnutzt (blau). (Weitere Details zum Simulationssetup und finden sich in der Arbeit.) Wie die Abbildung zeigt, können vor allem bei höheren Verkehrsaufkommen die Schwächen in der Scrambler-Implementierung die Tracking-Wahrscheinlichkeit signifikanten erhöhen. Wenn sich durchschnittlich 350 Fahrzeuge auf dem simulierten Autobahnabschnitt befinden, steigt die Tracking-Wahrscheinlichkeit beispielsweise von ca. 63 % auf ca. 95 %.

Inzwischen wurden unsere Untersuchungen von einer unabhängigen Gruppe auch auf andere kommerzielle WLAN-Karten ausgeweitet [Va16]. Auch bei diesen Karten wurden ähnliche Schwachstellen gefunden. Unserer Meinung nach ist es wichtig möglichst frühzeitig auf diese Problematik hinzuweisen, da die physikalische Schicht oft auf dem Chip implementiert ist und die Schwächen unter Umständen nicht durch spätere Firmware-Updates behoben werden können.

## 4 Schlussfolgerung

Schon in naher Zukunft werden Autos mit Funkmodulen ausgestattet, die eine direkte Kommunikation untereinander ermöglichen und damit die Grundvorraussetzung zum kooperativen Fahren schaffen. Um die Leistungsfähigkeit dieser Technologie besser zu verstehen und ihre Einsatzfähigkeit zu untersuchen, sind Experimente mit Prototypen von entscheidender Bedeutung. In dieser Arbeit haben wir einen SDR-basierten Transceiver entwickelt, der sowohl simulative als auch experimentelle Leistungsbewertung erlaubt und so einen einzigartigen Forschungsprozess ermöglicht. Wir denken, dass diese Implementierung einen wichtigen Beitrag leistet um die Eignung von WLAN-Technologie für Fahrzeugnetze zu evaluieren. Unsere Implementierung wurde mit Hilfe von Simulationsmodellen und Interoperabilitätstests mit IEEE 802.11p-Prototypen validiert. Darüber hinaus wurde gezeigt, dass zeitkritische Funktionen auf den FGPA des SDR ausgelagert werden können, ohne die Vorteile einer PC-Implementierung aufzugeben. Die Flexibilität des Transceivers wurde in Feldtests und zwei weiterführenden Studien gezeigt: Einerseits haben wir den Einfluss von Noise und Interferenz auf IEEE 802.11p untersucht und so ein häufig verwendetes Simulationsmodell validiert. Andererseits haben wir einen neuen Angriff auf die Privatsphäre vorgestellt, der erst durch unsere SDR-Implementierung ermöglicht wurde.

Um unser Arbeit der wissenschaftlichen Gemeinde zur Verfügung zu stellen, wurde der Transceiver unter einer Open-Source-Lizenz veröffentlicht.[3] Inzwischen basieren 72 Veröffentlichungen auf dieser Arbeit. 48 davon wurden von unabhängigen Gruppen publiziert.

## Literaturverzeichnis

[AHG07] Alexander, P.; Haley, D.; Grant, A.: Outdoor Mobile Broadband Access with 802.11. IEEE Communications Magazine 45/11, Nov. 2007.

[Bl15] Bloessl, B.; Sommer, C.; Dressler, F.; Eckhoff, D.: The Scrambler Attack: A Robust Physical Layer Attack on Location Privacy in Vehicular Networks. In: 4th IEEE International Conference on Computing, Networking and Communications (ICNC 2015), CNC Workshop. IEEE, Anaheim, CA, Feb. 2015.

[Bl17] Bloessl, B.; Klingler, F.; Missbrenner, F.; Sommer, C.: A Systematic Study on the Impact of Noise and OFDM Interference on IEEE 802.11p. In: 9th IEEE Vehicular Networking Conference (VNC 2017). IEEE, Torino, Italy, Nov. 2017.

[Bl18] Bloessl, B.: A Physical Layer Experimentation Framework for Automotive WLAN, PhD Thesis (Dissertation), Paderborn, Germany: Department of Computer Science, Juni 2018.

[3] `https://www.wime-project.net`

[Fe12] Fernandez, J. A.; Borries, K.; Cheng, L.; Vijaya Kumar, B. V. K.; Stancil, D. D.; Bai, F.: Performance of the 802.11p Physical Layer in Vehicle-to-Vehicle Environments. IEEE Transactions on Vehicular Technology 61/1, Jan. 2012.

[FR15] Fuxjaeger, P.; Ruehrup, S.: Validation of the NS-3 Interference Model for IEEE802.11 Networks. In: 8th IFIP Wireless and Mobile Networking Conference (WMNC 2015). IEEE, Munich, Germany, Okt. 2015.

[GAS06] Gallagher, B.; Akatsuka, H.; Suzuki, H.: Wireless Communications for Vehicle Safety: Radio Link Performance and Wireless Connectivity Methods. IEEE Vehicular Technology Magazine 1/4, Dez. 2006.

[Me11] Mecklenbräuker, C. F.; Molisch, A. F.; Karedal, J.; Tufvesson, F.; Paier, A.; Bernadó, L.; Zemen, T.; Klemp, O.; Czink, N.: Vehicular Channel Characterization and its Implications for Wireless System Design and Performance. Proceedings of the IEEE 99/7, Juli 2011.

[NBS14] Nagalapur, K. K.; Brännström, F.; Ström, E. G.: On Channel Estimation for 802.11p in Highly Time-Varying Vehicular Channels. In: IEEE International Conference on Communications (ICC 2014). IEEE, Sydney, Australia, Juni 2014.

[SD14] Sommer, C.; Dressler, F.: Vehicular Networking. Cambridge University Press, 2014.

[Sk16] Sklivanitis, G.; Gannon, A.; Batalama, S. N.; Pados, D. A.: Addressing Next-Generation Wireless Challenges with Commercial Software-Defined Radio Platforms. IEEE Communications Magazine 54/1, Jan. 2016.

[Va16] Vanhoef, M.; Matte, C.; Cunche, M.; Cardoso, L. S.; Piessens, F.: Why MAC Address Randomization is not Enough: An Analysis of Wi-Fi Network Discovery Mechanisms. In: 11th ACM Asia Conference on Computer and Communications Security (ASIACCS 2016). ACM, Xi'an, China, Mai 2016.

[WMG17] Wang, X.; Mao, S.; Gong, M. X.: An Overview of 3GPP Cellular Vehicle-to-Everything Standards. GetMobile: Mobile Computing and Communications 21/3, Sep. 2017.

**Bastian Bloessl** ist PostDoc am Trinity College Dublin in Irland, wo er durch ein Marie Skłodowska-Curie-Stipendium finanziert ist. Bastian hat den Diplomstudiengang Informatik an der Universität Würzburg 2011 abgeschlossen. Im selben Jahr begann er seine Promotion in der Gruppe von Prof. Falko Dressler an der Universität Innsbruck, die er ab 2014 in an der Universität Paderborn weiterführte. 2015 erhielt Bastian ein FitWeltweit-Stipendium des DAAD, das es ihm ermöglichte, sechs Monate als Gastwissenschaftler an der University of California, Los Angeles (UCLA) in der Gruppe von Prof. Mario Gerla zu arbeiten. Seit 2017 ist Bastian außerdem einer der Leiter des GNU Radio Projekts, einem Open-Source-Softwareprojekt zur Echtzeitsignalverarbeitung.

# 6D Posenschätzung mit gelernten, dichten Korrespondenzvorhersagen[1]

Eric Brachmann[2]

**Abstract:** Diese Arbeit befasst sich mit der der Schätzung von Position und Orientierung von Objekten oder Szenen aus einzelnen Kamerabildern (Posenschätzung). Es wird ein Verfahren vorgestellt, welches etablierte Lösungsstrategien der Computer Vision mit neuen Verfahren des maschinellen Lernens kombiniert. Zunächst wird das Korrespondenzproblem zwischen Eingabebild und Zielobjekt gelöst, dann wird die gesuchte Pose durch robuste, geometrische Optimierung ermittelt. Nur Teile des Verfahrens werden anhand von Trainingsdaten gelernt, was zu Generalisierung und Interpretierbarkeit des Systems führt. Etablierte Algorithmen, insbesondere der *RANSAC*-Algorithmus aus der robusten Optimierung, werden so erweitert, dass ein Trainieren des Gesamtsystems möglich ist. Mit dem DSAC-Algorithmus (*Differentiable RANSAC*) stellt diese Arbeit Forschern auf dem Gebiet des maschinellen Lernens ein neues, vielseitiges Werkzeug zur Verfügung.

## 1 Einleitung

In den letzten Jahrhunderten hat die Menschheit einen technischen Fortschritt erlebt, der in vielen Regionen der Welt zu einer Steigerung der Lebensqualität führte. Eine Grundlage für diesen Fortschritt wurde vor ca. 200 Jahren in der industriellen Revolution gelegt. Im Moment beobachten wir erneut einen rasanten Umbruch in der Menschheitsgeschichte, auch digitale Revolution genannt, ausgelöst durch die rapide Entwicklung von Computern. Berechnungen und Simulationen erfolgen in Sekundenbruchteilen, riesige Datenmengen können gespeichert und verarbeitet werden und das Internet verbindet Menschen und Geräte weltweit. Durch aktuelle Entwicklungen in der Forschung zur künstlichen Intelligenz (KI) können Maschinen immer komplexere Aufgaben immer selbstständiger lösen. Schon bald könnten Smart Homes, voll-automatische Fabriken und Warenhäuser, computergestützte Chirurgie oder autonomes Fahren Realität werden. Einige dieser Errungenschaften hätten das Potential, die Lebensqualität und Lebenserwartung weiter zu steigern, etwa durch Reduzierung von Verkehrsunfällen. Gleichzeitig wird unser bisheriges Verständnis von (menschlicher) Intelligenz und der soziale und wirtschaftliche Status von (menschlicher) Arbeit in Frage gestellt.

Soll eine Maschine eine komplexe Aufgabe autonom lösen, muss sie in den meisten Fällen ihre Umgebung wahrnehmen und interpretieren. Die Computer Vision ist ein Gebiet innerhalb der KI-Forschung, welches sich mit dem Verstehen von Bildern beschäftigt, d.h. mit dem Extrahieren von semantischen Informationen aus visuellen Daten. Für uns Menschen ist diese Fähigkeit so allgegenwärtig und selbstverständlich, dass uns die komplexen Prozesse der Bildentstehung, die visuelle Vielfalt und Mehrdeutigkeit unserer

[1] Englischer Titel der Dissertation: “Learning to Predict Dense Correspondences for 6D Pose Estimation”

[2] Universität Heidelberg, eric.brachmann@tu-dresden.de

alltäglichen Umgebung nicht bewusst sind. Das Aussehen von Gegenständen ändert sich enorm, je nach Blickwinkel, Beleuchtungssituation, Reflektionen oder teilweiser Verdeckung. Während man in den Anfangszeiten der Computer Vision versuchte, stabile Muster händisch zu definieren, durch die man trotz all dieser Faktoren Objekte in Bildern erkennen konnte, setzt man heute zunehmend auf das maschinelle Lernen. Aufgrund eines Trainingdatensatzes soll sich ein technisches System die komplexen Zusammenhänge zwischen Objektattributen und deren visueller Erscheinung selbständig erschließen. Neue Methoden des maschinellen Lernens, insbesondere des *Deep Learnings* mittels neuronalen Netzen mit Millionen lernbaren Parametern, haben die Leistungsfähigkeit von technischen Systemen für viele Aufgaben massiv verbessert. Etwa in der Erkennung von Gesichtern oder Verkehrsschildern sind diese Systems dem Menschen unter bestimmten Umständen inzwischen überlegen [St12].

Die Dissertation [Br18a] beschäftigt sich mit dem Schätzen der Position und der Orientierung von Objekten aus einzelnen Bildern. Aufgrund des komplexen Bildentstehungsprozesses ist die Lösung dieser Aufgabe für technische Systeme schwierig, gleichzeitig erfordern Anwendungen wie *Augmented Reality* eine hohe Stabilität und Präzision der Ergebnisse. Reines *Deep Learning*, d.h. die Lösung der gesamten Aufgabe durch ein neuronales Netz, liefert nur enttäuschende Ergebnisse. Die Dissertation stellt ein präzises, skalierbares und vielseitig einsetzbares Verfahren zur Posenschätzung vor, welches Methoden des maschinellen Lernens mit traditionellen Ansätzen der Computer Vision kombiniert.

## 1.1 Problemdefinition

Die Dissertation beschäftigt sich mit der Posenschätzung von Objektinstanzen. Im Unterschied zu einer Objektklasse bezeichnet eine Instanz ein bestimmtes physisches Objekt, einzigartig in Material und Form. Beispielsweise handelt es sich bei *Auto* um eine Objektklasse und bei *roter VW Golf VII* um eine Objektinstanz. Weiterhin sind Objektinstanzen in dieser Arbeit auf Starrkörperobjekte beschränkt, d.h. ihre Form ändert sich nicht. Im Unterschied dazu stehen artikulierte oder verformbare Objekte wie Laptops.

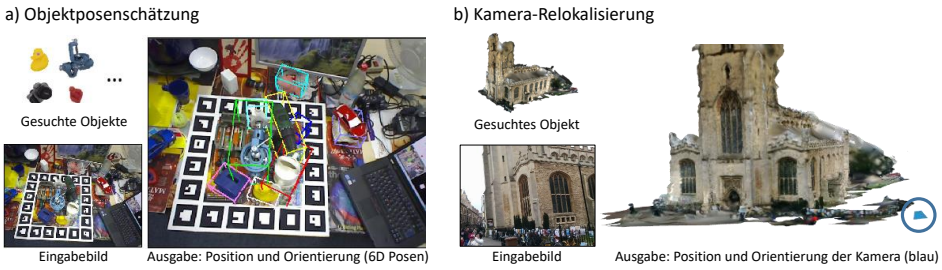



Abb. 1: **Posenschätzung. a)** Das System schätzt Position und Orientierung mehrerer sichtbarer Objekte, dargestellt als farbige 3D-Rahmen. **b)** Das System schätzt Position und Orientierung der Kamera, von der ein Bild einer bekannten Umgebung (eine Kirche in Cambridge, UK) aufgenommen wurde.

Gegeben ist ein Kamerabild $I$, das eine oder mehrere bekannte Objektinstanzen zeigt. Es handelt sich entweder um ein RGB-Bild einer üblichen Farbkamera, oder ein RGB-D-Bild einer speziellen Tiefenkamera, z.B. einer Kinect-Kamera, was die Posenschätzung erheblich vereinfacht. Für jedes Objekt, welches dem System bekannt ist, liefert es einen Sicherheitswert, ob das jeweilige Objekt zu sehen ist. Weiterhin bestimmt es für jedes Objekt die Pose $\mathbf{h}$ bestehend aus der Position $\mathbf{t}$ relativ zur Kamera und der Orientierung $\theta$ des Objekts, siehe Abb. 1 a). Position und Orientierung haben jeweils 3 Freiheitsgrade, d.h. bei der Pose $\mathbf{h}$ handelt es sich um einen 6D-Vektor.

Statt um einen Gegenstand kann es sich bei dem gesuchten Objekt auch um eine ganze Umgebung handeln, etwa eine bestimmte Wohnung oder ein Gebäude, siehe Abb. 1 b). In diesem Fall wird die Pose der Kamera relativ zum Objekt geschätzt, auch Kamera-Relokalisierung genannt. Objekt-Posenschätzung und Kamera-Relokalisierung sind methodisch äquivalent und werden in dieser Arbeit gleichermaßen behandelt.

## 1.2 Anwendungen

Die Anwendungsmöglichkeiten von Posenschätzung sind vielfältig. Kamera-Relokalisierung kann die Navigation von autonomen Fahrzeugen unterstützen, wenn GPS nicht zuverlässig funktioniert oder die Genauigkeit nicht ausreicht. Für Navigation in geschlossenen Räumen steht GPS außerdem oft nicht zur Verfügung. Soll ein Roboter mit Objekten interagieren, sie etwa in einem automatisierten Warenhaus greifen, muss deren Lage im Raum exakt bestimmt werden. *Augmented Reality*, also die Verschmelzung von realen und virtuellen Inhalten, erfordert die genaue Registrierung von Objekten und Umgebung mit der AR-Anzeige. Gleichermaßen wird in der Wahrnehmungspsychologie die Aufmerksamkeit von Probanden mittels tragbaren Eye-Trackern untersucht. Durch Kamera-Relokalisierung können in diesem Kontext Aufmerksamkeitskarten, etwa von sicherheitskritischen Arbeitsplätzen, erstellt werden. Im Folgenden wird eine Anwendung der Posenschätzung für computergestützte Chirurgie näher erläutert.

Die moderne Medizin ermöglicht besonders schonende Eingriffe durch laparoskopische, also minimal-invasive, Chirurgie, siehe Abb. 2 a). Dabei operiert der Chirurg mit speziellen Instrumenten durch die geschlossene Bauchdecke. Der Chirurg kann sich lediglich durch eine sehr eingeschränkte endoskopische Sicht orientieren, siehe Abb. 2 b), was diese Eingriffe sehr kompliziert macht. Durch das Tracken der 6D-Posen der Operationsinstrumente wäre es möglich, den Chirurgen zu unterstützen, etwa um Distanzen in der Bauchhöhle zu messen, Schnitttiefen zu bestimmen oder Operationsphasen zu erkennen. Wäre es darüber hinaus möglich, die Endoskop-Kamera innerhalb der Bauchhöhle zu lokalisieren, könnten dem Chirurgen Navigationshilfen angeboten werden, etwa über *Augmented Reality*, siehe Abb. 2 c).

Bei dem soeben erläuterten Anwendungsszenario von Posenschätzung in der laparoskopischen Chirurgie handelt es sich um eine Vision. Im Moment existiert kein Verfahren für das Intrumenten-Tracken oder die Kamera-Relokalisierung, das innerhalb des menschlichen Körpers verlässlich funktioniert. Mit dieser Arbeit werden bestehende Verfahren in

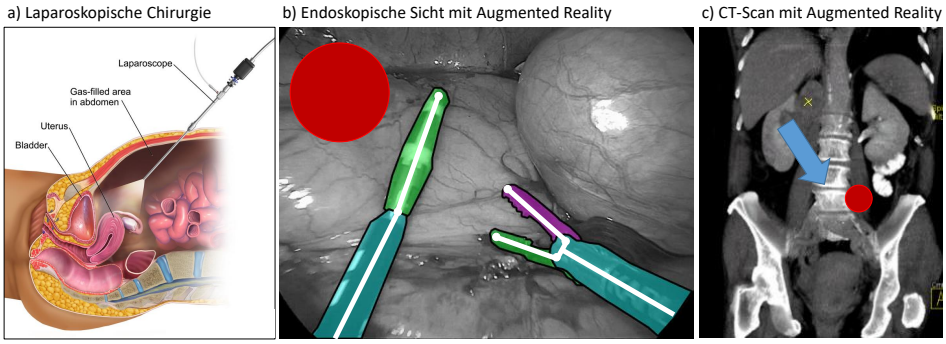
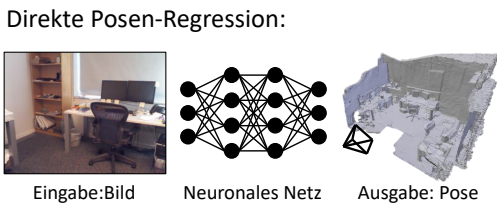



Abb. 2: **Computergestützte Chirurgie (Vision). a)** Schematische Darstellung einer minimalinvasiven Operation im Bauchraum. **b)** Sicht durch das Endoskop. Die Pose der Instrumente wird getrackt. Über *Augmented Reality* wird die Zielposition des Eingriffs als Navigationshilfe eingeblendet (rot). **c)** In einem präoperativen CT-Scan wird die geschätzte Endoskop-Position (blau) und die Zielposition des Eingriffs (rot) gezeigt.

vielen Aspekten verbessert, z.B. hinsichtlich der Robustheit gegenüber Objektverdeckung oder der Genauigkeit der Kamera-Relokalisierung. Andere wichtige Aspekte bleiben offen für zukünftige Forschung, etwa Relokalisierung in deformierbaren Umgebungen (etwa der Bauchhöhle), oder das Tracken von stark reflektiven Objekten (etwa Operationsinstrumenten).

## Forschungsbeitrag




Direkte Posen-Regression

| Methode (Jahr) | Genauigkeit |
|---|---|
| PoseNet v.1 [KGC15] (2015) | 45cm, 10° |
| Spatial LSTM [Wa17] (2017) | 31cm, 10° |
| PoseNet v.2 [KC17] (2017) | 23cm, 8° |
| PoseNet v.3 [Br18c] (2018) | 22cm, 8° |
| MapNet+ [Br18c] (2018) | 19cm, 7° |
| Traditioneller Ansatz [SLK12] (2012) | 5.1cm, 2.5° |
| Ansatz dieser Arbeit [ER18a] (2018) | 3.6cm, 1.1° |

Abb. 3: **Direkte Posen-Regression. Links:** Ein neuronales Netz erzeugt die gewünschte Ausgabe direkt. **Rechts:** Direkte Regression mit neuronalen Netzen erzielt in der Kamera-Relokalisierung keine guten Ergebnisse. Diese Arbeit kombiniert neuronale Netze mit traditionellen Ansätzen und ermöglicht damit eine sehr hohe Genauigkeit.

Im Jahr 2012 gewann ein *Convolutional Neural Network* (CNN) den bedeutenden ImageNet Wettbewerb für Bildklassifizierung mit einem großen Abstand zu allen konkurrierenden Methoden. Seitdem hat das sogenannte *Deep Learning*, also das maschinelle Lernen mittels neuronaler Netze einen beispiellosen Siegeszug in der Computer Vision und einigen anderen Wissenschaftszweigen angetreten. Für Aufgaben wie Bildklassifizierung,

2D-Objekt-Detektion und semantischer Segmentierung sind neuronale Netze im Moment unangefochten in ihrer Leistungsfähigkeit. Diese Methoden sind dabei oft sehr ähnlich aufgebaut. Das Eingabebild durchläuft die verschiedenen Schichten des neuronalen Netzes, wobei es schrittweise in die gewünschte Ausgabe transformiert wird, etwa in Wahrscheinlichkeiten für verschiedene Bildklassen. Das neuronale Netz erzeugt die gewünschte Ausgabe also direkt und unmittelbar aus der Eingabe, im folgenden *direkte Regression* genannt, siehe auch Abb. 3, links. Während dieses Vorgehen für viele Problemstellungen hervorragende Resultate erzielt, sind die entsprechenden Ergebnisse für 6D-Posenschätzung enttäuschend. Abb. 3, rechts führt die Ergebnisse einiger aktueller Ansätze von direkter Regression für das Problem der Kamera-Relokalisierung an. Die Genauigkeit stagniert seit 2017 bei ca. 20cm für die Lokalisierung innerhalb eines Zimmers und ist damit für Anwendungen wie *Augmented Reality* nicht brauchbar. Das ist insbesondere überraschend, da wesentlich ältere, traditionelle Ansätze (z.B. feature-basiertes Matching [SLK12]), welche keinerlei Form des maschinellen Lernens verwenden, diese Genauigkeit bei weitem übertreffen. Traditionellen Ansätze kombinieren von Menschen erdachte Bild-Features mit der Optimierung von geometrischen Bedingungen unter Kenntnis bestimmter Aspekte der Bildentstehung. Diese Verfahren haben den weiteren Vorteil, dass die geometrische Konsistenz der Vorhersagen geprüft werden kann. Damit sind die Vorhersagen zu gewissen Maße interpretierbar und mit einer Abschätzung ihrer Zuverlässigkeit verknüpft. Bei Ansätzen der direkten Regression handelt es sich dagegen zumeist um eine *Black Box*. Ihre Voraussagen resultieren aus einem nicht einsehbaren bzw. nicht interpretierbaren Prozess und werden ohne Sicherheitsabschätzung getroffen. Eine Voraussage kann also nur schwer bezüglich ihrer Vertrauenswürdigkeit abgeschätzt werden, was in kritischen Anwendungen wie dem autonomen Fahren ein essentielles Problem darstellt.

Ein wesentlicher Beitrag dieser Arbeit besteht darin, traditionelle Ansätze der Computer Vision mit den neuen Möglichkeiten des maschinellen Lernens zu kombinieren, ohne die traditionellen Ansätze aber vollständig zu ersetzen. Diese Arbeit zeigt am Beispiel der 6D-Posenschätzung, dass diese Kombination in einer erhöhten Genauigkeit resultiert und gleichzeitig wünschenswerte Eigenschaften der traditionellen Methoden erhält, beispielsweise eine teilweise Interpretierbarkeit und abschätzbare Vertrauenswürdigkeit der Vorhersagen. Im Folgenden werden die zentralen Forschungsbeiträge der Arbeit aufgeführt.

- Ein neues Verfahren zur 6D-Posenschätzung, das traditionelle Ansätze aus der robusten Optimierung und projektiven Geometrie mit Werkzeugen des maschinellen Lernens vereint.
- Das Verfahren zeichnet sich durch hohe Genauigkeit und Vielseitigkeit aus: Es unterstützt RGB- sowie RGB-D-Bilder als Eingabe und schätzt die Pose von texturierten oder nicht-texturierten Gegenständen sowie von ganzen Umgebungen.
- Ein zentraler Algorithmus der robusten Optimierung, welcher oft in traditionellen Verfahren der Computer Vision Verwendung findet, ist *Random Sample Consensus* (RANSAC). Diese Arbeit beschreibt eine differenzierbare Variante von RANSAC die in Kombination mit *Deep Learning* verwendet werden kann.

## 2 Lernen von Bild-Objekt-Korrespondenzen

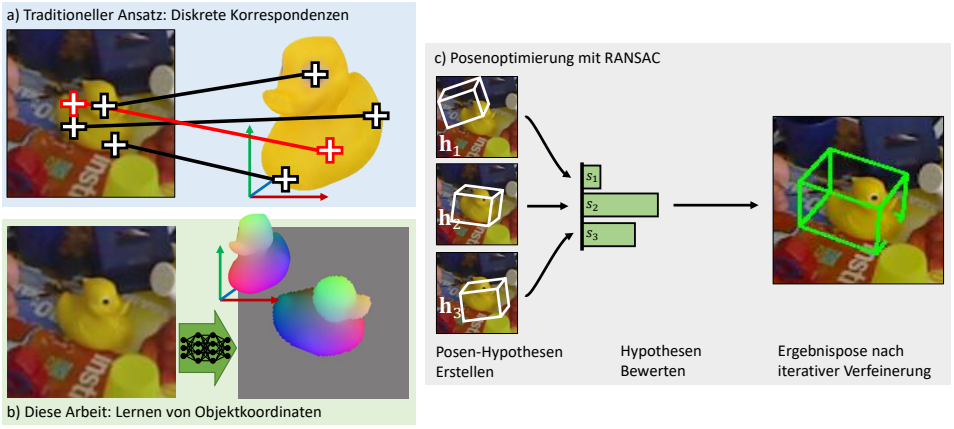



Abb. 4: **Objektkoordinaten-Regression. a)** Traditionell werden diskrete Korrespondenzen zwischen Bild und Objekt vorhergesagt. **b)** Diese Arbeit verwendet ein Lernverfahren um dichte, kontinuierliche Korrespondenzen zu schätzen. **c)** Die Posenschätzung erfolgt auf Grundlage der Korrespondenzen (a oder b) mit RANSAC.

Im traditionellen Ansatz erfolgt 6D-Posenschätzung in zwei Stufen. Zunächst wird mittels lokaler Bild-Features [Lo04] eine Anzahl von diskreten Korrespondenzen zwischen dem Eingabebild und dem Objekt gesucht, siehe Abb. 4 a). Am Beispiel der Abbildung konnten bestimmte Bildbereiche dem Auge, dem Flügel und der Vorderseite der Spielzeug-Ente richtig zugeordnet werden. In einem weiteren Fall (rot) schlug die Korrespondenzfindung fehl. Durch geometrische Optimierung kann aus den Korrespondenzen die 6D-Pose des Objekts geschätzt werden [Ka76, Ga03]. Die korrekte Pose sollte das gesuchte Objekt mit dem Bild entlang der Korrespondenzen in Übereinstimmung bringen. Erfolgt die geometrische Optimierung jedoch über alle Korrespondenzen, inklusive der stark fehlerbehafteten, wäre die geschätzte Pose nur von niedriger Qualität. Der RANSAC-Algorithmus [FB81] aus der robusten Optimierung ermöglicht genaue Schätzungen, auch wenn einige Korrespondenzen falsch sind, siehe Abb. 4 c). RANSAC wählt dazu mehrere zufällige Teilmengen von Korrespondenzen aus und erzeugt je eine Schätzung der Pose, sogenannte Hypothesen. Jede Hypothese wird dann bewertet hinsichtlich ihrer geometrischen Konsistenz mit allen übrigen Korrespondenzen. Die Hypothese mit der höchsten Konsistenz wird als finale Schätzung ausgewählt und eventuell noch durch einen iterativen Verfeinerungsprozess verbessert. Dieser Ansatz erzeugt genaue Ergebnisse und die geometrische Konsistenz des Ergebnisses lässt auf dessen Vertrauenswürdigkeit schließen. Das Verfahren kann fehlschlagen, wenn keine Korrespondenzen zwischen Objekt und Bild gefunden werden können, etwa wenn keine markanten Objekt-Strukturen für eine Zuordnung vorhanden sind.

In dieser Arbeit wird die oben genannte Strategie größtenteils beibehalten, jedoch der Schritt der Korrespondenzfindung durch ein Lernverfahren ersetzt. Zu diesem Zweck wird eine dichte, kontinuierliche Korrespondenzrepräsentation eingeführt, die sogenannten *Ob-*

*jektkoordinaten*. Jeder Punkt auf der Objektoberfläche hat eine eindeutige 3D-Koordinate im lokalen Koordinatensystem des Objekts, siehe auch Abb. 4 b) wo die Objektkoordinaten durch eine eindeutige Farbkodierung visualisiert werden. Das Lernverfahren entscheidet für jeden Bildbereich ob es sich um das Objekt oder den Hintergrund handelt. Falls der Bildbereich zum Objekt gehört, sagt das Lernverfahren weiterhin die korrespondierende 3D-Objektkoordinate voraus. Das Lernverfahren schätzt also eine Objektsegmentierung sowie ein dichtes, kontinuierliches Korrespondenzfeld in Form der Objektkoordinaten. Die Pose kann dann wie oben beschrieben mittels RANSAC geschätzt werden.

Als Lernverfahren für die Objektkoordinaten-Regression eignen sich Entscheidungsbäume [Br01], die mit gerenderten Ansichten des Objekts trainiert werden können. Dazu ist lediglich ein 3D-Modell des Objekts nötig, wie es in der Produktion oft verfügbar ist. Alternativ kann auch ein Datensatz mittels einer Tiefenkamera mit simultanem Posentracking erzeugt werden. Durch die Verwendung eines Lernverfahrens für die Korrespondenzfindung kann sich das technische System exakt auf das gewünschte Objekt spezialisieren. Damit wird auch die Posenschätzung von schwierigen, nicht-texturierten Objekten möglich, bei denen traditionelle Bild-Features versagen. Weiterhin können verschiedenste Umwelteinflüsse im Trainingsdatensatz simuliert werden, etwa starke Änderungen der Lichtverhältnisse. Das technische System lernt dann die entsprechende Robustheit. Einige experimentelle Ergebnisse können in Tabelle 1, links gesehen werden, wo das vorgeschlagene Verfahren mit einer nicht-gelernten Methode verglichen wird. Das vorgeschlagene Verfahren ist wesentlich robuster gegenüber teilweisen Verdeckungen und extremen Änderungen in der Beleuchtung.

## 3 Lernen von Unsicherheit

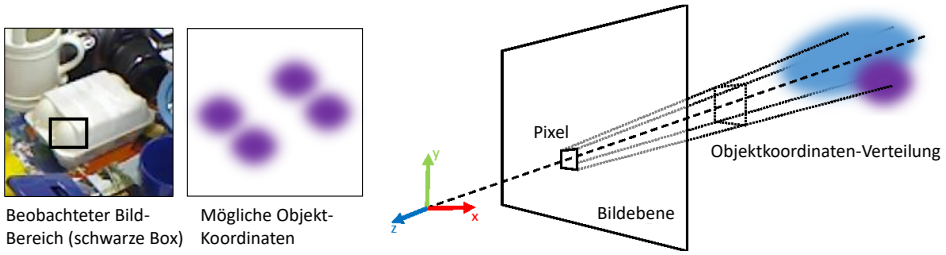



Abb. 5: **Unsichere Objektkoordinaten. Links:** Der beobachtete Bildbereich korrespondiert zu mehreren möglichen Objektkoordinaten. **Rechts:** Die optimale Pose maximiert die Likelihood der projizierten Objektkoordinaten-Verteilung.

Obwohl Lernverfahren selbst kleine Strukturmerkmale eines Objekts nutzen können, um eine Zuordnung zu Objektkoordinaten zu ermöglichen, gibt es Fälle, in denen keine eindeutige Zuordnung möglich ist, siehe Abb. 5, links. Das Objekt ist symmetrisch und der Bildbereich an der Objektecke kann nur auf vier mögliche Objektkoordinaten eingeschränkt werden. In dieser Arbeit wird die Korrespondenzvorhersage erweitert, indem *Verteilungen* von 3D-Objektkoordinaten in Form von *Gaussian Mixture Models* vorhergesagt werden. So kann das Lernverfahren Unsicherheit in der Objektkoordinaten-Schätzung

darstellen. Für eindeutige Korrespondenzen wird eine Verteilung mit einem Modus und geringer Standardabweichung vorhergesagt. Für mehrdeutige Korrespondenzen wird eine multimodale Verteilung mit gegebenenfalls hoher Standardabweichung vorhergesagt. In einer iterativen Optimierung wird die geschätzte Pose so verfeinert, dass die projizierte *Likelihood* unter den geschätzten Objektkoordinaten-Verteilungen maximiert wird, siehe Abb. 5, rechts. Die Posenverfeinerung mittels Objektkoordinaten-Unsicherheit führt insbesondere dann zu deutlich besseren Ergebnissen, wenn es sich beim Eingabebild lediglich um ein RGB-Bild handelt, siehe Tabelle 1, rechts.

| RGB-D | | | | RGB | |
|---|---|---|---|---|---|
| Methode | keine Verdeckung | mit Verdeckung | var. Lichtverhältnisse | Methode | |
| LINEMOD | 96.6% | 54.4% | 70.2% | LINE2D | 24.2% |
| Diese Arbeit | 98.1% | 67.3% | 91.8% | Diese Arbeit | 32.3% |
| | | | | Diese Arbeit (mit Unsicherheit) | 50.2% |

Tab. 1: **Ergebnisse für Objektposenschätzung.** Die Eingabe ist entweder ein RGB-D-Bild (links) oder ein RGB-Bild (rechts). Ergebnisse werden als Prozent richtig geschätzter Posen angegeben. LINEMOD [Hi12] bzw. LINE2D [Hi11] sind nicht-gelernte Verfahren.

## 4 Differenzierbare Robuste Optimierung

Ein wesentlicher Faktor für den Erfolg von *Deep Learning* ist das Ende-zu-Ende-Training. Alle Schichten eines neuronalen Netzes, von den Bild-Features bis zu den semantischen Repräsentationen, können sich optimal aneinander anpassen um eine maximale Genauigkeit zu erreichen. Gleichermaßen soll die Objektkoordinaten-Regression in dieser Arbeit möglichst so gelernt werden, dass sich ihre Voraussagen gut für die robuste Posenoptimierung eigenen. Leider ist ein Trainieren des Gesamtsystems ende-zu-ende nicht möglich, da der RANSAC-Algorithmus nicht differenzierbar ist. Das bedeutet, dass die Korrekturrichtungen für eine Fehlerminimierung nicht durch RANSAC hindurch an die Objektkoordinaten-Regression weitergeleitet werden können.

Diese Arbeit stellt eine Variante von RANSAC vor, welche differenzierbar ist und daher im Rahmen von *Deep Learning* verwendet werden kann. Im Zentrum des differenzierbaren RANSAC, kurz *DSAC*, steht die Minimierung des folgenden Erwartungswerts:

$$\mathbb{E}_{j\sim p(j)}\left[\ell(\mathbf{h}_j)\right]$$

Dabei bezeichnet $p(j)$ eine Wahrscheinlichkeitsverteilung über alle RANSAC-Hypothesen $\mathbf{h}_j$, die sich aus den individuellen Konsistenzbewertungen ergibt, und $\ell(\cdot)$ ist eine Fehlerfunktion, die die Genauigkeit einer Pose angibt. Durch die Minimierung dieses Ausdrucks lernt das System gute Hypothesen hoch zu bewerten und ihre Genauigkeit weiter zu steigern und schlechte Hypothesen schlechter zu bewerten wobei ihre Genauigkeit keine Rolle spielt. Die Genauigkeitssteigerung durch ein Trainieren des Systems ende-zu-ende mittels DSAC kann in Tabelle 2 nachvollzogen werden. Im Moment

des Schreibens dieses Textes ist das hier vorgestellte System weltweit führend hinsichtlich der Genauigkeit im Kamera-Relokalisierungsproblem auf Basis eines einzelnen RGB-Bildes [BR18b]. Der vorgestellte DSAC-Algorithmus ist nicht auf die Anwendung in der Posenschätzung beschränkt, sondern kann, ähnlich wie RANSAC, für viele Probleme der Computer Vision und anderen Wissenschaftszweigen verwendet werden.

| | Methode | Bilder korrekt | Genauigkeit (Median) |
|---|---|---|---|
| Verwandte Arbeiten | PoseNet [KGC15] | - | 44.6cm, 9.8 |
| | ORB-Feat. + RANSAC [Sh13] | 38.6% | - |
| | Active Search [SLK12] | - | 5.1cm, 2.5 |
| Diese Arbeit | Entscheidungswald + RANSAC | 55.2% | 4.5cm, 2.0 |
| | Neuronales Netz + RANSAC | 61.0% | 4.0cm, 1.6 |
| | Neuronales Netz + DSAC | 66.2% | 3.5cm, 1.6 |
| | Neuronales Netz + DSAC v.2 | 76.1% | 3.6cm, 1.1 |

Tab. 2: **Ergebnisse für Kamera-Relokalisierung.** Von allen aktuellen Verfahren erzielt der Ansatz dieser Arbeit die höchste Genauigkeit. DSAC v.2 wurde auf Grundlage dieser Arbeit in [BR18b] veröffentlicht.

## Fazit

Naturgemäß kann diese Kurzfassung nur einen oberflächlichen Überblick über die zugrundeliegende Dissertation geben. Insbesondere auf methodische, technische und experimentelle Details musste verzichtet werden. Beispielsweise kann das vorgestellte Verfahren die Posen von mehreren Dutzend Objekten gleichzeitig schätzen und weist damit eine gute Skalierbarkeit auf. Weiterhin kann das System über die Konsistenzprüfung einer Posenschätzung entscheiden, ob ein gesuchtes Objekt im Bild zu sehen ist oder nicht. Es wurde für artikulierte, also deformierbare, Objekte erweitert [Mi15] und in einen Echtzeittracker eingebettet [Kr14]. Die Dissertation gibt am Beispiel der 6D-Posenschätzung eine Antwort auf die Frage, wie die neuen Werkzeuge des *Deep Learning* genutzt werden können, ohne frühere Forschungsergebnisse, etwa aus der Computer Vision, komplett zu verwerfen. Insbesondere mit dem DSAC-Algorithmus stellt diese Arbeit ein wichtiges Werkzeug zur Verfügung, mit dem die mächtigen neuronalen Netze in etablierte technische Systeme eingebettet werden können.

## Literaturverzeichnis

[Br01] Breiman, Leo: Random Forests. Machine Learning, 2001.

[Br18a] Brachmann, Eric: Learning to Predict Dense Correspondences for 6D Pose Estimation. Dissertation, Dresden University of Technology, Germany, 2018.

[BR18b] Brachmann, Eric; Rother, Carsten: Learning Less Is More - 6D Camera Localization via 3D Surface Regression. In: CVPR. 2018.

[Br18c] Brahmbhatt, Samarth; Gu, Jinwei; Kim, Kihwan; Hays, James; Kautz, Jan: MapNet: Geometry-Aware Learning of Maps for Camera Localization. In: CVPR. 2018.

[FB81] Fischler, M. A.; Bolles, R. C.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Commun. ACM, 1981.

[Ga03] Gao, Xiao-Shan; Hou, Xiao-Rong; Tang, Jianliang; Cheng, Hang-Fei: Complete solution classification for the perspective-three-point problem. TPAMI, 2003.

[Hi11] Hinterstoisser, S. Holzer, S.; Cagniart, C.; Ilic, S.; Konolige, K.; Navab, N.; Lepetit, V.: Multimodal Templates for Real-Time Detection of Texture-less Objects in Heavily Cluttered Scenes. In: ICCV. 2011.

[Hi12] Hinterstoisser, S.; Lepetit, V.; Ilic, S.; Holzer, S.; Bradski, G.; Konolige, K.; Navab, N.: Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes. In: ACCV. 2012.

[Ka76] Kabsch, Wolfgang: A solution for the best rotation to relate two sets of vectors. Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography, 1976.

[KC17] Kendall, Alex; Cipolla, Roberto: Geometric loss functions for camera pose regression with deep learning. In: CVPR. 2017.

[KGC15] Kendall, A.; Grimes, M.; Cipolla, R.: PoseNet: A Convolutional Network for Real-Time 6-DoF Camera Relocalization. In: ICCV. 2015.

[Kr14] Krull, Alexander; Michel, Frank; Brachmann, Eric; Gumhold, Stefan; Ihrke, Stephan; Rother, Carsten: 6-DoF Model Based Tracking via Object Coordinate Regression. In: ACCV. 2014.

[Lo04] Lowe, David G.: Distinctive Image Features from Scale-Invariant Keypoints. IJCV, 2004.

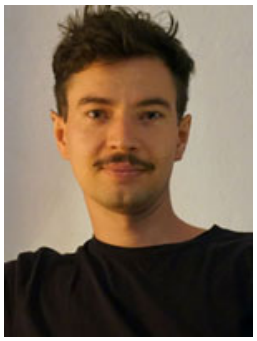
[Mi15] Michel, F.; Krull, A.; Brachmann, E.; Yang, M. Y.; Gumhold, S.; Rother, C.: Pose Estimation of Kinematic Chain Instances via Object Coordinate Regression. In: BMVC. 2015.

[Sh13] Shotton, J.; Glocker, B.; Zach, C.; Izadi, S.; Criminisi, A.; Fitzgibbon, A.: Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images. In: CVPR. 2013.

[SLK12] Sattler, Torsten; Leibe, Bastian; Kobbelt, Leif: Improving Image-Based Localization by Active Correspondence Search. In: ECCV. 2012.

[St12] Stallkamp, Johannes; Schlipsing, Marc; Salmen, Jan; Igel, Christian: Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. Neural networks, 2012.

[Wa17] Walch, Florian; Hazirbas, Caner; Leal-Taixé, Laura; Sattler, Torsten; Hilsenbeck, Sebastian; Cremers, Daniel: Image-based Localization with Spatial LSTMs. In: ICCV. 2017.



**Eric Brachmann**, geboren 1987, studierte Medieninformatik an der TU Dresden von 2006-2012 und schloss das Diplom mit Auszeichnung ab. Unmittelbar im Anschluss promovierte er am Lehrstuhl für Computergraphik und Visualisierung von Prof. Gumhold an der TU Dresden. Von 2015-2017 war er zusätzlich Mitarbeiter des Computer Vision Lab Dresden von Prof. Rother. Seit dem Abschluss der Promotion im Jahr 2018, arbeitet Eric Brachmann als PostDoc am Visual Learning Lab von Prof. Rother an der Universität Heidelberg. Sein Forschungsinteresse gilt der Verbindung traditioneller Verfahren der Computer Vision mit den Möglichkeiten des Deep Learnings.

# Verhaltenssensitive Nutzerschnittstellen für mobile Geräte mit berührungsempfindlichem Bildschirm[1]

Daniel Buschek[2]

**Abstract:** Diese Dissertation zeigt wie Interaktion mit mobilen Geräten mit Touchscreen verbessert werden kann, indem Nutzerschnittstellen in die Lage versetzt werden, zu erfassen und zu berücksichtigen *wie* Nutzer Interaktionen damit ausführen. Diese *verhaltenssensitiven* Nutzerschnittstellen verbessern Interaktion durch die Verwendung von Verhaltensdaten, welche bei der Nutzung ohnehin anfallen. Somit können dem Nutzer Zeit und Unterbrechungen erspart werden, die andernfalls nötig wären um dieselbe Information zu erfragen (z.B. Informationen zur Nutzeridentität, Handhaltung sowie weiterem Kontext). Zentrale Forschungsfragen adressieren deshalb das Verständnis von Verhaltenscharakteristika und -einflüssen, deren Modellierung, sowie Konzepte zu Inferenz und Reaktion und deren Integration in die Nutzerschnittstellen. Die Dissertation untersucht dies in mehreren Studien. Dabei werden sowohl grundlegende Verhaltensaspekte von Touch-Interaktionen auf mobilen Geräten analysiert, als auch Anwendungen mittels nutzbarer Prototypen betrachtet. Darüber hinaus stellt die Dissertation Konzepte und implementierte Software vor für das Sammeln von Verhaltensdaten, der Analyse und Modellierung von Interaktionsverhalten, sowie dem Erstellen von verhaltenssensitiven Nutzerschnittstellen. Diese Beiträge unterstützen Forscher und Entwickler bei der Untersuchung und praktischen Umsetzung von solchen Nutzerschnittstellen. Insgesamt zeigt die vorliegende Arbeit, wie Verhaltensmerkmale in der Interaktion mit mobilen Geräten mit Touchscreen zum Vorteil der Nutzer adressiert und genutzt werden können. Außerdem werden Transfer und Wiederverwendung von Verhaltensinformationen und -modellen diskutiert, zum Beispiel im Hinblick auf Kompromisse zwischen Bedienbarkeit und Privatsphäre bzw. Sicherheit. Schließlich reflektiert die Arbeit die generelle Rolle verhaltenssensitiver Nutzerschnitstellen. Dabei zeichnet sie eine Perspektive, in der solche Nutzerschnitstellen der direkten Einbettung von "Erwartungen" an das Bedienverhalten in intelligente interaktive Systeme dienen.

## 1 Einleitung und Überblick

Mobile Geräte mit berührungsempfindlichem Bildschirm ("Touchscreen") sind unerlässliche Alltagshelfer geworden, sei es zur Kommunikation, Informationssuche oder zum Erstellen und Abrufen persönlicher Daten. Insbesondere Smartphones werden oft als persönliche Geräte betrachtet, die einem bestimmten Nutzer gehören und für diesen den individuellen Zugang zum mobilen digitalen Leben bedeuten. Diese persönliche Nutzung, kombiniert mit der Bedeutung des Touchscreens zur Interaktion, motiviert die zentrale Aussage dieser Dissertation: Interaktion mit mobilen Geräten mit Touchscreen kann verbessert werden, indem Nutzerschnittstellen erfassen und berücksichtigen *wie* Nutzer Interaktionen damit ausführen. Diese Dissertation führt den Begriff "behaviour-aware" (d.h. "verhaltenssensitiv") ein, um solche Nutzerschnittstellen zu charakterisieren.

[1] Englischer Originaltitel: *Behaviour-Aware Mobile Touch Interfaces*

[2] LMU München, daniel.buschek@ifi.lmu.de

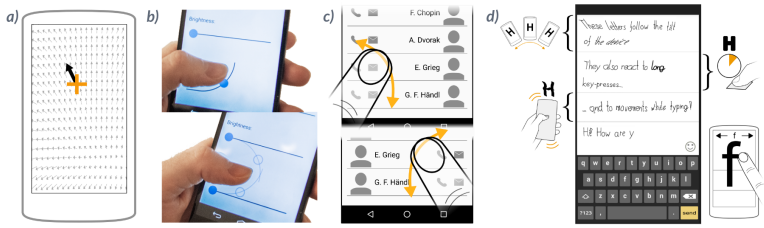



Abb. 1: Beispiele für in dieser Dissertation entwickelte verhaltenssensitive Nutzerschnittstellen auf mobilen Geräten: *(a)* Ein Modell des individuellen Verhaltens beim Zielen mit dem Finger korrigiert die Eingabe des Nutzers und erhöht so die Zielgenauigkeit um Eingabefehler zu verringern. *(b)* Die Regler passen ihre Form an den Radius des Daumens an und ermöglichen so ergonomische einhändige Bedienung (oberes Bild). Darüber hinaus geben sie dem Nutzer Rückmeldung über Unsicherheit in der Interpretation der Eingabe (durch transparente "Vorschau", unteres Bild). *(c)* Diese Kontaktliste vertauscht die links-/rechtsbündige Anordnung von Kontaktnamen und Schaltflächen, basierend auf der Trajektorie des Fingers beim Scrollen. Somit passt sich die Nutzerschnittstelle z.B. an links-/rechtshändige Bedienung an. *(d)* Diese Tastatur erzeugt eine individuelle Schriftart basierend auf Aspekten des Eingabeverhaltens, um persönliche und ausdrucksstarke Kommunikation per Textnachrichten zu ermöglichen, ähnlich den Einflüssen auf Handschrift in der analogen Welt.

Diese Nutzerschnittstellen verbessern Interaktion mittels Verhaltensdaten, die bei der Nutzung ohnehin anfallen. So können dem Nutzer Zeit und Unterbrechungen erspart werden, die sonst nötig wären um diese Daten zu erfragen (z.B. Informationen zur Nutzeridentität, Handhaltung, sowie weiterem Kontext). Verhaltenssensitive Nutzerschnittstellen können solche Informationen auf verschiedene Arten nutzen, insbesondere um sich an Nutzer anzupassen. Zentrale Forschungsfragen adressieren daher das Verständnis von Nutzerverhalten, dessen Modellierung, und Konzepte zu Inferenz, Reaktion und deren Integration in Nutzerschnittstellen. Diese Fragen werden in der Dissertation in mehreren Studien untersucht. Es werden sowohl grundlegende Aspekte von Interaktionen mit Touchscreens auf mobilen Geräten analysiert, als auch konkrete Anwendungen in drei Bereichen:

1) *Verbesserung der Eingabeleistung*, durch Modellierung des invidividuellen Nutzerverhaltens beim Zielen mit dem Finger, um zukünftige Eingaben zu korrigieren und die Genauigkeit der Eingabe zu verbessern. 2) *Förderung von Privatsphäre und Sicherheit*, durch Analyse von Mustern im Ziel-/Tippverhalten, womit Nutzer während der Eingabe identifiziert werden können. 3) *Verbesserung der Ausdrucksstärke* von Interaktionen, durch Verknüpfen von Verhaltensmerkmalen in der Eingabe mit Eigenschaften der Ausgabe.

Darüber hinaus bietet die Dissertation Konzepte und implementierte Werkzeuge für das Sammeln von Verhaltensdaten aus Interaktion, der Analyse und Modellierung von Verhalten, sowie dem Erstellen von verhaltenssensitiven und adaptiven Nutzerschnittstellen für mobile Geräte mit Touchscreen. Diese Beiträge unterstützen Forscher und Entwickler bei der Untersuchung und praktischen Umsetzung von solchen Nutzerschnittstellen.

Insgesamt zeigt die Arbeit, wie Verhaltensmerkmale in der Interaktion mit mobilen Geräten mit Touchscreen zum Vorteil der Nutzer genutzt werden können. Es werden außerdem Transfer und Wiederverwendung von Verhaltensinformationen und -modellen

diskutiert, zum Beispiel im Hinblick auf Kompromisse zwischen Bedienbarkeit und Privatsphäre bzw. Sicherheit. Schließlich reflektiert die Arbeit die generelle Rolle verhaltenssensitiver Nutzerschnitstellen. Dabei zeichnet sie eine Perspektive, in der diese der direkten Einbettung von "Erwartungen" an das Bedienverhalten in interaktive Systeme dienen.

## 2 Verhaltenssensitive Nutzerschnittstellen für mobile Touch-Geräte

### 2.1 Analyse von Eingabeverhalten auf mobilen Geräten mit Touchscreen

Die Grundlage für verhaltenssensitive Nutzerschnittstellen bilden drei Elemente: 1) Detailliertes Verständnis des Bedienverhaltens, 2) Modelle um dieses Verhalten zu repräsentieren, sowie 3) Methoden um es nutzbar zu machen und Informationen über Nutzer und Kontext zu erschließen. Die Dissertation deckt diese Elemente für Bedienverhalten auf mobilen Geräten mit Touchscreen mit den Fingern ab ("Touches"). Dazu wurden mehrere Studien mit Nutzern durchgeführt [BRMS13, BA15, BDLA16, BKA17, Bu18].

Die Arbeit untersucht als erste Einflüsse der Handhaltung (z.B. Eingabe per Daumen/Zeigefinger) [BRMS13, BA15, BDLA16] und grafischer Elemente [BA15, BDLA16] auf das Zielverhalten und dessen Modellierung ("Offset Models", siehe Abbildung 1a). Mit "Zielverhalten" ist die Fingerplatzierung und Genauigkeit beim Zielen auf grafische Elemente gemeint. Durch das in den Studien erlangte Wissen können die Modelle praktisch besser eingesetzt werden, um zuverlässig Zielgenauigkeit und Eingabeleistung zu verbessern.

Die Arbeit zeigt zudem zum ersten Mal [BRMS13], wie sich Muster im Zielverhalten eines Nutzers zwischen verschiedenen Gerätemodellen unterscheiden und übertragen werden können. Dies ist praktisch relevant um es Nutzern zu ermöglichen, auch auf einem neuen Gerät direkt mit personalisierten Verbesserungen der Eingabeleistung starten zu können.

Darüber hinaus überträgt die Dissertation zum ersten Mal diese Modelle auf Eingaben mit einem Stift [BKA17]. Es wird insbesondere gezeigt, dass auch die Eingabe mit dem vermeintlich genaueren Stift durch die Modellierung des Zielverhaltens weiter verbessert werden kann. Gleichzeitig ergeben sich durch die Gegenüberstellung der Muster von Finger und Stift neue Erkenntnisse über grundlegendes Eingabeverhalten: So ist zum Beispiel der Stift zwar im Durchschnitt genauer als der Finger, allerdings schwankt der Unterschied stark in Abhängigkeit der Position des Ziels auf dem Bildschirm.

Insgesamt zeigt die Dissertation die zentralen Herausforderungen bei der Verbesserung der Zielgenauigkeit durch Modellierung der individuellen Fingerplatzierung auf. Insbesondere erschweren mehrere Einflussfakoren robuste Verbesserungen der Zielgenauigkeit (z.B. wechselnde Handhaltungen des Geräts im Alltag). Gleichzeitig wandelt die Arbeit diese Herausforderungen in eine neue vielversprechende Perspektive um: Eingabeverhalten, das durch Kontextfaktoren beeinflusst wird, kann auch von verhaltenssensitiven Nutzerschnittstellen analysiert werden um eben jene Kontexte zu erkennen.

## 2.2 Modellierung und Inferenz auf Basis des Eingabeverhaltens

Die Dissertation trägt dazu bei diese Perspektive praktisch umzusetzen – mit Modellen und Inferenzmethoden, sowie Konzepten zu deren Einbettung in Benutzeroberflächen. Zum einen ergeben die Studien grundlegende Erkenntnisse für Modellentscheidungen und -parameter [BRMS13, BA15, BDLA16, BKA17, Bu18]. Zudem wird ein neues *inverses* Modell des Zielverhaltens vorgestellt, mit dem Nutzerschnittstellen vorhersagen können, wo auf dem Touchscreen Eingaben mit dem Finger zu erwarten sind. Dies ist nützlich um Nutzerverhalten zu simulieren [BDLA16] und Bedienelemente zu optimieren [BAA15].

Die Disseratation entwickelt zudem ein weiteres neues Modell des Zielverhaltens: Dieses betrachtet zum ersten Mal die Fingerplatzierung in *Sequenzen* über mehrere Eingaben (Touches) hinweg. Damit kann eine Veränderung im Verhalten über die Zeit erkannt werden. Dies hat hohe praktische Relevanz für verhaltensbiometrische Systeme, die automatisch erkennen möchten wenn ein Gerät einem anderen Nutzer übergeben wurde (z.B. automatischer "Gastmodus") oder entwendet wurde (z.B. Gerät automatisch sperren).

Die Arbeit stellt außerdem eine grundlegende Inferenzmethode mittels Modellen zum Zielverhalten vor: Damit kann von Mustern in der Fingerplatzierung auf dem Touchscreen auf die aktuelle Nutzungssituationen geschlossen werden. Konkret untersucht die Arbeit wie Nutzer identifiziert [BRMS13, BDLA16] und Handhaltung (links/rechts) sowie Finger (Daumen/Zeigefinger) bzw. Stifteingabe erkannt werden können [BRMS13, BA15, BKA17]. Diese Informationen können dann genutzt werden, um die Nutzerschnittstelle automatisch an die Situation anzupassen. Ein Vorteil gegenüber anderen Ansätzen besteht darin, dass die Informationen anhand der üblichen Bedienung des Touchscreens gewonnen werden. Somit sind zum Beispiel keine weiteren Sensoren am Gerät notwendig.

Das Erschließen von Informationen verlangt nach ausdrucksstarken Verhaltensmerkmalen ("Features"), insbesondere zur Unterscheidung von Nutzern in verhaltensbiometrischen Systemen. Die Arbeit trägt hier entscheidende Erkenntnisse bei [BDLA15a, BBA18]: So konnte gezeigt werden, dass gerade die Muster in der Fingerplatzierung höchst individuell sind, auch im Tippverhalten auf der Bildschirmtastatur (z.B. vereinfacht: Nutzer A tippt genauer in der oberen linken Ecke des Bildschirms bzw. der Tastatur, Nutzer B aber unten rechts). Insbesondere bieten diese Merkmale eine viel bessere Grundlage für Verhaltensbiometrie auf mobilen Geräten als die bislang genutzten zeitlichen Merkmale (z.B. Eingabeschwindigkeit). Ein Vergleich zwischen Labor- und Feldstudien [BDLA15a, BBA18] zeigt zudem, dass der Vorteil dieser Merkmale im Alltag noch höher ist.

Schließlich stellt die Dissertation auch ein grundlegendes Konzept vor (*ProbUI*), wie solche Modelle des Eingabeverhaltens direkt in grafische Bedienoberflächen eingebettet werden können [BA17]. Dieses Konzept wird auch im nächsten Abschnitt näher vorgestellt.

## 2.3 Anwendungsbereiche verhaltenssensitiver Nutzerschnittstellen

Die oberen erwähnten Anwendungsbereiche werden nun näher betrachtet. Als Motivation diskutiert die Arbeit in einem Essay [Bu16] eine neue weitergefasste Perspektive über die

technischen Konzepte hinaus: Russel Belk's Konzept des *"Extended Self"* [Be88, Be13] wird aufgegriffen ("erweitertes Selbst", kurz ES), wonach Menschen (digitale) Objekte nutzen, um ihre Identität zu definieren, zu reflektieren und zu kommunizieren – mit bewusst philosophischer Perspektive – in Bezug auf die abstrakten Funktionen "Haben", "Tun", und "Sein" [Be88]. Die Dissertation diskutiert, dass auch verhaltenssensitive Nutzerschnittstellen grundlegend an *individuellem* menschlichen Verhalten interessiert sind. Sie motiviert daher in Anlehnung an die drei Funktionen aus ES drei Anwendungsbereiche: 1) Die Nutzung von Verhaltensmerkmalen um Geräte und Daten zu schützen ("Haben"); 2) die Anpassung von Nutzerschnittstellen an das Verhalten, um die Bedienung zu verbessern ("Tun"); und 3) die Berücksichtigung von Verhaltensmerkmalen zur personalisierten Darstellung des Nutzers ("Sein"), zum Beispiel in Kommunikationsanwendungen. Die Beiträge der Dissertation in diesen Bereichen werden im Folgenden zusammengefasst.

Im Bereich der *Förderung von Privatsphäre und Sicherheit* untersucht die Dissertation in mehreren empirischen Studien biometrische Systeme, die Nutzer anhand des Eingabeverhaltens auf dem Touchscreen identifizieren. Metehodisch wird hier zum ersten Mal der Einfluss von bislang typischen Annahmen auf die Auswertung zugrunde liegender Modelle quantifiziert [BDLA15a]. Einige Annahmen werden als in realistischen Nutzungsszenarien im Alltag nicht haltbar identifiziert (z.B. Annahme fixer Handhaltung oder Kenntnis von Verhaltensdaten von einem spezifischen Angreifer noch *vor* dem Angriff). Damit zeigt die Arbeit quantifizierbar, dass bei einigen verbreiteteten Evaluationsschemata ein viel zu optimistisches Bild von verhaltensbiometrischen Systemen in diesem Kontext entsteht.

Außerdem quantifiziert die Dissertation in diesem Anwendungsbereich den Einfluss von Handhaltung und Eigenschaften der grafischen Bedienelemente auf den biometrischen Wert (d.h. Individualität) des Nutzerverhaltens [BDLA15a, BDLA16].

Schließlich stellt die Arbeit mit der entwickelten Software "*ResearchIME*" ein Werkzeug für Wissenschaftler bereit, mit dem zum ersten Mal "natürliche" verhaltensbiometrische Daten bei der Texteingabe im Alltag gesammelt werden können [BBA18]. Dabei spielt die Wahrung der Privatsphäre der Studienteilnehmer eine zentrale Rolle, was konkret durch ein neues mehrstufiges Filterkonzept umgesetzt wird (siehe auch Abschnitt 2.4).

Im Anwendungsbereich der *Verbesserung der Eingabeleistung* untersucht die Dissertation adaptive Nutzerschnittstellen mit zwei Ansätzen: 1) Adaption der Interpretation der Eingabe im Hintergrund; sowie 2) sichtbare Adaption der grafischen Bedienoberfläche. Der erste Ansatz setzt auf vorigen Arbeiten auf [We12] und nutzt die oben beschriebenen Modelle des Zielverhaltens. Diese werden auf den Eingabedaten ("Touches") eines Nutzers trainiert um dann die Genauigkeit zukünftiger Eingaben individuell zu verbessern [BRMS13, BA15, BKA17]. Somit können Fehler bei der Eingabe reduziert werden. Die Dissertation entwickelt diese Modelle dabei entscheidend weiter und untersucht wichtige Einflussfaktoren (siehe oben, Abschnitte 2.1 und 2.2).

Für den zweiten Ansatz stellt die Dissertation das oben erwähnte neue Konzept "*ProbUI*" vor, welches die dynamische Adaption von grafischen Bedienoberflächen an das Nutzerverhalten während der Nutzung ermöglicht [BA17]. Im Gegensatz zu vorherigen Arbeiten geschieht dies auf einer systematischen und formal einheitlichen Grundlage, die

auch explizit Entwickler bei der Umsetzung unterstützt. Das Konzept wurde zudem als "open-source" Implementierung für Entwickler nutzbar gemacht. Somit wird eine Brücke von theoretischer Modellierung zu praktischer Anwendung geschlagen. Konkret ersetzt *ProbUI* die üblichen internen Repräsentationen von Zielbereichen in grafischen Nutzerschnittstellen mit einem probabilistischen Modell. Dies ermöglicht es Entwicklern, Varianten im Nutzerverhalten zu antizipieren und mit geringem Aufwand Adaptionen und Rückmeldungen der Bedienoberfläche für die Nutzer zu implementieren. Abbildungen 1b) und c) zeigen konkrete Beispiele für solche verhaltenssensitive Nutzerschnittstellen.

Zur *Verbesserung der Ausdrucksstärke* von Interaktionen wird ein neues Konzept ("*TapScript*") für eine Bildschirmtastatur als verhaltenssensitive Nutzerschnittstelle vorgestellt [BDLA15b]. Diese Tastatur passt die Textdarstellug automatisch an, in Abhängigkeit des aktuellen Tippverhaltens (siehe Abbildung 1d). Berücksichtigt werden unter anderem Zielgenauigkeit, Tippgeschwindigkeit, sowie Bewegung des Geräts/Körpers. Die Idee ist inspiriert von der Art und Weise wie sich Einflüsse im handschriftlichen Schriftbild zeigen. In empirischen Untersuchungen wird gezeigt, dass Nutzer Nachrichten mit *TapScript* als persönlicher empfinden und anhand des Schriftbilds andere Nutzer wiedererkennen können. Außerdem können einfache Kontexte (stationär/unterwegs) unterschieden werden. Neben der praktischen Anwendung zeigt *TapScript* auf theoretischer Ebene ein neues Konzept: Ursprünglich unbewusste und "unsichtbare" Verhaltensvariationen werden für den Nutzer sichtbar gemacht. Dies ermöglicht neue Eingabedimensionen.

Schließlich beschreibt diese Dissertation eine umfassende Perspektive über Anwendungen und Geräte hinweg: In allen drei untersuchten Bereichen können ähnliche Verhaltensmerkmale und Modelle eingesetzt werden. Dies impliziert eine Systemarchitektur, die unter Beachtung von Datenschutz einen solchen Austausch über Anwendungen und Geräte hinweg möglich macht. Als einen ersten technischen Schritt stellt die Arbeit einen Ansatz vor, mit dem Nutzermodelle automatisch von einem auf ein anderes Gerät angepasst werden können [BRMS13]. Darüber hinaus werden mögliche Vorteile einer solchen Perspektive auf verhaltenssensitive Nutzerschnittstellen im Ausblick beleuchtet (siehe Abschnitt 3).

### 2.4 Methodik und Werkzeuge für verhaltenssensitive Nutzerschnittstellen

Die Dissertation stellt Methoden und Werkzeuge vor, mit denen Wissenschaftler und Entwickler verhaltenssensitive Nutzerschnittstellen erforschen und implementieren können. Ein erster wichtiger Schritt ist das Erfassen von Verhaltensdaten. Um Wissenschaftler dabei zu unterstützen, wurde eine neue Methode und Software ("*ResearchIME*") entwickelt [BBA18]: Diese Tastatur-App erfasst Merkmale der Texteingabe im Alltag. Zuvor wurde dies in verwandten Arbeiten ausschließlich im Labor oder in künstlich gestellten Aufgaben untersucht. Eine zentrale Herausforderung im Alltag ist die Wahrung der Privatsphäre der Nutzer. Daher wurde ein neues dreistufiges Filterkonzept entwickelt: 1) *Automatischer Filter* – Sensible Eingabefelder (z.B. Namen, Passwörter, Telefonnummern) werden automatisch erkannt und nicht aufgezeichnet. 2) *Zufallsfilter ("Subsampling")* – Für andere Eingaben werden Verhaltensmerkmale für eine stark beschränkte Zufallsauswahl an Tastenanschlägen gespeichert, wodurch der eingegebene Text nicht rekonstruiert

werden kann. Die Zufallsauswahl ist dabei so gestaltet, dass die Daten dennoch für viele Forschungsfragen nützlich sind. 3) *Filter durch Teilnehmer selbst* – Schließlich stellt die entwickelte Tastatur auch eine Schaltfläche bereit, mit der die Teilnehmer selbst die Erfassung der Daten komplett unterbrechen können. Die Auswirkung der Filterparameter auf die Rekonstruierbarkeit und damit Privatsphäre wurde empirisch untersucht [BA17].

Neben der Datenerfassung werden auch Verhaltensmodelle benötigt. Um Wissenschaftler und Entwickler hier zu unterstützen wurde das *TouchML* Softwarepaket entwickelt. Dieses implementiert die entwickelten Modelle der Fingerplatzierung bei der Eingabe. Es unterstützt Datenanalyse und mobile Anwendungen sowie Einbettung in mobile Webseiten. Somit kann die Eingabegenauigkeit in solchen Anwendungen verbessert werden. Zudem können die Modelle für das Erkennen von Nutzern oder Kontexten verwendet werden.

Schließlich unterstützt die Arbeit mit dem beschriebenen *ProbUI* Konzept und Softwarepaket die praktische Erstellung von verhaltenssensitiven Nutzerschnittstellen. Abbildungen 1b) und c) zeigen Beispiele. Zentral ist die neue Verknüpfung von drei in früheren Arbeiten getrennten Ansätzen: 1) Entwickler nutzen eine *deklarative Sprache* um Nutzerverhalten zu referenzieren. 2) Das System leitet daraus sowie aus Eigenschaften der grafischen Bedienoberfläche automatisch eine *wahrscheinlichkeitsbasierte* interne Repräsentation ab. 3) Diese Modelle schätzen während der Interaktion kontinuierlich die Nutzerintention, was Entwickler im Programmcode nutzen können, um Feedback anzuzeigen oder Bedienelemente dynamisch anzupassen. Durch diese neue Verknüpfung von deklarativen und probabilistischen Ansätzen müssen Entwickler keine Experten in probabilistischer Modellierung sein, um solche verhaltenssensitiven Nutzerschnittstellen formal systematisch umzusetzen.

## 3 Schlussfolgerungen

Selbst für eine scheinbar simple Aktion, wie das Berühren einer Schaltfläche auf einem Touchscreen, zeigt sich, dass Bedienverhalten stark zwischen Menschen und Kontexten variiert. Verhaltenssensitive Nutzerschnittstellen bieten deshalb die Möglichkeit nichttriviale Informationen über Nutzer und Kontext aus dem Bedienverhalten abzuleiten. Solche Informationen auf andere Weise zu gewinnen bedingt oft zusätzliche “Kosten” für die Nutzer, wie zum Beispiel eine Verzögerung der Hauptaufgabe (z.B. Email schreiben) durch einen Authentifizierungsvorgang (z.B. Passworteingabe). Kann die Nutzeridentität hingegen anhand des Nutzungsverhaltens während der Hauptaufgabe direkt erschlossen werden, so spart das den Nutzern Zeit und Unterbrechungen.

Konkret zeigt die Arbeit damit auch eine neue Perspektive auf, wie Mensch-Maschine-Interaktion im Sinne von Informationsübertragung optimiert werden kann: Dies kann nicht nur durch die Entwicklung von neuen Bedienkonzepten und -geräten geschehen, sondern auch dadurch, dass bestehenden Nutzerschnittstellen technisch ermöglicht wird, reichhaltigere Informationen aus dem gewohnten Nutzungsverhalten zu gewinnen. In der vorliegenden Arbeit geschieht dies zudem ohne zusätzliche Anforderungen an die Hardware zu stellen, weshalb die erarbeiteten Verbesserungen direkt auf Millionen von Geräten per Software-Update zur Anwendung kommen können.

Es wurden drei Anwendungsbereiche von verhaltenssensitiven Nutzerschnittstellen auf mobilen Geräten mit Touchscreen untersucht: 1) Verbesserung der Eingabeleistung, 2) Förderung von Privatsphäre und Sicherheit, und 3) Verbesserung der Ausdrucksstärke von Interaktionen. Abschließend macht die Arbeit nun deutlich, dass (dasselbe) Eingabeverhalten durch verhaltenssensitive Nutzerschnittstellen und die zu Grunde liegenden Modelle über *mehrere* Anwendungen und Geräte hinweg genutzt werden können. Ein Beispiel verdeutlicht die Vorteile: Um praktische Kompromisse aus Benutzbarkeit und Sicherheit für eine Zielgruppe zu verbessern, wird häufig versucht die für die Sicherheit verantwortlichen Teile von Nutzerschnittstellen zu verbessern (z.B. neue Authentifizierungskonzepte wie Eingabe von Mustern statt Passwörtern). Im Gegensatz dazu zeigt diese Arbeit auf, dass durch verhaltenssensitive Nutzerschnittstellen eine breitere Perspektive auf solche Abwägungen ermöglicht wird: Im Beispiel könnte das Verhalten bei der Eingabe eines Passworts auch dazu verwendet werden, um die Handhaltung zu erschließen und damit die auf die Authentifizierung folgenden Nutzerschnittstellen zu optimieren (z.B. Layout von Schaltflächen oder Interpretation von Tippverhalten). Somit wird eine Verbesserung der Nutzbarkeit des Gesamtsystems erreicht ohne dass Designer, Entwickler oder Forscher dabei auf eine Veränderung der Passworteingabe an sich beschränkt werden.

Genereller formuliert zeigt diese Dissertation daher auf, dass Erfassung und Nutzung von Informationen zum Bedienverhalten nicht am selben Punkt stattfinden müssen. Damit zeichnet die Arbeit eine Vision von *umfassenden* verhaltenssensitiven Nutzerschnittstellen: Ein solcher Transfer von Verhaltensinformation und -modellen über Anwendungsbereiche und Geräte hinweg ist besonders vielversprechend in allgegenwärtigen, vernetzten Computersystemen (“ubiquitous computing”).

Als Ausblick soll an dieser Stelle auch beleuchtet werden, welche Rolle verhaltenssensitive Nutzerschnittstellen für die Zukunft von Mensch-Maschine-Interaktion in Forschung und Industrie spielen. Eine Paneldiskussion auf der *Computer-Human-Interaction (CHI)* Konferenz stellte kürzlich die Frage: Zu welchem Grad sollen Maschinen versuchen Menschen in Echtzeit zu verstehen, im Gegensatz zum rein fixen Verständnis, das der Mensch bei ihrer Entwicklung in ihr Design einfließen lässt? [Fa17]

Diese Dissertation gibt eine klare Antwort: Durch verhaltenssensitive Nutzerschnittstellen können *Erwartungen* an das Nutzerverhalten direkt in interaktive Systeme eingebettet werden, um Nutzerverhalten zu antizipieren und darauf zu reagieren. Die Vorteile gegenüber einem fixen Design vorab betreffen sowohl gesteigerte Effizienz, Effektivität und Ausdrucksstärke alltäglicher Bedienkonzepte, als auch die Förderung von Privatsphäre und Sicherheit persönlicher Daten und Geräte.

Diese Perspektive der eingebetteten Erwartungen weist über den Fachbereich der Mensch-Maschine-Interaktion hinaus: Zum Beispiel beschreiben Arbeiten zu “*Predictive Processing*” [Cl16] wie (menschliche) Wahrnehmung mit der *Vorhersage* der eigenen Sinneseingaben beginnt. Lernen geschieht demnach durch Vergleich mit tatsächlichen Eindrücken, um zukünftige Reaktionen zu informieren. Diese Dissertation greift diesen Gedanken im Kontext von Nutzerschnittstellen auf: Verhaltenssensitive Nutzerschnittstellen, wie hier entwickelt, nutzen ebenso Modelle von erwarteten Eingabecharakteristika, die durch tatsächliche Nutzereingaben verfeinert werden können und Reaktionen der Nutzerschnitt-

stelle informieren. Das vorgestellte *ProbUI* Konzept zeigt diese Verbindung am deutlichsten [BA17]. Insgesamt können solche Systeme insbesondere auf individuelle menschliche Verhaltensweisen und variierende Nutzungskontexte eingehen.

Schließlich bieten verhaltenssensitive Nutzerschnittstellen einen Ansatz um die wachsenden Möglichkeiten maschinellen Lernens und künstlicher Intelligenz in einer kollaborativen Rolle in zukünftige interaktive Systeme einzubetten: Durch verhaltenssensitive Nutzerschnittstellen wird der Mensch nicht *ersetzt*, sondern durch die "Intelligenz" des Systems in der Bedienung *unterstützt*. Verhaltenssensitive Nutzerschnittstellen tragen in diesem Sinne dazu bei, dass viele Menschen von den Vorteilen solcher Systeme im Alltag und Berufsleben profitieren können, und auch auf deren diverse Fähigkeiten, Erfahrungen und Vorlieben bei der Bedienung individuell vom System eingegangen werden kann.

## Literaturverzeichnis

[BA15] Buschek, Daniel; Alt, Florian: TouchML: A Machine Learning Toolkit for Modelling Spatial Touch Targeting Behaviour. In: Proceedings of the 20th International Conference on Intelligent User Interfaces. IUI '15, ACM, New York, NY, USA, S. 110–114, 2015.

[BA17] Buschek, Daniel; Alt, Florian: ProbUI: Generalising Touch Target Representations to Enable Declarative Gesture Definition for Probabilistic GUIs. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '17, ACM, New York, NY, USA, S. 4640–4653, 2017.

[BAA15] Buschek, Daniel; Auch, Alexander; Alt, Florian: A Toolkit for Analysis and Prediction of Touch Targeting Behaviour on Mobile Websites. In: Proceedings of the 7th ACM SIGCHI Symposium on Engineering Interactive Computing Systems. EICS '15, ACM, New York, NY, USA, S. 54–63, 2015.

[BBA18] Buschek, Daniel; Bisinger, Benjamin; Alt, Florian: ResearchIME: A Mobile Keyboard Application for Studying Free Typing Behaviour in the Wild. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '18, ACM, New York, NY, USA, 2018.

[BDLA15a] Buschek, Daniel; De Luca, Alexander; Alt, Florian: Improving Accuracy, Applicability and Usability of Keystroke Biometrics on Mobile Touchscreen Devices. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '15, ACM, New York, NY, USA, S. 1393–1402, 2015.

[BDLA15b] Buschek, Daniel; De Luca, Alexander; Alt, Florian: There is More to Typing Than Speed: Expressive Mobile Touch Keyboards via Dynamic Font Personalisation. In: Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services. MobileHCI '15, ACM, New York, NY, USA, S. 125–130, 2015.

[BDLA16] Buschek, Daniel; De Luca, Alexander; Alt, Florian: Evaluating the Influence of Targets and Hand Postures on Touch-based Behavioural Biometrics. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '16, ACM, New York, NY, USA, S. 1349–1361, 2016.

[Be88] Belk, Russell W.: Possessions and the Extended Self. Journal of Consumer Research, 15(2):139, 1988.

[Be13] Belk, Russell W.: Extended Self in a Digital World. Journal of Consumer Research, 40(3):477–500, 2013.

[BKA17] Buschek, Daniel; Kinshofer, Julia; Alt, Florian: A Comparative Evaluation of Spatial Targeting Behaviour Patterns for Finger and Stylus Tapping on Mobile Touchscreen Devices. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., 1(4), 2017.

[BRMS13] Buschek, Daniel; Rogers, Simon; Murray-Smith, Roderick: User-specific Touch Models in a Cross-device Context. In: Proceedings of the 15th International Conference on Human-Computer Interaction with Mobile Devices and Services. MobileHCI '13, ACM, New York, NY, USA, S. 382–391, 2013.

[Bu16] Buschek, Daniel: There is more to biometrics than user identification: Making mobile interactions personal, secure and representative. it - Information Technology, 58(5):247–253, 2016.

[Bu18] Buschek, Daniel: A Model for Detecting and Locating Behaviour Changes in Mobile Touch Targeting Sequences. In: Proceedings of the 23rd International Conference on Intelligent User Interfaces. IUI '18, ACM, New York, NY, USA, Marz 2018.

[Cl16] Clark, Andy: Surfing Uncertainty: Prediction, Action, and the Embodied Mind. Oxford University Press, 2016.

[Fa17] Farooq, Umer; Grudin, Jonathan; Shneiderman, Ben; Maes, Pattie; Ren, Xiangshi: Human Computer Integration versus Powerful Tools. In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems. ACM Press, 2017.

[Kh17] Khamis, Mohamed; Buschek, Daniel; Thieron, Tobias; Alt, Florian; Bulling, Andreas: EyePACT: Eye-Based Parallax Correction on Touch-Enabled Interactive Displays. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., 1(4), Dezember 2017.

[We12] Weir, Daryl; Rogers, Simon; Murray-Smith, Roderick; Löchtefeld, Markus: A User-specific Machine Learning Approach for Improving Touch Accuracy on Mobile Devices. In: Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology. UIST '12, ACM, New York, NY, USA, S. 465–476, 2012.

**Daniel Buschek** ist wissenschaftlicher Mitarbeiter (Post-doc) an der Ludwig-Maximilians-Universität München (LMU). Seine Forschungsinteressen liegen an der Schnittstelle von Mensch-Maschine-Interaktion und Maschinellem Lernen / Künstlicher Intelligenz. Insbesondere interessiert er sich dafür wie nutzer- und kontextspezifische Verhaltensmuster genutzt werden können um sichere, personalisierte, verständliche und ausdrucksstarke Interaktion mit intelligenten Systemen zu ermöglichen. Seine Promotion an der LMU München (Abschluss 2018) beschäftigte sich mit diesen Themen im Kontext mobiler Geräte mit Touchscreens. Während seines Studiums und der Promotion arbeitete er zudem an Forschungsprojekten bei Siemens München, der University of Glasgow, und dem Helsinki Institute for Information Technology / Aalto University. Seine Forschung wurde unter anderem publiziert in CHI, MobileHCI, IUI, TOCHI und IMWUT, und erhielt dort mehrere Auszeichnungen.

# Die Analyse kleiner Moleküle mittels Methoden der Kombinatorik und des maschinellen Lernens[1]

Kai Dührkop[2]

**Abstract:** Massenspektrometrie ist eine Technik für die Analyse kleiner Moleküle im Hochdurchsatz. Aber sie liefert nur Informationen über die Masse der gemessenen Moleküle und, mittels Tandem-Massenspektrometrie, über die Massen der gemessenen Fragmente. Die automatisierte Auswertung von Massenspektren beschränkt sich oft auf die Suche in Spektrendatenbanken, so dass nur Moleküle derepliziert werden können, die bereits in einer solchen Datenbank gemessen wurden. In dieser Dissertation präsentieren wir zwei Methoden zur Beantwortung zweier zentraler Fragen: Was ist die Molekülformel eines gemessenen Ions? Und was ist seine Molekülstruktur? SIRIUS ist eine Methode der kombinatorischen Optimierung für die Annotation von Massenspektren und der Identifikation der Molekülformel. Dazu berechnet sie hypothetische Fragmentierungsbäume. Wir stellen ein neues Scoring Modell für die Berechnung von Fragmentierungsbäumen vor, welches die kombinatorische Optimierung als einen Maximum-a-posteriori-Schätzer auffasst. Dieses Modell ermöglicht es uns, Parameter und Hyperparameter des Scorings direkt aus den Daten abzuschätzen. Wir zeigen, dass dieses statistische Modell, dessen (Hyper)Parameter auf einem kleinen Datensatz geschätzt wurden, allgemeingültig für viele Datensätze und sogar für verschiedene Massenspektrometriegeräte ist. Wir stellen außerdem CSI:FingerID vor, eine Methode, die Kernel Support Vector Maschinen zur Vorhersage von molekularen Fingerabdrücken aus Tandem-Massenspektren nutzt. Diese vorhergesagten Fingerabdrücke können in Strukturdatenbanken gesucht werden. Dies ermöglicht erstmals die Aufklärung von Molekülstrukturen mittels Massenspektrometrie, ohne dabei auf den Abgleich von Massenspektren mittels Datenbanksuche angewiesen zu sein. Beide Methoden, SIRIUS und CSI:FingerID, sind als Kommandozeilenprogramm und als Benutzeroberfläche verfügbar. Die Vorhersage molekularer Fingerabdrücke ist als Webservice implementiert, der über eine Millionen Anfragen pro Monat erhält.

## 1 Einführung: Metabolomik und Massenspektrometrie

Die Bioinformatik beschäftigt sich mit der automatisierten Auswertung biologischer Daten. Im Fokus steht dabei üblicherweise die Genetik. Wir wissen heute, dass der Mensch mehr ist als nur die Summe seiner Gene und das viele Einflüsse, teils epigenetisch, teils umweltbedingt, auf uns und die Regulation unserer Gene einwirken. Unser Verhalten und unser Gesundheitszustand hängt nicht nur von den Genen und Proteinen, sondern von einer Vielzahl kleiner Moleküle ab. Diese werden mit der Nahrung, über die Haut, über Medikamente oder die Luft aufgenommen. Manche Stoffe werden sogar von Bakterien in unserem Darm erzeugt und wirken unmittelbar auf unseren Stoffwechsel [Co17]. Es wird geschätzt, dass jeder Mensch im Laufe seines Lebens zwei bis drei Millionen verschiedener kleiner Moleküle aufnimmt [IG07], und dass 80 bis 85 % aller Krankheiten mit kleinen

[1] Computational Methods for Small Molecule Identification

[2] Lehrstuhl Bioinformatik, Friedrich-Schiller-Universität, kai.duehrkop@uni-jena.de

Molekülen in Verbindung stehen [Up16]. Wie aktuell das Thema *kleine Moleküle* ist, zeigen die Diskussionen um Stickoxid und Feinstaub, oder um die krebserregende Wirkung des Herbizids Glyphosat.

Doch Moleküle sind nicht nur Auslöser von Krankheiten, sie können auch Krankheiten heilen. Insbesondere in Bakterien, Pilzen und Pflanzen dienen kleine Moleküle oft zur Verteidigung gegen Parasiten, Nahrungskonkurrenten und Krankheitserregern. Der bekannteste natürliche Wirkstoff ist das Antibiotikum Penicillin, der in bestimmten Schimmelpilzen gebildet wird, und dessen Entdeckung zu den bedeutendsten Entwicklungen der Medizin gehört. Seit 1970 wurden nur noch wenige neue Antibiotika entdeckt, weswegen die Weltgesundheitsorganisation 2014 eine Warnung über die *Post-antibiotische Ära* herausgab [Wo14]: Immer mehr Bakterien entwickeln Resistenzen gegen die bekannten Antibiotika. Dies macht die Suche nach neuen Antibiotika äußerst wichtig. Dabei suchen Forscher für gewöhnlich in Biomen, die noch wenig erforscht sind, wie zum Beispiel am Meeresboden [HF10, Am10]. Doch wie lassen sich potenzielle neue Antibiotika unter den tausenden von Molekülen, die in einer solchen Bodenprobe enthalten sind, aufspüren?

Mit einer ähnlichen Frage beschäftigt sich auch die Umweltforschung. Welche und wie viele potenziell giftige oder krebserregende Stoffe gelangen, zum Beispiel über das Trinkwasser, in unseren Körper? Wie weist man Stoffe, wie zum Beispiel Glyphosat, im Trinkwasser nach? Die wichtigste analytische Technik zum Nachweis kleiner Moleküle ist die Massenspektrometrie (MS). Man kann sich ein Massenspektrometer wie eine molekulare Waage vorstellen. Moleküle werden ionisiert (also mit einer Ladung versehen) und dann in einem elektrischen Feld beschleunigt. Dabei trennen sich schwere und leichte Moleküle auf, so dass die Masse (oder genauer, das Masse-Ladungs-Verhältnis m/z) jedes Moleküls bestimmt werden kann. Die Ausgabe ist ein MS Spektrum, bestehend aus einer Vielzahl von Peaks. Jeder Peak ist ein m/z Wert mit einer bestimmten Intensität, welche mit der Anzahl der gemessenen Ionen mit diesem Masse-Ladungs-Verhältnis korreliert. Hier liegt aber auch das Problem: Ein Massenspektrometer kann nur die Masse eines Moleküls bestimmen, nicht aber seine Zusammensetzung. Das bedeutet, dass man zwischen Molekülen mit gleicher Masse aber unterschiedlicher Struktur nicht unterscheiden kann.

Dieses Problem lässt sich mittels Tandem-Massenspektrometrie (MS/MS) lösen. Dabei wird zuerst ein MS gemessen und dann alle Ionen mit einem gewünschten m/z selektiert. Diese Ionen werden dann in einer Kollisionskammer in kleinere Bruchstücke fragmentiert. Diese Bruchstücke werden daraufhin in einem zweiten MS gemessen und ergeben das MS/MS Spektrum. Kleine Moleküle nehmen während der Ionisierung üblicherweise nur einen einzelnen Ladungsträger auf. Zerfällt ein solches einfach geladenes Ion in Bruchstücke, so kann nur ein Bruchstück den Ladungsträger enthalten, während die anderen Bruchstücke üblicherweise ungeladen sind. Da ein MS nur geladene Teile messen kann, gehen die ungeladenen Bruchstücke im Messvorgang verloren. Sie werden daher *Neutralverluste* genannt. Nur die geladenen Bruchstücke, *Fragmente* genannt, werden im zweiten MS gemessen.

Ein weiteres Problem ist, dass die genaue Fragmentierungsweise eines Moleküls nicht im Voraus bekannt ist. Überraschenderweise ist diese Fragestellung ein enorm schweres Problem und bis heute gibt es keine Methode die mit Sicherheit die Fragmentierung von Molekülen vorhersagen kann. Selbst quantenchemische Berechnungen liefern nur ungenaue

Ergebnisse [Sp18]. Stattdessen werden Datenbanken von MS/MS Messungen bekannter Moleküle angelegt. Durch die Suche in solch einer Datenbank lassen sich Stoffe identifizieren - aber eben nur solche, die zuvor gemessen und in einer Datenbank abgespeichert wurden. Die Folge ist, dass nur ein winziger Bruchteil der Moleküle, die in einer Probe gemessen werden, auch einem bekannten Stoff zugeordnet werden können [dDQ15, Up16].

Man unterteilt Experimente in der Massenspektrometrie grob in zwei Kategorien: Der gerichteten und ungerichteten Analyse. Bei der gerichteten sucht man gezielt nach bestimmten Stoffen in einer Probe (Beispiel: die Suche nach Drogen oder Dopingmitteln). Bei der ungerichteten Analyse versucht man alle Stoffe in einer Probe aufzuklären. Bei biologischen Proben spricht man dann auch von der Metabolomik, die, analog zur Genomik in der Genetik, die Gesamtheit aller Metabolite (also kleinen Moleküle) in einer biologischen Probe untersuchen will. Wie bereits erwähnt, arbeitet die Metabolomik heute weitestgehend mit Datenbanksuchen. Das heißt aber eben auch, dass nur Moleküle identifiziert werden können, die bereits bekannt sind und gemessen wurden. Für neue biologische Erkenntnisse braucht man jedoch Methoden, die auch völlig neue Stoffe identifizieren können. Hierfür sind in den letzten Jahren verschiedene Methoden entwickelt wurden, die üblicherweise nur noch eine Liste an Kandidatenstrukturen benötigen, aber nicht mehr eine Datenbank von tatsächlichen Messungen dieser Moleküle [HSB14, Bl18]. Im Laufe meiner Doktorarbeit haben wir an der Entwicklung zweier solcher Methoden gearbeitet: SIRIUS [BD16] und CSI:FingerID [Du15]. Dabei werden Techniken der kombinatorischen Optimierung, der Bayesschen Statistik und des maschinellen Lernens miteinander vereint. Zusammen ermöglichen beide Methoden die Strukturaufklärung kleiner Moleküle. In verschiedenen unabhängigen Evaluationen konnten sie alle anderen Methoden zur Strukturaufklärung mittels Massenspektrometrie deutlich schlagen [Du19]. section*Analyse der Summenformel mittels SIRIUS Die Masse eines Moleküls hängt nur von seiner Zusammensetzung, also den Atomen und ihren Elementen, nicht aber von der Anordnung der Atome oder den Bindungen ab. Eine solche Zusammensetzung kann in Form einer Summenformel dargestellt werden. So beschreibt die Summenformel $C_6H_{12}O_6$ ein Molekül mit 6 Kohlenstoff-, 12 Wasserstoff- und 6 Sauerstoffatomen. Der erste Schritt der Strukturaufklärung eines Moleküls liegt in der Bestimmung seiner Summenformel. Überraschenderweise ist bereits das ein schweres Problem: Obgleich moderne MS Geräte die Masse eines kleinen Moleküls auf die fünfte Nachkommastelle genau bestimmen können, gibt es doch, für große Massen, oft hunderte bis tausende Summenformeln die annähernd die gleiche Masse haben. Zur Unterscheidung dieser Summenformeln reicht die Masse allein nicht aus. Mit SIRIUS haben wir eine Software entwickelt, welche die Summenformel des Ions bestimmt. Dabei werden zwei Analyseverfahren kombiniert: Die Isotopenmusteranalyse und die Fragmentierungsanalyse.

Isotopen sind Atome mit gleicher Protonen- und Elektronenzahl (und damit von gleichem Element), aber unterschiedlicher Neutronenzahl. Sie unterscheiden sich also in ihrer Masse. Beispielsweise ist im Schnitt jedes 100te Kohlenstoffatom ein $^{13}C$ Isotop, während die anderen Kohlenstoffatome üblicherweise vom Typ $^{12}C$ sind. Entsprechend kann man jeder Summenformel nicht eine einzelne, sondern eine Verteilung von Massen zuordnen. Diese Verteilung lässt sich für eine Summenformel berechnen und mit dem gemessenen Isotopenmuster vergleichen. Wir haben ein Maximum Likelihood Verfahren entwickelt, wel-

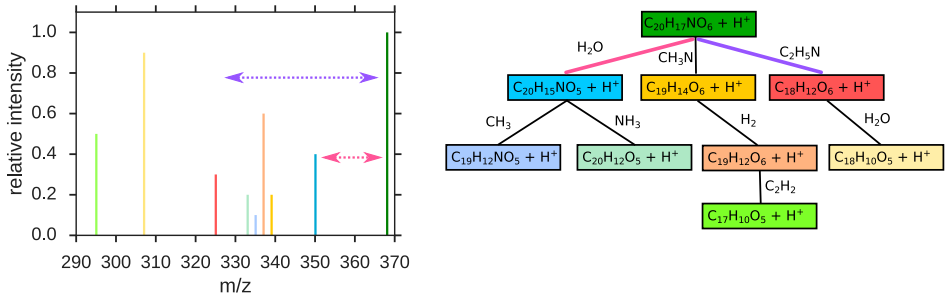



Abb. 1: Beispiel eines Fragmentierungsbaums (rechts) und dem zugehörigen MS/MS Spektrum (links) des Moleküls Bicucullin. Jeder Knoten im Baum ist die Erklärung eines Peaks im Spektrum. Jede Kante erklärt eine Fragmentierungsreaktion zwischen zwei Peaks im Spektrum.

ches einem Isotopenmuster die wahrscheinlichste Summenformel zuordnet [Bo09, Du19]. Isotopenmuster sind auch relevant, wenn man seltene Elemente wie Chlor oder Brom in einer Probe nachweisen will [Me16]. Dazu haben wir ein tiefes neuronales Netz entwickelt, welches solche seltenen Elemente aus einem Isotopenmuster vorhersagt [Du19].

Die Fragmentierungsanalyse berechnet Fragmentierungsbäume zur Aufklärung der Summenformel. Ein Fragmentierungsbaum beschreibt den Prozess der Fragmentierung als Baum, dessen Knoten die Summenformeln der Fragmente und dessen Kanten die Neutralverluste sind (Abbildung 1). Die Wurzel des Baums ist zugleich die Summenformel des gemessenen Ions. Zur Berechnung eines Fragmentierungsbaumes werden für alle Peaks im MS/MS Spektrum alle möglichen Summenformeln enumeriert. Dies ist über dynamische Programmierung effizient möglich [Bo08, Du13]. Danach wird ein gerichteter Fragmentierungsgraph erzeugt, dessen Knoten alle Summenformeln sind und der eine Kante zwischen je zwei Summenformeln enthält, die einander in einer Teilmengenbeziehung stehen. Jeder induzierte Teilbaum dieses Graphen ist eine mögliche Interpretation des Spektrums. Damit jedem Peak maximal eine Summenformel zugeordnet wird, färbt man den Graphen, so dass alle Knoten des gleichen Peaks die gleiche Farbe haben. Nun fordert man einen farbenfrohen Teilbaum, dessen Kanten- und Knotengewichte maximal sind. Als Gewichte wird ein beliebiges Scoring verwendet, welches wahrscheinlichen bzw. plausiblen Knoten und Kanten positive Gewichte, sowie den Unwahrscheinlichen negative Gewichte zuordnet. Das resultierende Problem „Maximaler Farbenfroher Teilbaum“ ist NP-schwer, kann aber in der Praxis mittels Integer Linearer Programmierung effizient gelöst werden [Ra13].

Die Fragmentierungsbaumberechnung geht auf die Arbeit von Rasche et al. [Ra11] zurück. Allerdings war das Scoring für die Kanten eher ad hoc gewählt und nur teilweise statistisch interpretierbar. Während meiner Doktorarbeit haben wir dieses algorithmische Problem in ein Maximum A Posteriori Problem umformuliert. Die Kantengewichte werden dabei zu *A Priori* Wahrscheinlichkeiten, die Knotengewichte zu *Likelihoods*. Dadurch bekommen die Kantengewichte eine statistische Interpretation. Darüber hinaus lassen sich diese Wahrscheinlichkeiten direkt aus echten Daten abschätzen: Für die A Priori Wahrscheinlichkeiten haben wir Wahrscheinlichkeitsverteilungen abgeschätzt und mittels Maximum

Likelihood Schätzern an gemessene Daten angepasst. Für die Likelihoods haben wir die Messfehler des MS Gerätes modelliert, sowie die Verteilung des Hintergrundrauschens.

Das resultierende neue Scoring wurde auf verschiedenen unabhängigen Datensätzen evaluiert. Die Identifikationsrate, also der Anteil korrekt bestimmter Summenformeln, stieg mit dem neuen Scoring um das doppelte auf 73,8 % an und schlägt damit nicht nur die vorherige Fragmentierungsbaumanalyse, sondern mit großem Abstand auch alle alternativen Methoden zur Summenformelbestimmung. Zusammen mit der Isotopenmusteranalyse erreicht SIRIUS Identifikationsraten von 86,36 %, 93,30 % und 93,75 % auf drei verschiedenen, unabhängigen Datensätzen. Neben dem neuen Scoring haben wur auch an Heuristiken und algorithmischen und technischen Optimierungen gearbeitet, was in einer 200fachen Beschleunigung der Fragmentierungsbaumberechnung resultierte [Du18, Wh15, Du19].

**Strukturaufklärung mittels CSI:FingerID**

Der zweite Teil meiner Doktorarbeit beschäftigt sich mit der Vorhersage von Molekülstrukturen mittels überwachtem maschinellem Lernen. Dabei stellt sich das Problem, dass sowohl MS Spektren als auch Moleküle keine einfachen numerischen Vektoren sind, wie sie normalerweise beim maschinellen Lernen als Eingabe und Ausgabe verwendet werden. Der erste Schritt für eine Strukturvorhersage aus MS Spektren liegt also darin, Spektren und Moleküle als Vektoren zu kodieren. Hier haben Markus Heinonen et al. Pionierarbeit geleistet [He12]: Sie kodieren Moleküle als binäre Vektoren, in denen jede Position für eine bestimmte Teilstruktur oder funktionelle Gruppe steht. Im Vektor steht eine 1, wenn das Molekül diese Teilstruktur enthält, ansonsten eine 0. Diese Form der Kodierung bezeichnet man auch als *Molekularen Fingerprint*. Da der Vektor binär ist, lässt sich die Strukturvorhersage als Menge von Klassifizierungsproblemen beschreiben: Für jede Teilstruktur wird ein Prediktor trainiert, der aus einem MS Spektrum das Vorhandensein dieser Struktur vorhersagt [He12]. Support Vektor Maschinen sind ein effizientes Verfahren zur Klassifizierung. Ihr Vorteil ist, dass sie nicht zwingend numerische Vektoren als Eingabe benötigen, sondern alternativ einen sogenannten Kernel: Dies ist eine Funktion, die das innere Produkt der Eingabevektoren berechnet. Beispielsweise muss das Spektrum nicht als Vektor kodiert werden, solange die Ähnlichkeit zweier Spektren direkt berechnet werden kann.

Auf diesen Grundlagen von Heinonen et al haben wir die Methode CSI:FingerID entwickelt [Sh14, Du15]. Statt nur das MS/MS Spektrum, verwenden wir Fragmentierungsbäume als Eingabe. Entsprechend haben wir Ähnlichkeitsfunktionen auf Fragmentierungsbäumen entwickelt. Beispielsweise eine Ähnlichkeitsfunktion, welche die Anzahl gemeinsamer Fragmente oder Neutralverluste zwischen zwei Bäumen zählt. Mittels dynamischer Programmierung lassen sich auch die Zahl gemeinsamer Pfade oder Teilbäume zählen. Besonders effektiv sind Ähnlichkeitsfunktionen, welche Eigenschaften auf den Molekülformeln bewerten, die aus der Masse des Peaks nicht ablesbar sind. Beispielsweise haben wir einen Kernel definiert, der aus der Summenformel eines Fragments die Zahl der kovalenten Atombindungen abschätzt. Ein solcher Kernel ist beispielsweise in der Lage, langkettige Moleküle von Molekülen mit vielen Ringen zu unterscheiden. Andere Ähnlichkeitsfunktionen suchen nach Teilsummenformeln, die zwei Fragmente gemeinsam haben, und die auf

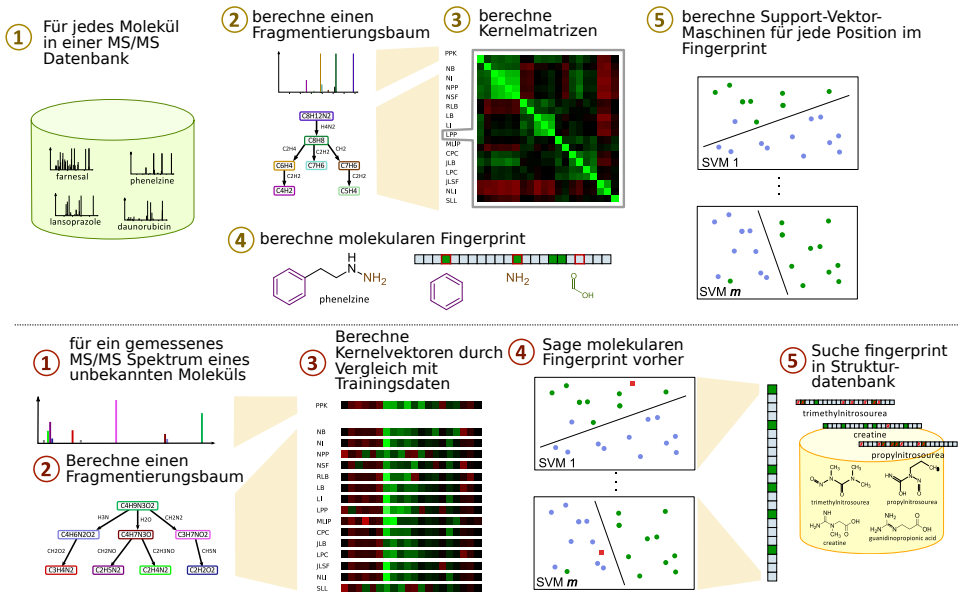



Abb. 2: Schematische Darstellung der Arbeitsweise von CSI:FingerID. In der Trainingsphase (Oben) wird das Maschinelle Lernverfahren auf einer Datenbank von Referenzspektren trainiert. In der Vorhersagephase (Unten) wird für ein unbekanntes Molekül dessen molekularer Fingerprint vorhergesagt und in einer Strukturdatenbank gesucht.

bestimmte Teilstrukturen hinweisen: So deutet das Vorhandensein der Summenformel $H_3PO_4$ auf eine Phosphatsäure im Molekül hin. Statt nur einer einzigen Ähnlichkeitsfunktion, definieren wir eine ganze Reihe von Kernels. Die 17 besten Kernels haben wir dann mittels multiplem Kernellernen kombiniert [Sh14]: Dabei werden die Kernels auf den Bäumen so linear kombiniert, dass die resultierende Funktion möglichst gut der Ähnlichkeitsfunktion auf den molekularen Fingerprints entspricht. Der resultierende Kernel hat eine deutlich verbesserte Qualität gegenüber dem ursprünglichen Kernel von Heinonen et al. Vor allem aber generalisiert er besser, wie sich auf unabhängigen Datensätzen gezeigt hat.

Um ein Molekül zu identifizieren, bestimmen wir erst seinen molekularen Fingerprint: Dabei kommen 2 937 Supportvektormaschinen zum Einsatz, die für jede Substruktur aus dem MS Spektrum und dem zugehörigen Fragmentierungsbaum die Wahrscheinlichkeit vorhersagen, das diese Struktur im gemessenen Molekül enthalten ist. Dieser vorhergesagte molekulare Fingerprint wird dann mittels Maximum Likelihood in einer Strukturdatenbank von Fingerprints gesucht. Siehe Abbildung 2 für den schematischen Aufbau von CSI:FingerID.

Wir haben CSI:FingerID auf verschiedenen Datensätzen evaluiert. Es ist konsistent und mit großem Abstand die beste Methode zur Identifizierung von Strukturen mittels Massenspektrometrie. In einer Evaluation mit Kreuzvalidierung konnte CSI:FingerID 3,3 Mal so viele Moleküle identifizieren wie die zweitbeste Methode. Die Identifikationsrate stieg

von 12,12 % auf 40,37 % an. Die Methode (zusammen mit ihrer Input-Output-Kernel-Regression Variante IOKR [Br16]) wurde auch zwei Mal blind in einem Wettbewerb, dem Critical Assessment of Small Molecule Identification (CASMI), evaluiert [Sc17]. In CASMI 2016 konnte die Methode 1,5 Mal so viele Moleküle wie die zweitbeste Methode identifizieren. In CASMI 2017 war der Abstand sogar noch größer - CSI:FingerID hatte 4,8 Mal so viele korrekte Identifikationen wie die nächstbeste Methode.

## Fazit und Ausblick

Wir haben eine Kommandozeilenanwendung und grafische Nutzerschnittstelle für SIRIUS und CSI:FingerID implementiert. Die Integration mit CSI:FingerID geschieht über eine REST-Schnittstelle, so dass die Supportvektormaschinen auf unserem Server verbleiben und regelmäßig auf neuen Referenzdaten trainiert werden können. Unsere Methode wird mittlerweile weltweit eingesetzt und es werden Millionen Anfragen pro Monat an den CSI:FingerID Server gestellt. Verschiedene wichtige Massenspektrometrie Toolkits wie OpenMS [Rö16] und MZmine [Pl10], sowie die Global Natural Products Social Molecular Networking (GNPS) Datenbank [Wa16] integrieren SIRIUS mittlerweile in ihre Workflows. Auch wird unsere Methode zur Analyse kleiner Moleküle bereits von mehreren Pharmafirmen eingesetzt. CSI:FingerID ist in der Lage 40 % der MS/MS Spektren korrekt zu identifizieren. Schränkt man die Suche auf eine Datenbank biologischer Moleküle ein, steigt die Identifikationsrate sogar auf 75 % an. Dies macht eine qualitative Analyse im Hochdurchsatz überhaupt erst möglich.

Aufbauend auf SIRIUS und CSI:FingerID, entwickeln wir nun auch weitere Methoden, die nicht nur die Struktur, sondern auch abgeleitete Eigenschaften von Molekülen vorhersagen: Beispielsweise, ob ein Molekül zu einer bestimmten Klasse von Naturstoffen gehört. Oder ob ein gemessenes Molekül, ausgehend nur von seinem Massenspektrum, einem bekannten Wirkstoff oder Medikament ähnlich ist. Dies soll die Auswertung und Interpretation metabolomischer Daten sowie die Suche nach neuen Wirkstoffen vereinfachen und beschleunigen.

## Literaturverzeichnis

[Am10] Aminov, Rustam: A brief history of the antibiotic era: lessons learned and challenges for the future. Front Microbiol, 1:134, 2010.

[BD16] Böcker, Sebastian; Duehrkop, Kai: Fragmentation trees reloaded. J Cheminform, 8:5, 2016.

[Bl18] Blaženović, Ivana; Kind, Tobias; Ji, Jian; Fiehn, Oliver: Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics. Metabolites, 8(2), 2018.

[Bo08] Böcker, Sebastian; Lipták, Zsuzsanna; Martin, Marcel; Pervukhin, Anton; Sudek, Henner: DECOMP–from interpreting Mass Spectrometry peaks to solving the Money Changing Problem. Bioinformatics, 24(4):591–593, 2008.

[Bo09] Böcker, Sebastian; Letzel, Matthias; Lipták, Zsuzsanna; Pervukhin, Anton: SIRIUS: Decomposing isotope patterns for metabolite identification. Bioinformatics, 25(2):218–224, 2009.

[Br16] Brouard, Céline; Shen, Huibin; Dührkop, Kai; d'Alché-Buc, Florence; Böcker, Sebastian; Rousu, Juho: Fast metabolite identification with Input Output Kernel Regression. Bioinformatics, 32(12):i28–i36, 2016. Proc. of *Intelligent Systems for Molecular Biology* (ISMB 2016).

[Co17] Cohen, Louis J; Esterhazy, Daria; Kim, Seong-Hwan; Lemetre, Christophe; Aguilar, Rhiannon R; Gordon, Emma A; Pickard, Amanda J; Cross, Justin R; Emiliano, Ana B; Han, Sun M; Chu, John; Vila-Farres, Xavier; Kaplitt, Jeremy; Rogoz, Aneta; Calle, Paula Y; Hunter, Craig; Bitok, J Kipchirchir; Brady, Sean F: Commensal bacteria make GPCR ligands that mimic human signalling molecules. Nature, 549(7670):48–53, 2017.

[dDQ15] da Silva, Ricardo R.; Dorrestein, Pieter C.; Quinn, Robert A.: Illuminating the dark matter in metabolomics. Proc Natl Acad Sci U S A, 112(41):12549–12550, 2015.

[Du13] Dührkop, Kai; Ludwig, Marcus; Meusel, Marvin; Böcker, Sebastian: Faster mass decomposition. In: Proc. of Workshop on Algorithms in Bioinformatics (WABI 2013). Jgg. 8126 in Lect Notes Comput Sci. Springer, Berlin, S. 45–58, 2013.

[Du15] Duehrkop, Kai; Shen, Huibin; Meusel, Marvin; Rousu, Juho; Böcker, Sebastian: Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proc Natl Acad Sci U S A, 112(41):12580–12585, 2015.

[Du18] Duehrkop, Kai; Lataretu, Marie A.; White, W. Timothy J.; Böcker, Sebastian: Heuristic algorithms for the Maximum Colorful Subtree problem. In: Proc. of Workshop on Algorithms in Bioinformatics (WABI 2018). Jgg. 113 in Leibniz International Proceedings in Informatics (LIPIcs), Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, S. 23:1–23:14, 2018.

[Du19] Duehrkop, Kai; Fleischauer, Markus; Ludwig, Marcus; Aksenov, Alexander A.; Melnik, Alexey V.; Meusel, Marvin; Dorrestein, Pieter C.; Rousu, Juho; Böcker, Sebastian: SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat Methods, Marz 2019.

[He12] Heinonen, Markus; Shen, Huibin; Zamboni, Nicola; Rousu, Juho: Metabolite identification and molecular fingerprint prediction via machine learning. Bioinformatics, 28(18):2333–2341, 2012.

[HF10] Hughes, Chambers C.; Fenical, William: Antibacterials from the sea. Chem Eur J, 16(42):12512–12525, 2010.

[HSB14] Hufsky, Franziska; Scheubert, Kerstin; Böcker, Sebastian: New kids on the block: Novel informatics methods for natural product discovery. Nat Prod Rep, 31(6):807–817, 2014.

[IG07] Idle, Jeffrey R.; Gonzalez, Frank J.: Metabolomics. Cell Metab, 6(5):348–351, 2007.

[Me16] Meusel, Marvin; Hufsky, Franziska; Panter, Fabian; Krug, Daniel; Mueller, Rolf; Böcker, Sebastian: Predicting the presence of uncommon elements in unknown biomolecules from isotope patterns. Anal Chem, 88(15):7556–7566, 2016.

[Pl10] Pluskal, Tomáš; Castillo, Sandra; Villar-Briones, Alejandro; Oresic, Matej: MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinf, 11:395, 2010.

[Ra11] Rasche, Florian; Svatoš, Aleš; Maddula, Ravi Kumar; Boettcher, Christoph; Böcker, Sebastian: Computing fragmentation trees from tandem mass spectrometry data. Anal Chem, 83(4):1243–1251, 2011.

[Ra13] Rauf, Imran; Rasche, Florian; Nicolas, François; Böcker, Sebastian: Finding Maximum Colorful Subtrees in practice. J Comput Biol, 20(4):1–11, 2013.

[Rö16] Röst, Hannes L.; Sachsenberg, Timo; Aiche, Stephan; Bielow, Chris; Weisser, Hendrik; Aicheler, Fabian; Andreotti, Sandro; Ehrlich, Hans-Christian; Gutenbrunner, Petra; Kenar, Erhan; Liang, Xiao; Nahnsen, Sven; Nilse, Lars; Pfeuffer, Julianus; Rosenberger, George; Rurik, Marc; Schmitt, Uwe; Veit, Johannes; Walzer, Mathias; Wojnar, David; Wolski, Witold E.; Schilling, Oliver; Choudhary, Jyoti S.; Malmström, Lars; Aebersold, Ruedi; Reinert, Knut; Kohlbacher, Oliver: OpenMS: a flexible open-source software platform for mass spectrometry data analysis. Nat Methods, 13(9):741–748, 2016.

[Sc17] Schymanski, Emma Louise; Ruttkies, Christoph; Krauss, Martin; Brouard, Céline; Kind, Tobias; Dührkop, Kai; Allen, Felicity Ruth; Vaniya, Arpana; Verdegem, Dries; Böcker, Sebastian; Rousu, Juho; Shen, Huibin; Tsugawa, Hiroshi; Sajed, Tanvir; Fiehn, Oliver; Ghesquière, Bart; Neumann, Steffen: Critical Assessment of Small Molecule Identification 2016: Automated Methods. J Cheminf, 9:22, 2017.

[Sh14] Shen, Huibin; Duehrkop, Kai; Böcker, Sebastian; Rousu, Juho: Metabolite Identification through Multiple Kernel Learning on Fragmentation Trees. Bioinformatics, 30(12):i157–i164, 2014. Proc. of *Intelligent Systems for Molecular Biology* (ISMB 2014).

[Sp18] Spackman, Peter R.; Bohman, Bjoern; Karton, Amir; Jayatilaka, Dylan: Quantum chemical electron impact mass spectrum prediction for de novo structure elucidation: assessment against experimental reference data and comparison to competitive fragmentation modeling. Int J Quantum Chem, 118(2), 2018.

[Up16] Uppal, Karan; Walker, Douglas I.; Liu, Ken; Li, Shuzhao; Go, Young-Mi; Jones, Dean P.: Computational metabolomics: a framework for the million metabolome. Chem Res Toxicol, 29(12):1956–1975, 2016.

[Wa16] Wang, Mingxun et al.: Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol, 34(8):828–837, 2016.

[Wh15] White, W. Timothy J.; Beyer, Stephan; Duehrkop, Kai; Chimani, Markus; Böcker, Sebastian: Speedy Colorful Subtrees. In: Proc. of Computing and Combinatorics Conference (COCOON 2015). Jgg. 9198 in Lect Notes Comput Sci. Springer, Berlin, S. 310–322, 2015.

[Wo14] World Health Organization et al.: Antimicrobial resistance: global report on surveillance. World Health Organization, 2014.

**Kai Dührkop** wurde am 21.10.1988 in Rudolstadt geboren, ging bis 2007 auf das Friedrich-Froebel Gymnasium in Bad Blankenburg und studierte anschließend bis 2012 an der Friedrich-Schiller-Universität in Jena. Sein Studium schloss er mit einem Diplom in Bioinformatik ab. Die Diplomarbeit wurde zum Thema „A Sparse Dynamic Programming Algorithm for Fragmentation Tree Alignments“ verfasst. Von 2012 bis 2018 promovierte er am Lehrstuhl für Bioinformatik an der Friedrich-Schiller-Universität und schrieb seine Dissertation zum Thema „Computational Methods for Small Molecule Identification“. Seine Disputation fand am 20. September 2018 statt. Während seiner Promotion hat er 13 Publikationen verfasst, 7 davon als Erstautor. Er hielt wissenschaftliche Vorträge auf drei internationalen Konferenzen und besuchte zwei Dagstuhl- sowie ein Shonan-Seminar. Seit April 2019 arbeitet er im Rahmen eines wissenschaftlichen Austauschs an der University of California San Diego.

# Differenzielle Kryptanalyse von Symmetrischen Primitiven[1]

Maria Eichlseder[2]

**Abstract:** Symmetrische Kryptographie stellt hocheffiziente Verfahren zur Verfügung, die Vertraulichkeit und Integrität von Daten in diversen Anwendungen schützen. Die Sicherheit dieser Verfahren ruht auf deren Primitiven, wie Blockchiffren oder Permutationen. Die spezifischen Anforderungen an Funktionalität, Effizienz und Robustheit dieser Primitive sind im steten Wandel, was sich in einer regen Publikation neuer Designs und Designprinzipien niederschlägt.

In dieser Dissertation entwickeln wir Techniken zur differenziellen Kryptanalyse verschiedener Primitive. Unsere Ergebnisse zeigen dabei signifikante Schwachstellen in einigen neuen Designs auf. Zusätzlich entwickeln und evaluieren wir verbesserte Strategien in der automatisierten Kryptanalyse.

**Keywords:** Kryptographie Differenzielle Kryptanalyse Hashfunktionen Authenticated Encryption

## 1 Einleitung

IT-Systeme verarbeiten heute mehr sensible Informationen denn je, darunter private persönliche Daten und vertrauliche Geschäftsdaten. Um diese Daten zu schützen, werden kryptographische Algorithmen wie Verschlüsselungsverfahren und Hashfunktionen benötigt. Ein zentraler Faktor für die Sicherheit dieser kryptographischen Algorithmen ist die Widerstandsfähigkeit ihrer Kernbestandteile, der sogenannten Primitive wie beispielsweise Blockchiffren, gegen Angriffe wie differenzielle, lineare und algebraische Kryptanalyse. Das Ziel der Kryptanalyse ist es, unsichere Primitive zu identifizieren sowie den sogenannten „Security Margin“ (ein Maß für die Sicherheit ausgehend von den derzeit besten Angriffen) von sicheren Primitiven möglichst genau zu schätzen. Beides ist wichtig, um sichere Verfahren für die Zukunft zu gewährleisten.

Differenzielle Kryptanalyse ist dabei eine der bedeutendsten Angriffsmethoden und versucht, geheime Informationen zu extrahieren oder Nachrichten zu fälschen, indem das Verhalten der Primitive statistisch untersucht wird, wenn zwei ähnliche, aber leicht unterschiedliche Eingaben verarbeitet werden. Seit Einführung Anfang der 1990er [BS90] hat sich diese Methode als ausgesprochen effektiv und vielseitig bewährt, und entsprechende Gegenmaßnahmen gehören daher zu den Grundpfeilern beim Design von modernen Blockchiffren.

Während symmetrische Kryptographie jahrzehntelang quasi synonym zum Design und der Analyse von Blockchiffren war, sind in den letzten Jahren auch alternative Primitive ins akademische Rampenlicht gerückt: Insbesondere kryptographische Permutationen

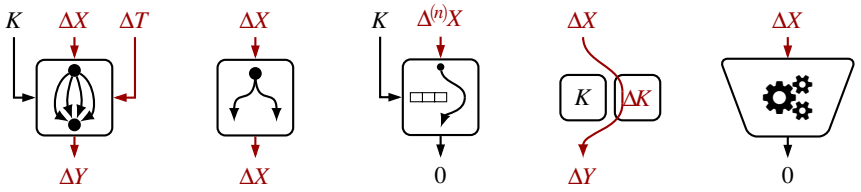
[1] Englischer Titel der Dissertation: „Differential Cryptanalysis of Symmetric Primitives“

[2] Technische Universität Graz, maria.eichlseder@iaik.tugraz.at

und tweakbare Blockchiffren (tweakable blockciphers, TBCs) machen Blockchiffren ihre Rolle als ideale Primitive für effiziente, sichere und elegante Verfahren streitig. Diese Primitive bieten dem Kryptanalysten allerdings eine veränderte Angriffsoberfläche, beispielsweise durch schlüssellose Rundenfunktionen oder durch zusätzliche kontrollierbare Tweak-Inputs in jeder Runde. Die Implikationen dieser Unterschiede sind bislang nur unzureichend untersucht und der entsprechende kryptanalytische Werkzeugkasten noch nicht ausgereift.

Als zentrale Ergebnisse der Dissertation [Ei18] zeigen wir, dass die von Blockchiffren übernommenen Design-Strategien nicht immer ausreichenden Schutz gegen differenzielle Angriffe bieten können. Wir zeigen neue Analysetechniken und potenzielle Schwachstellen auf, die bei neuen Designs berücksichtigt werden sollten. Mehrere in den letzten Jahren publizierte vielversprechende Primitive, beispielsweise die in Software besonders performante Permutation Simpira oder die leichtgewichtige TBC MANTIS, konnten auf diesem Wege geknackt werden. Abb. 1 gibt einen symbolischen Überblick über die vorgeschlagenen Analysetechniken, -szenarien, und -ziele. Zusätzlich entwickeln wir Techniken zur Verbesserung der computergestützten differenziellen Analyse von schlüssellosen Primitiven und erreichen damit die besten praktischen Kollisionsangriffe auf mehrere reduzierte Varianten der Hashfunktion SHA-2, was maßgeblich dazu beiträgt, den Security Margin dieses weltweit intensiv genutzten Standards einzuschätzen – SHA-2 kommt unter anderem im Großteil aller https-gesicherten Web-Verbindungen oder in Bitcoin zum Einsatz.

Weitere Beiträge und Ergebnisse von Kollaborationen im Rahmen des Doktorats, die nicht im Dissertationstext ausgeführt werden, betreffen die praktische Sicherheit von kryptographischen Implementierungen und Designs. Wir nutzen unter anderem Techniken aus der differenziellen Kryptanalyse, um zu zeigen, dass ein Angreifer mit speziellen physikalischen Fehlerangriffen (Statistical Ineffective Fault Attacks, SIFA: Abb. 3) durch Stören der Berechnungen geheime Informationen lernen kann – selbst wenn die Implementierung eigentlich mit den gebräuchlichen Gegenmaßnahmen gegen genau solche Angriffe ausgestattet ist. Ein bedeutendes Ergebnis ist die Mitentwicklung und Analyse des authentifizierten Verschlüsselungsverfahrens Ascon, das besonders effiziente und ressourcenschonende Implementierungen ermöglicht, die dabei robust gegen verschiedene Implementierungsangriffe sind. Ascon wurde 2019 als Gewinner der 2014 gestarteten „CAESAR Competition“ für kryptographische Designs in der Kategorie „Lightweight Applications“ ausgezeichnet.



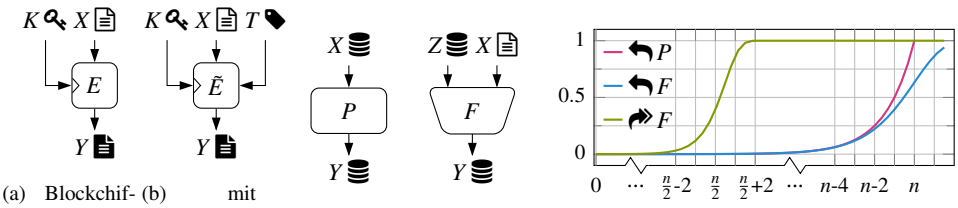

(a) MANTIS (b) Simpira (c) LowMC (d) Prøst (e) SHA-2

Abb. 1: Überblick über die entwickelten differenziellen Analysetechniken und analysierte Primitive.

## 2 Hintergrund

Kryptographische Primitive sind die kleinsten Bausteine in einem kryptographischen System, mit denen ein Sicherheitslevel assoziiert werden kann. Die darüberliegenden Schichten, wie Modes of Operation oder Protokolle, führen ihre Sicherheit jeweils – idealerweise in Form eines Reduktionsbeweises – auf diese Primitive zurück. In der asymmetrischen Kryptographie sind diese Primitive typischerweise (mehr oder weniger nahe) verwandt mit klassischen Problemen der Komplexitätstheorie, während die symmetrischen Kryptographie immer wieder neue, effiziente Circuits dazu entwickelt. Die meisten klassischen symmetrischen Primitive sind Blockchiffren, d.h. Familien von Permutationen mit einer fixen Inputgröße von beispielsweise 128 Bits, wobei ein geheimer Schlüssel $K$ eine Permutation $E_K$ aus dieser Familie auswählt, mit der dann die Klartextblöcke $X$ in Ciphertextblöcke $Y$ übersetzt werden. Der Angreifer versucht, aus beobachteten Daten $(X,Y)$ entweder den Schlüssel $K$ oder andere Informationen über die Permutation $E_K$ abzuleiten, um damit die Sicherheit der darüberliegenden Ebenen zu untergraben, wie etwa die Vertraulichkeit der Nachricht in einem authentifizierten Verschlüsselungsverfahren.

Eine limitierende Eigenschaft des Blockchiffren-Modells ist die fehlende Abbildung des Kontexts der übersetzten Daten, etwa der Adresse des Blocks innerhalb der Nachricht. Das muss auf darüberliegenden Ebenen kompensiert werden, was diverse Probleme mit sich bringt. Als Lösungsansatz bieten tweakbare Blockchiffren eine explizite zusätzliche Inputmöglichkeit, den Tweak $T$, während Permutationen mit vergrößertem Input die Grenzen zwischen Daten, Schlüssel und Kontext für die Berechnung völlig verschwimmen lassen. Intern folgen die Primitive grundsätzlich einer ähnlichen Struktur, bei der eine einfache, durch $K,T$ parametrisierte Rundenfunktion oft iteriert wird; durch die verschiedene Parametrisierung und typische Blockgrößen ergeben sich jedoch Unterschiede im konkreten Design. Abb. 2 illustriert die entstehenden Interfaces der Primitive sowie einige generische Schranken für deren Sicherheit, die sich aus diesen Interfaces ergeben.



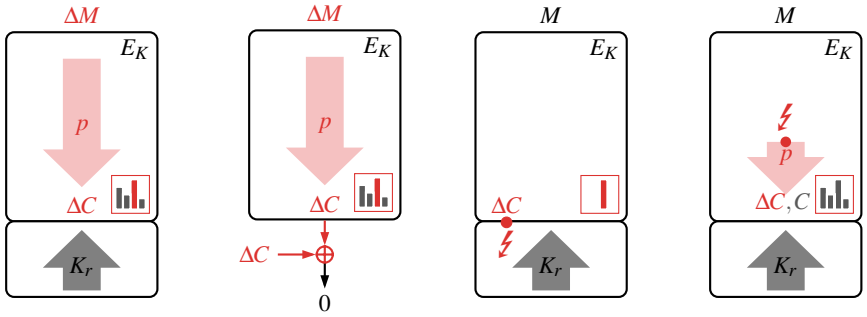

(a) Blockchiffre (b) mit Tweak (c) Permutation (d) Kompression (e) $\mathbb{P}$[Erfolg] nach $2^x$ Versuchen mit $n = |Y|$

Abb. 2: Verschiedene Typen symmetrischer Primitive sowie generische Erfolgswahrscheinlichkeit von Preimage- (↶) und Kollisionsangriffen (↷) auf Permutationen ($P$) und Funktionen ($F$).

Das Ziel differenzieller Kryptanalyse [BS90] ist, Schwachstellen in den Primitiven zu identifizieren, die effizientere Angriffe als diese generischen Schranken erlauben. Sei $f : \mathbb{F}_2^n \to \mathbb{F}_2^n, x \mapsto y$ eine vektorielle Boolesche Funktion, etwa eine Blockchiffre mit einem fixen unbekannten Schlüssel. Wir betrachten Paare von Inputvariablen $(x,x^*)$ mit einer Differenz $\Delta x = x^* \oplus x$ und interessieren uns für die Outputdifferenz $\Delta y = y^* \oplus y$, genauergesagt, für die von einer fixen Inputdifferenz $\alpha = \Delta x$ induzierte Ableitungsfunktion

$$\Delta_\alpha f(x) := f(x \oplus \alpha) \oplus f(x)\,.$$

Selbst wenn der Wert $x$ bzw. der Schlüssel $K$ mit $f = E_K$ unbekannt sind, lassen sich gegebenenfalls Aussagen über die statistische Verteilung $\mathsf{edp}(\alpha,\beta) := \mathbb{P}_{K,x}[\Delta_\alpha f_K(x) = \beta]$ von $\beta$ treffen. Dazu werden mögliche Charakteristiken $\chi$, die resultierende Differenzen nach jeder Runde in $f$ beschreiben, sowie ihre erwartete Wahrscheinlichkeit $p = \mathsf{edp}(\chi)$ untersucht. Gibt es eine Charakteristik $\chi$ und damit ein Differential $(\alpha,\beta)$ mit $-\log_2(\mathsf{edp}(\alpha,\beta)) < \min(n,|K|)$ für (fast) die volle Rundenanzahl von $f$, so lässt sich $f$ mit einer Komplexität in der Größenordnung $\mathsf{edp}(\alpha,\beta)^{-1}$ von einer zufälligen Funktion unterscheiden und damit beispielsweise der Schlüssel $K$ ableiten oder eine Nachricht fälschen (Abb. 3). Eine notwendige Bedingung für ein sicheres Design ist also die Nicht-Existenz so einer Charakteristik mit hoher Wahrscheinlichkeit. Diese klassische Bedingung ist aber keineswegs hinreichend, wie mehrere Ergebnisse dieser Dissertation zeigen.




(a) DC: Schlüssel (b) DC: Fälschungen (c) DFA: Schlüssel (d) SIFA: Schlüssel

Abb. 3: Differenzen in Angriffen: Differenzielle Kryptanalyse (DC) und Fehlerangriffe (DFA, SIFA).

## 3 Differenzielle Kryptanalyse neuer symmetrischer Primitive

Im ersten Teil der Dissertation werden vier kürzlich vorgeschlagene neue Primitive und darauf basierende Modes analysiert. Die Analysetechniken bedienen sich dabei im weiteren Sinne der differenziellen Kryptanalyse. Da klassische differenzielle Analyse selbstverständlich bei allen untersuchten Primitiven im Designprozess berücksichtigt wurde und daher die Existenz geeigneter Charakteristiken durch die Designer ausgeschlossen wurde, schlagen wir verschiedene neue Techniken vor, die das Interface sowie konkrete Designentscheidungen ausnutzen. Die Ergebnisse beinhalten Full Breaks, also Widerlegung des Security Claims der Autoren in Form praktischer oder theoretischer Angriffe, aber auch Schwachstellen in reduzierten Primitiven mit verringerter Rundenzahl oder in einer Verwendungsweise, die nicht vom Security Claim abgedeckt wird. Mehrere unserer Angriffe konnten inzwischen weiterentwickelt und verbessert oder auf weitere Ziele angewendet werden. Zwei der Designs wurden in Reaktion auf die Angriffe aktualisiert, um die Schwachstellen zu beheben.

Unser erstes Analyseziel ist die leichtgewichtige tweakbare Blockchiffre MANTIS, publiziert auf der CRYPTO 2016 [Be16], der Top-Konferenz im Bereich Kryptographie. Durch den zusätzlichen Tweak-Input $T$, der der Kontrolle des Angreifers unterliegt, können unabhängig vom Inputblock $X$ weitere Differenzen in die Ausführung eingebracht werden. Die Designer berücksichtigen dies grundsätzlich und beweisen, dass trotzdem keine einzelne Charakteristik der Variante MANTIS-5 eine Wahrscheinlichkeit über $2^{-68}$ bei einer Blockgröße von $n = 64$ Bits aufweist. Zusätzlich schränken sie den Angreifer sicherheitshalber auf ein Datenlimit von $2^{30}$ Chosen-Plaintext Queries ein.

Eine erste Beobachtung zu dem Design ist, dass sich aufgrund spezieller differenzieller Eigenschaften der leichtgewichtigen Hauptbausteine von MANTIS, nämlich der involutiven S-box und der Near-MDS-Matrix im linearen Layer, vergleichsweise leicht verschiedene passende Charakteristiken $\chi$ zum selben Differential $(\alpha, \beta)$ finden lassen. In Folge entwickeln wir ein Framework zum Finden und Bewerten von großen, strukturierten Clustern aus Charakteristiken, die wir als Semi-Truncated Differential Characteristics bezeichnen. Diese kombinieren Vorteile von Truncated Characteristics (die etwa mit Mixed-Integer Linear Programming oder SMT-Solvern gesucht werden und einfach approximativ zu bewerten sind) mit der präziseren Auswertung und höheren Wahrscheinlichkeit von einzelnen Charakteristiken (per Hand oder mit dem in section 4 beschriebenen Tool gesucht) und sind besonders bei der Analyse von leichtgewichtigen Blockchiffren mit Tweak nützlich. Zusätzlich beschreiben wir, wie damit besonders effizient der geheime Schlüssel abgeleitet werden kann. Auf diese Weise finden wir einen Cluster mit einer hohen Wahrscheinlichkeit von $2^{-39}$, dessen spezielle Struktur es außerdem erlaubt, passende Inputs mit noch geringerer Datenkomplexität von etwa $2^{25}$ und damit unter dem von den Designern gesetzten Limit zu finden. In einer detaillierten Analyse identifizieren wir noch weitere Eigenschaften von MANTIS, die die Komplexität in die eine oder andere Richtung beeinflussen. Eine praktische, nicht parallelisierte Implementierung der Attacke findet schlussendlich den Schlüssel in unter einer Stunde, während die Designer von einer Mindestlaufzeit von astronomisch hohen $2^{96}$ Verschlüsselungsäquivalenten ausgingen. Die Publikation [Do17] wurde als eine der drei besten auf der Konferenz FSE 2017 ausgezeichnet.

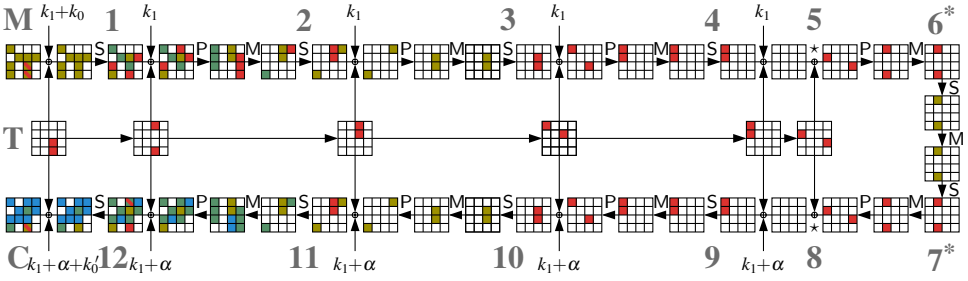



Abb. 4: Semi-Truncated Differential Characteristic für MANTIS-5 mit verschieden großen Mengen an erlaubten Differenzen pro Zelle ($|\chi_i| \in \{1$ ■, 4 ■, 13 ■, 15, 16 ■$\}$)

Unser zweites Analyseziel ist die Permutation Simpira [GM16], deren Designer (Intel, NIST) besonders auf hohe Performance auf modernen Desktop-CPUs abzielen. Dazu schlagen sie ein generalisiertes Feistel-Netzwerk (GFN) aus rundenreduzierten AES-Instanzen vor, das mit Intels AES-NI und vergleichbaren Instruktionserweiterungen anderer Plattformen äußerst effizient in Software implementierbar ist. Gleichzeitig soll die Verwendung der bewährten und gut analysierten AES-Rundenfunktion sowie des beweisbar sicheren Feistelnetzwerks einen hohen Security Margin bieten.

Wir können zeigen [DEM16], dass die computergestützte Sicherheitsanalyse der Designer (ebenso wie der Sicherheitsbeweis des GFNs) mehrere Abhängingkeiten zwischen Zwischenergebnissen der Permutation außer Betracht lässt. Dadurch sind die hergeleiteten Schranken für die maximale differenzielle Wahrscheinlichkeit einzelner Charakteristiken von $2^{-450}$ ungültig. Tatsächlich können mit einer theoretischen Komplexität von „nur“ $2^{110}$, also weniger als dem Security Claim von $2^{128}$, differenzielle Fixpunkte für die volle 512-bit-Permutation Simpira-4 gefunden werden (Abb. 5). In der von den Designern vorgeschlagenen Verwendung mit Feed-Forward als Hashfunktion entspricht das einer Kollision. Die Ergebnisse zeigen, wie unvorhersehbar der Security Margin bei schlüssellosen Primitiven wegen interner Abhängigkeiten sowie der Möglichkeit für den Angreifer, Zwischenergebnisse zu kontrollieren und deterministische Startstrukturen zu konstruieren, sein kann. Die Designer haben mittlerweile eine aktualisierte Variante Simpira v2, die den Fehler durch eine Neukonstruktion des GFNs behebt, publiziert.

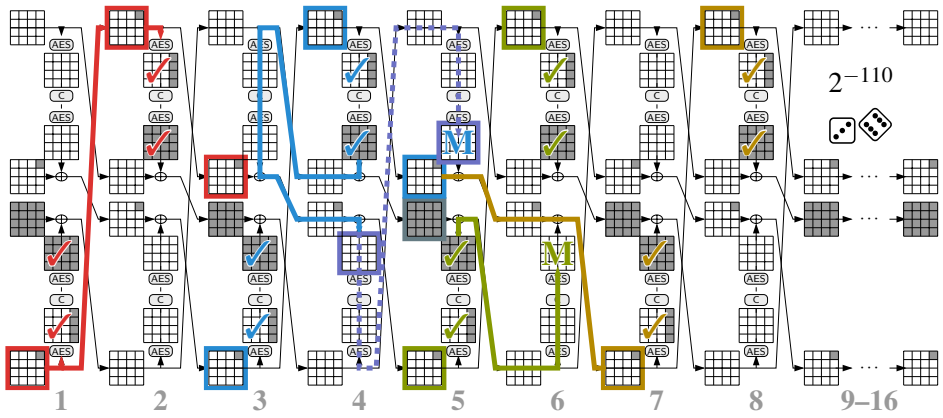



Abb. 5: Differenzieller Fixpunkt für Simpira-4 mit 16 von 15 Runden: 40 statt 80 aktive S-boxen, davon 20 deterministisch erfüllbar (markiert) und 20 probabilistisch mit Wahrscheinlichkeit $2^{-110}$

Weitere Analyseziele, die in der Dissertation diskutiert werden, sind die Blockchiffre LowMC (EUROCRYPT 2015) sowie das permutationsbasierte authentifizierte Verschlüsselungsverfahren Prøst (ein Kandidat in Runde 1 der CAESAR Competition). LowMC ist durch sein Designziel – die Minimierung der multiplikativen Komplexität – besonders anfällig gegen höherdimensionale differenzielle Angriffe, die nicht nur die erste Ableitung $\Delta_\alpha f(x)$, sondern höhere Ableitungen $\Delta^{(d)}_{\alpha_1,\ldots,\alpha_d} f(x) := \Delta_{\alpha_d} \cdots \Delta_{\alpha_1} f(x) = \bigoplus_{\alpha \in \langle \alpha_1 \ldots \alpha_d \rangle} f(x \oplus \alpha)$ betrachten und so den bereits niedrigen algebraischen Grad der Funktion weiter reduzieren. Wir erweitern dabei den von den Designern bereits analysierten einfachen Distin-

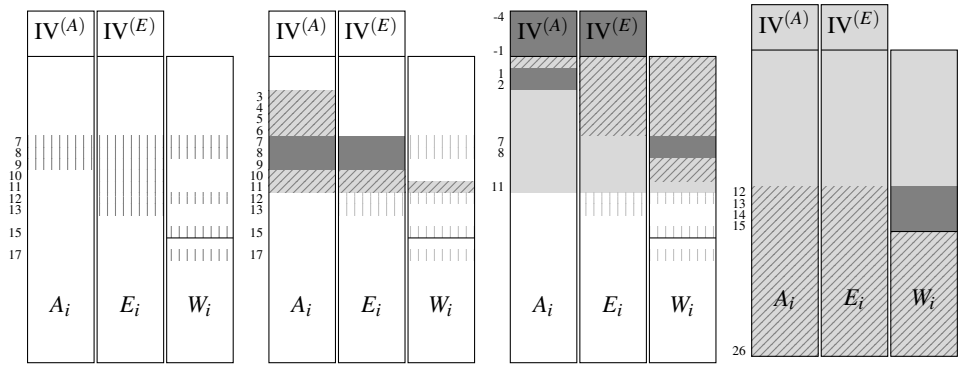
guisher um bis zu 4 Runden (bei einem Security Margin von nur 5 Runden), indem wir mehrere differenziell invariante Unterräume konstruieren und verketten. Die Designer haben mittlerweile eine aktualisierte Version LowMC v2 mit einem größeren Security Margin vorgeschlagen. Bei Prøst analysieren wir ausnahmsweise nicht das Primitiv, sondern den Mode of Operation, und zeigen, wie ein Angreifer, der Nachrichten unter zwei verwandten Schlüsseln beobachten kann, sehr einfach Fälschungen konstruieren kann. Diese Beobachtung ist zwar unerwartet, ist aber keine Bedrohung für die Sicherheit von Prøst.

Zusammenfassend zeigen unsere Ergebnisse, dass Design und Sicherheitsanalyse von schlüssellosen, tweakbaren, oder anderweitig atypischen Primitiven für aktuelle Anwendungs-Bedürfnisse nach wie vor eine Herausforderung sind. Insbesondere strikte Performance-Anforderungen können zu Designs führen, deren hochoptimierte Bausteine unerwünschte Eigenschaften mit sich bringen, die in klassischen Analysemethoden schwer einkalkulierbar sind. Während jeder unserer Angriffe spezifische Schwachstellen der jeweiligen Primitive identifiziert und ausnutzt, sind mehrere der Techniken, beispielsweise Semi-Truncated Differential Characteristics oder die Linearisierungstechniken für algebraische Analysen, von allgemeinerem Interesse für die Anwendung auf neue Primitive.

## 4 Automatische Tools für differenzielle Kryptanalyse

Der zweite Teil der Dissertation widmet sich der Verbesserung und Anwendung von automatischen Tools zur differenziellen Kryptanalyse von Hashfunktionen, insbesondere SHA-2. Während im ersten Teil hauptsächlich externe Solver für Mixed-Integer Linear Programming oder Boolean Satisfiability zum Einsatz kamen, wenden wir uns hier einem spezialisierten Such-Tool zu, das von Mendel, Nad und Schläffer [MNS11]für die Kollisionssuche in SHA-2 entwickelt wurde und intern eine spezialisierte Guess-and-Determine-Suche verwendet, die Parallelen zu SAT-Solvern aufweist. Die SHA-2-Familie von Hashfunktionen wird als aktueller US- und internationaler Standard allgegenwärtig eingesetzt, ob für die Integrität in TLS-Handshakes, Zertifikaten, SSH, oder Bitcoin. Die Familie besteht aus zwei Hauptzweigen: SHA-256 (bevorzugt für 32-bit-Plattformen) und SHA-512 (64-bit), wobei der Großteil der bisherigen Analyse sich auf SHA-256 bezieht, SHA-512 auf aktuellen Plattformen aber effizienter ist und gleichzeitig ein höheres Sicherheitsniveau bietet. Unser Fokus sind dabei die Herausforderungen bei der Anwendung vorher für SHA-256 entwickelter Analysestrategien auf SHA-512, die sich aus der doppelten Größe der Kompressionsfunktion von SHA-512 ergeben.

Wir erweitern das bestehende Such-Tool um Techniken zum schnelleren Propagieren von Informationen besonders in linearen Operationen sowie für eine zielgerichtetere Suche durch eine Look-Ahead-Heuristik, die Widersprüche früher identifizieren soll und somit die Zeit minimiert, die in „toten“ Abschnitten des Suchbaums verbracht wird. Mit dem verbesserten Tool können signifikant mehr Runden analysiert werden und so die besten praktische Kollisionsangriffe auf die Varianten von SHA-512 gefunden werden, insbesondere Kollisionen für 27 von 80 Runden (Abb. 6), Semi-Free-Start Collisions für 39 Runden, und Free-Start Collisions für bis zu 44 Runden je nach Variante (vorher bei SHA-512 alle nur für bis zu 24 Runden, bei SHA-256 auch Semi-Free-Start Collisions für 38 Runden).




(a) Charakteristik (b) Dichter Abschnitt (c) IV matchen (d) Sparser Abschnitt

Abb. 6: Phasen der Kollisionssuche für 27-Step SHA-512 mit Message $W_i$ und State-Wörtern $A_i, E_i$ in Step $i$: Geratene Werte ■ und Differenzen □, abgeleitete Werte ▨, fixe Werte ■ und Differenzen □

Zusätzlich verwenden wir das Tool, um aufzuzeigen, wie ein bösartiger Designer einige Rundenkonstanten im Vorgänger SHA-1 so wählen könnte, dass er ein Kollisions-Backdoor einbaut. Der Wert dieser Rundenkonstanten spielt normalerweise für die Sicherheit keine Rolle, doch wenn der Designer die Konstanten während der Durchführung eines Angriffsversuchs wählt (und diese Konstanten für das Design fixiert), kann er zukünftig ohne weiteren Aufwand Kollisionen mit einem bestimmten Präfix generieren. Wir konstruieren in einer Kollaboration [All14] Konstanten und passende Präfixe, mit denen nachher gültige, kollidierende Dateien in mehreren Formaten (z.B. `.sh`, `.rar`, `.mbr`, `.jpg`) erzeugt werden können (Abb. 7).

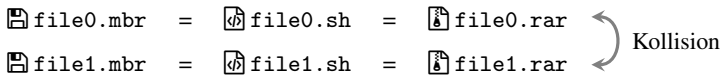



Abb. 7: Malicious SHA-1 mit Polyglot-Kollisionen.

## 5 Design und Implementierungssicherheit

Neben den obigen Kernthemen, die in der Dissertation ausgeführt werden, lieferte die Forschung im Rahmen des Doktorats auch Beiträge zu etwas breiter gestreuten Themen, insbesondere im Kontext von Implementierungssicherheit und neuen Designs.

Wie in Abb. 3 angedeutet ist ein Angreifer in der Praxis nicht darauf angewiesen, für eine Erlangung des Schlüssels mit differenziellen Methoden die notwendigen Differenzen wie bisher diskutiert über die Inputdaten einzuführen und dann mühsam, mit immer geringerer Wahrscheinlichkeit Runde um Runde, durch die Berechnungen zu verfolgen: Stattdessen kann er auch „schummeln“ und die Differenzen mit physikalischen Mitteln, etwa einem Laser oder durch kurzfristiges Manipulieren der Stromversorgung, direkt kurz vor Ende der Berechnung einfügen und die Effekte wesentlich günstiger auswerten [BS97] (Differential Fault Attack, DFA). Um solche Angriffe zu verhindern, werden in der Praxis diverse Implementierungsgegenmaßnahmen eingesetzt, die fehlerhafte Berechnungen

erkennen (Fault Detection) sowie das Lernen einzelner Zwischenergebnisse durch Seitenkanäle für den Angreifer nutzlos machen (Masking). Gemeinsam mit Kollegen und unter Verwendung der statistischen Analysemethoden der differenziellen Kryptanalyse konnten wir zeigen, dass ein Angreifer trotz dieser üblichen Gegenmaßnahmen ans Ziel kommen und den Schlüssel erlangen kann [Do18b, Do18a], indem er eine Differenz einführt, die abhängig von mehreren Datenpunkten in der weiteren Berechnung vor der Detection wieder verschwinden kann, und dann nur die fehlerfreien Fälle analysiert (Statistical Ineffective Fault Attack, SIFA).

Last but definitely not least ist die Mitentwicklung von Ascon zu erwähnen, einem authentifizierten Verschlüsselungsverfahren, das sich besonders durch ein leichtgewichtiges Design bei hoher Robustheit gegenüber diversen Implementierungsangriffen und -fehlern auszeichnet [Do19]. Ascon wurde 2014 als Kandidat zur CAESAR Competition, einem internationalen Wettbewerb für kryptographische Designs, eingereicht, und 2019 nach jahrelanger Analyse von der Jury als primäre Empfehlung für ressourcenbeschränkte Anwendungen ausgezeichnet. Aktuell ist Ascon Kandidat im „Lightweight Cryptography Standardization Process“ der US-Standardisierungsbehörde NIST.

## Literaturverzeichnis

[Al14] Albertini, Ange; Aumasson, Jean-Philippe; Eichlseder, Maria; Mendel, Florian; Schläffer, Martin: Malicious Hashing: Eve's Variant of SHA-1. In (Joux, Antoine; Youssef, Amr M., Hrsg.): Selected Areas in Cryptography – SAC 2014. Jgg. 8781 in LNCS. Springer, S. 1–19, 2014.

[Be16] Beierle, Christof; Jean, Jérémy; Kölbl, Stefan; Leander, Gregor; Moradi, Amir; Peyrin, Thomas; Sasaki, Yu; Sasdrich, Pascal; Sim, Siang Meng: The SKINNY Family of Block Ciphers and Its Low-Latency Variant MANTIS. In (Robshaw, Matthew; Katz, Jonathan, Hrsg.): Advances in Cryptology – CRYPTO 2016. Jgg. 9815 in LNCS. Springer, S. 123–153, 2016.

[BS90] Biham, Eli; Shamir, Adi: Differential Cryptanalysis of DES-like Cryptosystems. In (Menezes, Alfred; Vanstone, Scott A., Hrsg.): Advances in Cryptology – CRYPTO 1990. Jgg. 537 in LNCS. Springer, S. 2–21, 1990.

[BS97] Biham, Eli; Shamir, Adi: Differential Fault Analysis of Secret Key Cryptosystems. In (Kaliski Jr., Burton S., Hrsg.): Advances in Cryptology – CRYPTO '97. Jgg. 1294 in LNCS. Springer, S. 513–525, 1997.

[DEM16] Dobraunig, Christoph; Eichlseder, Maria; Mendel, Florian: Cryptanalysis of Simpira v1. In (Avanzi, Roberto; Heys, Howard M., Hrsg.): Selected Areas in Cryptography – SAC 2016. Jgg. 10532 in LNCS. Springer, S. 284–298, 2016.

[Do17] Dobraunig, Christoph; Eichlseder, Maria; Kales, Daniel; Mendel, Florian: Practical Key-Recovery Attack on MANTIS5. IACR Transactions on Symmetric Cryptology, 2016(2):248–260, 2017.

[Do18a] Dobraunig, Christoph; Eichlseder, Maria; Groß, Hannes; Mangard, Stefan; Mendel, Florian; Primas, Robert: Statistical Ineffective Fault Attacks on Masked AES with Fault Countermeasures. In (Peyrin, Thomas; Galbraith, Steven, Hrsg.): Advances in Cryptology – ASIACRYPT 2018. Jgg. 11273 in LNCS. Springer, S. 315–342, 2018.

[Do18b] Dobraunig, Christoph; Eichlseder, Maria; Korak, Thomas; Mangard, Stefan; Mendel, Florian; Primas, Robert: SIFA: Exploiting Ineffective Fault Inductions on Symmetric Cryptography. IACR Transactions on Cryptographic Hardware and Embedded Systems, 2018(3):547–572, 2018.

[Do19] Dobraunig, Christoph; Eichlseder, Maria; Mendel, Florian; Schläffer, Martin: , Ascon v1.2. Submission to NIST's Lightweight Cryptography Standardization Process, 2019. `https://csrc.nist.gov/CSRC/media/Projects/Lightweight-Cryptography/documents/round-1/spec-doc/ascon-spec.pdf`.

[Ei18] Eichlseder, Maria: Differential Cryptanalysis of Symmetric Primitives. Dissertation, Technische Universität Graz, 2018.

[GM16] Gueron, Shay; Mouha, Nicky: , Simpira: A Family of Efficient Permutations Using the AES Round Function. IACR Cryptology ePrint Archive, Report 2016/122, 2016.

[MNS11] Mendel, Florian; Nad, Tomislav; Schläffer, Martin: Finding SHA-2 Characteristics: Searching through a Minefield of Contradictions. In (Lee, Dong Hoon; Wang, Xiaoyun, Hrsg.): Advances in Cryptology – ASIACRYPT 2011. Jgg. 7073 in LNCS. Springer, S. 288–307, 2011.

**Maria Eichlseder**, geboren 1988, studierte an der Technischen Universität Graz Informatik (B.Sc., M.Sc.) und Technische Mathematik (B.Sc.). Im Anschluss begann sie 2013 ihr Doktoratsstudium unter der Betreuung von Florian Mendel und Christian Rechberger. Sie promovierte 2018 sub auspiciis praesidentis und wurde für ihre Arbeit mit dem Staatspreis für die besten Dissertationen 2018 (Award of Excellence des Bundesministeriums) ausgezeichnet. Das authentifizierte Verschlüsselungsverfahren Ascon, das sie im Rahmen ihres Doktorats mit entwickelte, wurde 2019 zum Gewinner der CAESAR Competition for Authenticated Encryption in der Kategorie „Lightweight Applications" gekürt. Sie ist derzeit als Postdoc an der TU Graz und forscht im Bereich symmetrische Kryptographie an Kryptanalyse und Design von effizienten Hashfunktionen, authentifizierten Verschlüsselungsverfahren, und deren Primitiven. Schwerpunkte sind dabei mathematische Aspekte der Kryptanalyse, Zusammenhänge zwischen kryptanalytischen und physikalischen Angriffen, sowie automatisierte Werkzeuge und Heuristiken für Kryptanalyse.

# Programmiermodelle und Unterstützung für umfassende Evaluationen im Umfeld von MPTCP Schedulern, Adaptionsentscheidungen und DASH Video Streaming[1]

Alexander Frömmgen[2]

**Abstract:** In der Dissertation wird aufgezeigt, dass die Analyse, die Umsetzung und die Evaluation von Kommunikationssystemen durch *i)* fehlende Abstraktionen und die resultierende Implementierungskomplexität sowie *ii)* die benötigten umfassenden Evaluationen für die Vielzahl an Konfigurationsmöglichkeiten und Netzwerkumgebungen erschwert werden.

Zur Überwindung dieser beiden Hindernisse präsentiert die Dissertation die folgenden Beiträge: *i)* drei Programmiermodelle für die Bereiche des Multipath TCP Paketschedulings, der adaptiven Kommunikationssysteme, sowie der Topologie-Adaption in Kommunikationssystemen, *ii)* 13 neue, ausführbare Multipath TCP Paketscheduler, sowie *iii)* ein wiederverwendbares Rahmenwerk zur nahtlosen Ausführung und Analyse umfassender Netzwerkexperimente.

## 1 Einführung

Kommunikationssysteme und das Internet sind in den letzten Jahren zu der zentralen Infrastruktur geworden. Dieser Beitrag gibt einen Überblick über die Dissertation des Autoren, welche sich der Weiterentwicklung heutiger Kommunikationssysteme, insbesondere mit Methoden und Werkzeugen zur Weiterentwicklung heutiger Kommunikationssysteme, beschäftigt.[3] In der Arbeit wird aufgezeigt, dass die Analyse, die Umsetzung und die Evaluation von Kommunikationssystemen durch *i)* fehlende Abstraktionen und die resultierende Implementierungskomplexität sowie *ii)* die benötigten umfassenden Evaluationen für die Vielzahl an Konfigurationsmöglichkeiten und Netzwerkumgebungen erschwert werden. Multipath TCP (MPTCP) [Fo13], das de facto Transportprotokoll zur Nutzung mehrerer Netzwerkpfade, stellt ein prominentes Beispiel dar. Innovationen im Bereich der Multipath TCP Paketscheduler werden durch die Implementierungskomplexität innerhalb des Linux-Kernels und den benötigten umfassenden Analysen für die Vielzahl an Anwendungen und Netzwerkumgebungen erschwert.

Zur Überwindung des ersten Hindernisses schlagen wir *ProgMP*, das erste Programmiermodell für Multipath TCP Paketscheduler, als Abstraktion für den Entwurf und die Entwicklung von Multipath TCP Paketschedulern vor [Fr17]. *ProgMP* beinhaltet eine aus-

[1] Englischer Originaltitel der Dissertation: Programming Models and Extensive Evaluation Support for MPTCP Scheduling, Adaptation Decisions, and DASH Video Streaming

[2] Die Dissertation ist im Rahmen der Tätigkeit im Sonderforschungsbereich MAKI an der TU Darmstadt an den Fachgebieten DVS und KOM entstanden. Kontakt: alexander.froemmgen@kom.tu-darmstadt.de

[3] Für eine detaillierte Darstellung der Beiträge und eine korrekte Würdigung der Grundlagen und der verwandten Arbeiten verweisen wir an dieser Stelle für das gesamte Dokument auf die Dissertationsschrift [Fr18a].

druckstarke Spezifikationssprache und eine einfache Programmierschnittstelle zur Spezifikation ausführbarer Multipath TCP Paketscheduler. Wir zeigen die Stärken von *ProgMP* am Beispiel 13 neuer Paketscheduler mit unterschiedlichen Optimierungszielen auf [Fr17, FHK18]. Wir schlagen den ersten redundanten Multipath TCP Paketscheduler vor und zeigen, dass dieser die Latenz für Anwendungen mit strikten Latenz- und moderaten Durchsatzanforderungen signifikant reduziert [Fr16]. Wir nutzen *ProgMP* für eine detaillierte Analyse von Entwurfsentscheidungen für die Verwendung von Redundanz zur Abwägung von Latenz und Durchsatz. Des Weiteren schlagen wir Paketscheduler vor, welche feingranulare Durchsatz- oder Latenzziele einhalten, sowie die Interaktion mit darüberliegenden Protokollen wie beispielsweise HTTP/2 optimieren und dabei Pfadpräferenzen einhalten. Unsere detaillierten Evaluationen mittels Netzwerkemulation sowie Echtweltmessungen zeigen, dass *ProgMP* effiziente Schedulingentscheidungen und eine Vielzahl neuer, ausführbarer Multipath TCP Paketscheduler ermöglicht. Neben *ProgMP*, unserem Hauptbeitrag zur Überwindung fehlender Abstraktionen, stellen wir des Weiteren Programmiermodelle als Abstraktionen für die Adaptionsentscheidung adaptiver Kommunikationssysteme vor. So schlagen wir vor, Adaptionsentscheidungen mittels *event condition action* Regeln (Ereignis, Bedingung, Aktion) zu spezifizieren und diese Regeln basierend auf genetischer Programmierung in umfassenden Netzwerkexperimenten automatisch für eine gegebene Nutzenfunktion zu lernen [Fr15]. Schließlich stellen wir ein Programmiermodell für die Spezifikation von Topologie-Adaptionen in Kommunikationssystemen mittels Graphmustern in der Topologie vor [St16a] und zeigen, wie die darunterliegende Suche nach Graphmustern verteilt ausgeführt werden kann [St18].

Zur Überwindung des zweiten identifizierten Hindernisses haben wir das *MACI* Rahmenwerk zur Verwaltung, skalierbaren Ausführung und interaktiven Analyse von umfassenden Netzwerkexperimenten entwickelt [Fr18b]. Im Kern ist *MACI* eine Kombination und Integration etablierter Werkzeuge zur Förderung gründlicher, nahtloser Evaluationen während des gesamten Forschungsprozesses. Wir diskutieren unsere *MACI* Erfahrungen während *i)* der Entwicklung und Evaluation unserer vorgeschlagenen *ProgMP* Scheduler [Fr17, FHK18], *ii)* der Entwicklung einer Multipatherweiterung für das Transportprotokoll QUIC [Vi18] *iii)* der Analyse eines verteilten Protokolls zur Auffindung von Graphmustern in Topologien [St18], sowie *iv)* eines systematischen Vergleichs von *DASH Video Streaming* Implementierungen [St17]. Unsere Erfahrungen bestätigen, dass *MACI* wiederkehrende Aufgaben in der Evaluation diverser Kommunikationssysteme unterstützt und dadurch die Forschungseffizienz signifikant erhöht. Die Experimente mit *MACI* für *ProgMP*, die Multipatherweiterung für QUIC, die Topologiemustererkennung und die Videoübertragung mittels *DASH* gehen über eine Evaluation von *MACI* hinaus und stellen signifikante Beiträge zum Verständnis der jeweiligen Gebiete dar.

Im Folgenden gehen wir exemplarisch auf zwei Beiträge der Dissertation genauer ein: *i)* das *ProgMP* Programmiermodell für Multipath TCP Paketscheduler sowie *ii)* das *MACI* Rahmenwerk für umfassende Netzwerkexperimente.

# 2 ProgMP: Ein Programmiermodell für MPTCP Paketscheduler

## 2.1 Einleitung und Motivation

Das Transportprotokoll TCP liegt den meisten heutigen Kommunikationssystemen zugrunde und ermöglicht die Kommunikation in verteilten Systemen und Anwendungen. Multipath TCP ist eine kürzlich im `RFC 6824` vorgeschlagene TCP-Erweiterung, welche es mittels des Konzepts sogenannter *Subflows* ermöglicht mehrere Netzwerkpfade für eine logische MPTCP Transportprotokollverbindung zu verwenden [Fo13]. Es wurde gezeigt, dass Multipath TCP den Durchsatz und die Zuverlässigkeit der Transportprotokollverbindung erhöhen kann und dabei dynamisch auf Veränderungen und Schwankungen der verfügbaren Netzwerkresourcen reagiert.

Multipath TCP erhöht die Komplexität im Vergleich zu traditionellem TCP, da Pakete eines Datenstroms auf mehrere *Subflows* verteilt werden müssen, bevor diese auf der Empfangsseite wieder zusammengeführt werden (Abbildung 1). Dafür benötigt Multipath TCP eine Instanz, welche für jedes zu übertragende Datenpaket entscheidet, auf welchem *Subflow* bzw. Netzwerkpfad dieses übertragen werden soll. Diese Entscheidungsinstanz wird Paketscheduler oder auch vereinfacht Scheduler genannt. Die *Paketschedulingentscheidung* hat einen großen Einfluss auf die Leistungsmerkmale des Transportprotokolls. So können ungeeignete Entscheidungen die Vorteile von Multipath TCP verschwinden lassen oder sogar die Gesamtleistung verringern.

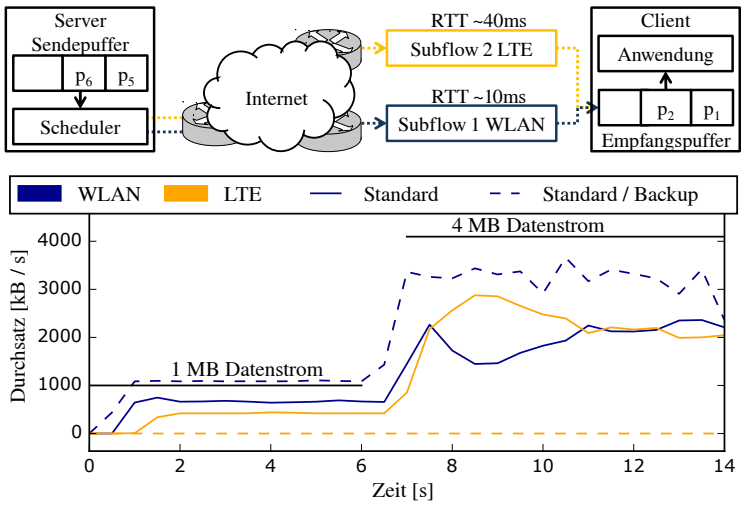



Abb. 1: **Bedarf an flexibleren Schedulern:** Versuchsaufbau und reproduzierbare Messergebnisse für eine interaktive Netzwerkanwendung mit dem MPTCP Standardscheduler *MinRTT* unter Verwendung eines WLAN- und eines LTE-*Subflows*. Weder der Standardscheduler noch der *Backup Modus* des Standardschedulers ermöglichen es, die Videobitrate von 4MB/s sicherzustellen und dabei präferenzbewusst den WLAN-*Subflow* auszuschöpfen bevor der LTE-*Subflow* verwendet wird.

Die heutigen MPTCP Implementierungen sind zumeist auf hohen Durchsatz optimiert. Viele Anwendungen benötigen jedoch für die jeweiligen Anforderungen und Charakteristika angepasste Schedulingstrategien. So haben Anwendungen beispielsweise unterschiedliche Anforderugnen bezüglich der erfahrenden Netzwerklatenz und dem erreichten Durchsatz und viele Anwender präferieren die Verwendung von WLAN über LTE für mobile Anwendungen. Bis zur Erstellung dieser Dissertation fehlten sowohl in der Forschung als auch in den Implementierungen Konzepte und Methoden zur Spezifikation und Verwendung optimierter anwendungs- und präferenzbewusster MPTCP Paketscheduler.

Abbildung 1 illustriert diese Limitationen anhand einer exemplarischen Messung einer interaktiven Videoübertragung. Für die Messung wurde eine MPTCP Verbindung mit einem WLAN- und einem LTE-*Subflow* zwischen einer AWS Serverinstanz und einem Laptop verwendet. Die ersten sechs Sekunden des Videostroms sind mit 1MB/s kodiert, der übrige Videostrom mit 4MB/s. Obwohl der 1MB/s Datenstrom vollständig auf dem WLAN-*Subflow* mit 10ms Umlaufzeit übertragen werden könnte, beobachten wir, dass der heutige Standardscheduler *MinRTT* ungefähr 30% des 1MB/s Stroms über den LTE-*Subflow* mit 40ms Umlaufzeit übertragt. Dies ist eine Folge der Durchsatz- und Ressourcenausgleichsoptimierungen des *MinRTT* Schedulers im Zusammenspiel mit der Staukontrolle und einer TCP-Optimierung für kleine Paketpuffer. Eine weitere Messung unter Verwendung des *Backup Modus* für den LTE-*Subflow* zeigt, dass dieser Modus den benötigten Durchsatz für den 4MB/s Strom nicht bereitstellen kann, da er effektiv zu einer Deaktivierung des LTE-*Subflows* führt.

Dieses symptomatische Beispiel zeigt das Fehlen eines einzelnen, optimalen Schedulers für alle Anwendungsfälle auf. Stattdessen benötigt MPTCP optimierte Paketscheduler zur Berücksichtigung von Anwendungsanforderungen und Pfadpräferenzen. Gleichzeitig beobachten wir, dass die Entwicklung, die Evaluation und die Verwendung neuer Multipath TCP Scheduler durch die hohe Implementierungskomplexität, etwa in der Standardimplementierung im Linux-Kernel, behindert wird. Entsprechend finden sich viele Vorschläge für optimierte Scheduler, welche nicht in einer Echtweltumgebung evaluiert wurden bzw. komplexe Änderungen an der zugrundeliegenden MPTCP Implementierung vermeiden.[4] Der Bedarf an optimierten Schedulern und die fehlenden Konzepte zur Erstellung dieser zeigt die Notwendigkeit an Abstraktionen zum effizienten Entwurf und zur Umsetzung von Multipath TCP Schedulern auf.

## 2.2 Das ProgMP Programmiermodell

Im Rahmen der Dissertation schlagen wir *ProgMP*, das erste Programmiermodell für den Entwurf und die Umsetzung von Multipath TCP Schedulern [Fr17], vor.[5] *ProgMP* besitzt eine Schedulerspezifikationssprache, welche es Kommunikationssystemforschern und Anwendungsentwicklern ermöglicht, Multipath TCP Scheduler auf einem hohen Abstraktionsniveau zu beschreiben. Neben der Spezifikationssprache beinhaltet *ProgMP* eine Programmierschnittstelle, welche es Anwendungen erlaubt, zur Laufzeit Informatio-

[4] Die Dissertation [Fr18a] beinhaltet eine Übersicht über 27 Paketscheduler im weiteren Multipath TCP Umfeld.
[5] Dokumentation, Implementierung und Beispiele zu *ProgMP* sind auf `https:/progmp.net` verfügbar.

nen und Hinweise an den Scheduler zu übergeben. Schließlich stellen wir eine effiziente Ausführungsumgebung für die spezifizierten Scheduler in der Multipath TCP Linux Kernelimplementierung bereit. Somit schließt *ProgMP* die Lücke zwischen einer einfachen, ausdrucksstarken Beschreibung der Multipath TCP Scheduler und der Erprobung und Nutzung von Schedulerinnvoationen in realen Anwendungen. Im Folgenden stellen wir weitere Details von *ProgMP* vor.

**Spezifikation von Schedulern** Die Spezifikationssprache von *ProgMP* ist eine domänenspezifische Sprache, welche Ausdrucksstärke, Verständlichkeit und Ausführbarkeit abwägt. Es sind alle relevanten Entitäten aus der Paketschedulingdomäne in der Sprache enthalten. So gibt es in der Sprache die drei Paketpuffer *i)* `Q` für die zu sendenden Pakete, *ii)* `QU` für die sich in der Übertragung befindenden Pakete, sowie *iii)* `RQ` für Pakete welche erneut übertragen werden müssen. Die Sprache bietet einfache Kontrollflussoperationen sowie deklarative Sprachkonstrukte zur Auswahl der Pakete aus den Puffern beziehungsweise des zu nutzenden Subflows aus der Menge aller Subflows `SUBFLOWS`. Das implizite, statische Typsystem reduziert im Zusammenspiel mit unveränderlichen Variablen mögliche Seiteneffekte während der Ausführung und ermöglicht eine effiziente Ausführung sowie eine einfache Fehlerbehandlung. Darüber hinaus kann eine einzelne Schedulerausführung kein, ein, oder mehrere Pakete unterschiedlichen Subflows zuweisen. Dies vereinfacht viele Schedulerspezifikationen erheblich und reduziert die Komplexität und somit die Fehlerwahrscheinlichkeit im Bereich der Zustandsverwaltung für die Schedulerentwicklerin.

Abbildung 2 stellt zur Illustration der Schedulerspezifikationssprache Auszüge aus zwei alternativen Schedulern zur Verwendung von Redundanz dar. Beide Scheduler nehmen an, dass `sbfCandidates` eine Menge an potenziellen Subflowkandidaten beinhaltet. Die erste Alternative (links) sendet (`PUSH`t) das nächste zu sendende Paket (`Q.TOP`) auf allen Subflowkandidaten und entfernt das Paket aus dem Puffer falls zumindest ein Subflowkandidat vorhanden ist. Die zweite Alternative (rechts) nutzt zwei unterschiedliche Strategien in Abhängigkeit vom Zustand des Sendepuffers. Befinden sich ungesendete Pakete in diesem (`!Q.EMPTY`) bevorzugt der Scheduler diese neuen Pakete über bereits gesendete Pakete und überträgt sie auf dem Subflow mit der minimalen Umlaufzeit (`RTT`). Sind alle Pakete

```
1 /* Alternative 1 für Redundanz */
2 ...
3 IF(!sbfCandidates.EMPTY) {
4   FOREACH(VAR sbf IN sbfCandidates) {
5     sbf.PUSH(Q.TOP);
6   }
7   DROP(Q.POP());
8 }
```

```
1  /* Alternative 2 für Redundanz */
2  ...
3  IF (!Q.EMPTY) {
4    sbfCandidates.MIN(sbf => sbf.RTT).PUSH(Q.POP());
5    RETURN;
6  }
7
8  VAR packetCandidate = QU.FILTER(packet =>
9    !sbfCandidates.FILTER(sbf =>
10     !packet.SENT_ON(sbf)).EMPTY).TOP;
11 sbfCandidates.FILTER(sbf =>
12   !packetCandidate.SENT_ON(sbf)).MIN(sbf => sbf.RTT).
13     PUSH(packetCandidate);
14 }
```

Abb. 2: Ausschnitte aus zwei *ProgMP* Schedulern, welche Pakete redundant auf mehreren Pfaden übertragen. Der linke Scheduler überträgt das Paket auf allen *Subflows*, welche zum Zeitpunkt der Entscheidung verfügbar sind. Der rechte Scheduler hingegen überträgt nur dann Pakete redundant wenn keine ungesendeten Pakete vorhanden sind.

zumindest auf einem *Subflow* übertragen nutzt der Scheduler Redundanz und übertragt die Pakete erneut auf einem anderen Subflow. Die beiden Beispiele zeigen, dass *ProgMP* es ermöglicht, komplexe Schedulingentscheidungen präzise in wenigen Zeilen zu spezifizieren. Im Vergleich dazu benötigt der traditionelle Ansatz mit Linux-Kernelmodulen 251 Zeilen C Programmcode für den redundanten Scheduler.

**Die Ausführungsumgebung** Wir haben eine *ProgMP*-Ausführungsumgebung im Multipath TCP Linux-Kernel implementiert und evaluiert. Die Ausführungsumgebung besteht aus einem Interpreter sowie einem eBPF-basierten Just-in-Time Compiler zur Sprachausführung, einer Implementierung der Pufferabstraktion und einer Umsetzung der Fehlerbehandlungsabstraktion. Darüber hinaus wird Domänenwissen genutzt, um die Schedulerausführung zur Laufzeit zu optimieren. So kann ein Scheduler etwa zur Laufzeit unter der Annahme einer festen Anzahl an *Subflows* optimiert und erneut kompiliert werden.

**Vorgestellte und Evaluierte Scheduler** Im Rahmen der Dissertation haben wir *ProgMP* zur Analyse und Verbesserung etablierter sowie zum systematischen Entwurf und zur Evaluation neuer MPTCP Scheduler verwendet. Die betrachteten Scheduler (Tabelle 1) optimieren dabei Leistungseigenschaften entsprechend unterschiedlicher Anwendungsanforderungen und Nutzer- und Pfadpräferenzen. Die Vorstellung und Evaluation der Scheduler stellt nicht nur eine Evaluation der Ausdrucksstärke, Tragweite und Anwendbarkeit von *ProgMP* dar, sondern ist darüber hinaus ein fundamentaler Beitrag zum Verständnis des Entwurfsraums an MPTCP Schedulern und eine signifikante Steigerung der MPTCP Leistung in vielen Anwendungsgebieten.

| | | Diskutierte Varianten | Anzahl Zeilen mit Präfix |
|---|---|---|---|
| **Etablierte Scheduler** | Präferenzen | | |
| Standard *MinRTT* | | 1 | 15 |
| *Backup Modus* | Binär | 1 | 15 + 7 = 22 |
| Round robin | Binär | 2 | 21 und 35 |
| Redundant[6] | Binär | 1 | 21 |
| **Neue Scheduler und Funktionen** | | | |
| Aktives Ausprobieren für zeitnahe RTT-Abschätzungen | | 3 | 31, 34 und 38 |
| Abwägen von Latenz und Ressourceneinsatz mittels Redundanz | | 3 | 17, 21 und 24 |
| Übertragungszeitminimierung in heterogenen Umgebungen | | 1 | 28 |
| Präferenzbewusstes Einhalten tolerierbarer RTTs | | 2 | 23 und 26 |
| Präferenzbewusstes Bereitstellen des benötigten Durchsatzes | | 2 | 22 und 35 |
| Einwegeverzögerungsbewusstes Scheduling | | 1 | 15 |
| Adaptives Scheduling für HTTP/2 | | 1 | 26 |

Tab. 1: Mit *ProgMP* analysierter Entwurfsraum an Multipath TCP Schedulern. Da die meisten Scheduler den gleichen Präfix von 11 Zeilen nutzen, werden effektiv zwischen 4 und 27 weitere Spezifikationszeilen benötigt. Die *ProgMP*-Spezifikationen der etablierten Scheduler benötigten 3,6% bis 12.4% der Anzahl an Zeilen der C-basierten Kernelmodule.

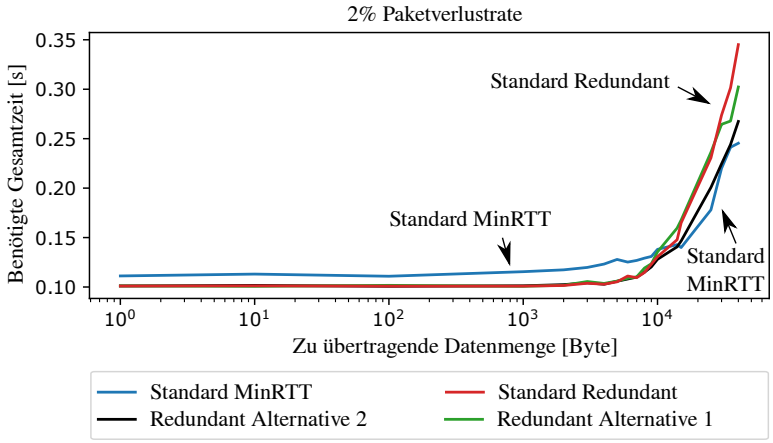



Abb. 3: Durchschnittlich benötigte Übertragungszeit in Abhängigkeit der zu übertragenden Datenmenge in jeweils 20 Experimenten mit Mininet, zwei homogenen *Subflows* und 2% Paketverlustrate.

Abbildung 3 zeigt exemplarisch einen experimentellen Vergleich der in Abbildung 2 vorgestellten Redundanzalternativen sowie des Standard *MinRTT* und redundanten Schedulers. Die Messung zeigt, dass alle drei Redundanzalternativen für geringe Datenmengen einen Vorteil gegenüber dem Standardscheduler haben. Für größere Datenmengen hingegen unterscheiden sich die Alternativen. Hier ist der Standardscheduler und Alternative 2 den beiden anderen redundanten Schedulern überlegen. Die exemplarische Messung zeigt, dass Scheduler durch die direkte Ausführbarkeit von *ProgMP*-Spezifikationen empirisch evaluiert und analysiert werden können und somit Schedulerinnovationen signifikant vereinfacht werden.

## 3 MACI: Ein Rahmenwerk für die wiederkehrenden Anforderungen umfassender Netzwerkexperimente

Die Forschung an Kommunikationssystemen ist auf Netzwerkexperimente angewiesen. Entsprechend wurden innerhalb der Forschungsgemeinschaft Methoden und Werkzeuge für kontrollierte und reproduzierbare Netzwerkexperimente entwickelt – etwa Netzwerksimulatoren und deren zugrundeliegenden Netzwerkmodellen. So ist eine Vielzahl an Simulatoren und Emulatoren verfügbar, welche auf unterschiedliche Anwendungen, unterliegende Abstraktionen und Netzwerkmodelle ausgerichtet sind. Experimente mit diesen Ausführungsumgebungen sind die Grundlage für den Entwurf und die Entwicklung heutiger Kommunikationssysteme und ermöglichen eine frühe und wiederkehrende Rückmeldung und Analyse der Systeme.

[6] Der redundante Scheduler wurde im Rahmen der Dissertation entwickelt [Fr16]. Er wird hier dennoch als *etabliert* klassifiziert, da er ursprünglich ohne *ProgMP* entwickelt wurde und einer von drei verfügbaren Schedulern für den MPTCP Linux-Kernel ist (`https://github.com/multipath-tcp/mptcp/blob/mptcp_v0.94/net/mptcp/mptcp_redundant.c`).

Im Rahmen der Dissertation haben wir beobachtet, dass Forscher und Entwickler wiederkehrend unterstützende Infrastruktur und Werkzeuge entwickeln, um die Ausführung und die Auswertung von Experimenten zu automatisieren. Die Entwicklung dieser Werkzeuge startet üblicherweise für jedes Forschungsprojekt von Neuem. Die Entwicklung beginnt meist mit kleinen Hilfsprogrammen, welche im Verlauf des Forschungsprojekts immer umfassender werden. Insgesamt nehmen sie schließlich einen wesentlichen Anteil am Gesamtaufwand des Forschungsprojekts ein. Auch wenn die Entwicklung dieser Werkzeuge eine geringe Komplexität aufweist, lenkt sie von der eigentlichen Forschungsarbeit ab und verzögert das Projekt.

In der Dissertation identifizieren wir die *wiederkehrenden* Anforderungen von und Aufgaben für netzwerkexperimentbasierte Studien, etwa *i)* die Spezifikation, die Verwaltung und die Dokumentation der Experimente und ihrer abhängigen und unabhängigen Parameter, *ii)* die skalierbare, parallele Ausführung umfassender Experimentstudien, sowie *iii)* die interaktive Analyse der Experimentergebnisse basierend auf den zuvor spezifizierten Parametern. Darüber hinaus argumentieren wir, dass ein integrierter Ansatz für diese Aufgaben die (Forschungs-) Effizienz deutlich erhöht.

Basierend auf diesen Beobachtungen präsentieren wir *MACI*[7], das erste maßgeschneiderte Rahmenwerk um die zuvor identifizierten wiederkehrenden Anforderungen und Aufgaben zu erfüllen [Fr18b]. *MACI* ist ein Rahmenwerk für die nahtlose Verwaltung, skalierbare Ausführung und interaktive Auswertung umfassender Netzwerkexperimente. Die nahtlose Integration dieser Funktionen ermöglicht es beispielsweise, die Ergebnisse der Experimente ohne manuelle Aufbereitung der Daten visuell darzustellen und basierend auf den damit gewonnen Einsichten weitere Experimentkonfigurationen in wenigen Schritten zu starten. *MACI* entstand als Ergebnis und basierend auf den Erfahrungen aus mehreren Forschungsprojekten [St16b, Fr15, Kh16].

*MACI* stellt eine geschickte Integration und Kombination etablierter Werkzeuge dar um eine gründliche, experimentbasierte Evaluation während des gesamten Forschungsprozesses zu fördern. *MACI* wendet beispielsweise die Konzepte der interaktiven Datenanalyse aus dem Bereich des Business Intelligence und des Data Science auf Netzwerkexperimente an. *MACI* folgt dem Zeitgeist der agilen Softwareentwicklung und der kontinuierlichen Integration in der Softwareentwicklung, indem es Hindernisse für kurze Iterationszyklen eliminiert. Dabei wird die zugrundeliegende Ausführungsumgebung für das einzelne Netzwerkexperiment als austauschbare Blackbox betrachtet um eine Vielzahl an etablierten Ausführungsumgebungen, etwa Netzwerksimulatoren und Emulatoren, zu unterstüzen.

Wir haben *MACI* erfolgreich für die in Tabelle 2 dargestellten Forschungsprojekte verwendet. Diese nutzen verschiedene Ausführungsumgebungen und befassen sich mit unterschiedlichen Protokollen auf diversen Netzwerkschichten. Dies zeigt, das *MACI* für diverse Forschungsprojekte im Bereich der Kommunikationssysteme anwendbar ist. *MACI* hat dabei alle wiederkehrenden Aufgaben im Bereich der experimentellen Evaluation signifikant vereinfacht und es uns somit ermöglicht, uns auf die eigentlichen Forschungsfragen zu fokussieren.

[7] *MACI* ist auf `https://maci-research.net` öffentlich verfügbar.

| | Ausführungsumgebung | Schicht/Protokoll |
|---|---|---|
| **Von Bachelor- und Masterstudenten genutzt** | | |
| Nachbildung einer bekannten MPTCP Experimentstudie | Mininet | Transport/MPTCP |
| Entwicklung einer Multipath-Erweiterung für QUIC [Vi18] | Mininet | Transport/MPQUIC |
| Lernen und Evaluieren von Staukontrollen für QUIC | Mininet | Transport/QUIC |
| **Von Doktoranden genutzt** | | |
| Entwicklung neuer MPTCP Scheduler [Fr17, FHK18] | Mininet | Transport/MPTCP |
| Analyse einer verteilten Topologiemustererkennung [St18] | Java | Diverse |
| Analyse und Vergleich von DASH Implementierungen [St17] | Mininet | Anwendung/DASH |

Tab. 2: Übersicht über bisherige *MACI*-Verwendungen.

Exemplarisch möchten wir hier den experimentellen Vergleich der unterschiedlichen Redundanzalternativen für Multipath TCP Scheduler mittels *MACI* nennen. Neben der direkten visuellen Analyse (Abbildung 3 basiert auf einer der direkt verfügbaren Visualisierungen) haben wir von der skalierbaren, parallelen Ausführung der Experimente profitiert. So basiert die genannte Abbildung auf 30.720 Experimenten, von denen jedes fast eine Minute benötigt. Mit *MACI* waren wir in der Lage, diese Experimente mit 20 AWS Instanzen an einem Tag für 95$ durchzuführen. Eine sequenzielle Ausführung hätte hingegen über 20 Tage und die selben finanziellen Aufwendungen benötigt.

Schließlich stellen wir fest, dass die gemeinsame Verwendung eines Werkzeugs und das dadurch geförderte Folgen einer gemeinsamen Evaluationsmethode die Zusammenarbeit in den Forschungsprojekten deutlich vereinfacht hat. Darüber hinaus hat das Bereitstellen eines Werkzeugs und einer bewährten Methoden Bachelor- und Masterstudenten bei ihren ersten Forschungsschritten unterstützt.

## Literaturverzeichnis

[FHK18] Frömmgen, Alexander; Heuschkel, Jens; Koldehofe, Boris: Multipath TCP Scheduling for Thin Streams: Active Probing and One-way Delay-awareness. In: Proceedings of the IEEE International Conference on Communications (ICC). 2018.

[Fo13] Ford, Alan; Raiciu, Costin; Handley, Mark; Bonaventure, Olivier: TCP Extensions for Multipath Operation with Multiple Addresses, RFC 6824, IETF. 2013.

[Fr15] Frömmgen, Alexander; Rehner, Robert; Lehn, Max; Buchmann, Alejandro: Fossa: Learning ECA Rules for Adaptive Distributed Systems. In: Proceedings of the IEEE International Conference on Autonomic Computing (ICAC). S. 207–210, 2015.

[Fr16] Frömmgen, Alexander; Erbshäußer, Tobias; Zimmermann, Torsten; Wehrle, Klaus; Buchmann, Alejandro: ReMP TCP: Low Latency Multipath TCP. In: Proceedings of the IEEE International Conference on Communications (ICC). 2016.

[Fr17] Frömmgen, Alexander; Rizk, Amr; Erbshäußer, Tobias; Weller, Max; Koldehofe, Boris; Buchmann, Alejandro; Steinmetz, Ralf: A Programming Model for Application-defined

Multipath TCP Scheduling. In: Proceedings of the ACM/IFIP/USENIX Middleware Conference, **Best Paper Award**, `https://progmp.net`. ACM, S. 134–146, 2017.

[Fr18a] Frömmgen, Alexander: Programming Models and Extensive Evaluation Support for MPTCP Scheduling, Adaptation Decisions, and DASH Video Streaming, Dissertation, TU Darmstadt. 2018.

[Fr18b] Frömmgen, Alexander; Stohr, Denny; Rizk, Amr; Koldehofe, Boris: Dont Repeat Yourself: Seamless Execution and Analysis of Extensive Network Experiments. In: Proceedings of the ACM International Conference on emerging Networking EXperiments and Technologies (CoNEXT), `https://maci-research.net`. S. 20–26, 2018.

[Kh16] Khuda Bukhsh, Wasiur; Rizk, Amr; Frömmgen, Alexander; Koeppl, Heinz: Optimizing Stochastic Scheduling in Fork-Join Queuing Models: Bounds and Applications. In: Proceedings of the IEEE INFOCOM. 2016.

[St16a] Stein, Michael; Frömmgen, Alexander; Kluge, Roland; Löffler, Frank; Schürr, Andy; Buchmann, Alejandro; Mühlhäuser, Max: TARL: Modeling Topology Adaptations for Networking Applications. In: Proceedings of the International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS). ACM, S. 57–63, 2016.

[St16b] Stohr, Denny; Frömmgen, Alexander; Fornoff, Jan; Zink, Michael; Buchmann, Alejandro; Effelsberg, Wolfgang: QoE Analysis of DASH Cross-Layer Dependencies by Extensive Network Emulation. In: Proceedings of the SIGCOMM Workshop on QoE-based Analysis and Management of Data Communication Networks (Internet-QoE). ACM, S. 25–30, 2016.

[St17] Stohr, Denny; Frömmgen[8], Alexander; Rizk, Amr; Zink, Michael; Steinmetz, Ralf; Effelsberg, Wolfgang: Where are the Sweet Spots?: A Systematic Approach to Reproducible DASH Player Comparisons. In: Proceedings of the ACM Conference on Multimedia (MM), `https://maci-research.net/dash`. S. 1113–1121, 2017.

[St18] Stein, Michael; Frömmgen, Alexander; Kluge, Roland; Lin, Wang; Wilberg, Augustin; Koldehofe, Boris; Mühlhäuser, Max: Scaling Topology Pattern Matching: A Distributed Approach. In: Proceedings of the ACM/SIGAPP Symposium on Applied Computing (SAC). 2018.

[Vi18] Viernickel, Tobias; Frömmgen, Alexander; Rizk, Amr; Koldehofe, Boris; Steinmetz, Ralf: Multipath QUIC: A Deployable Multipath Transport Protocol. In: Proceedings of the IEEE International Conference on Communications (ICC). 2018.

**Alexander Frömmgen** schloss im Jahr 2013 das Informatik-Studium an der Technischen Universität Darmstadt ab. Anschließend war Herr Frömmgen an den Lehrstühlen von Prof. Alejandro Buchmann und von Prof. Ralf Steinmetz als wissenschaftlicher Mitarbeiter im Sonderforschungsbereich MAKI tätig. Im Jahr 2018 wurde er am Fachbereich Informatik der TU Darmstadt zum Dr.-Ing. (mit Auszeichnung) promoviert. Herr Frömmgen arbeitet nun bei Google im Umfeld interaktiver Kommunikationsanwendungen.

[8] Die beiden ersten Autoren haben in gleichem Maße beigetragen.

# Methoden und Messverfahren für Mechanismen des automatischen Skalierens in elastischen Cloudumgebungen[1]

Nikolas Herbst[2]

**Abstract:** Auto-Skalierungsmechanismen für Cloud-Umgebungen versprechen stabile Servicequalität bei niedrigen Kosten und wechselnder Auslastung. Die großen, öffentlichen Cloud-Anbieter bieten regelbasierte Auto-Skalierer auf Basis von Schwellenwerten an. Diese Art des Auto-Skalierens hat jedoch Reaktionszeiten in der Größenordnung von Minuten. Neuartige Auto-Skalierungsmechanismen aus der Literatur versuchen, die Grenzen reaktiver Mechanismen durch den Einsatz proaktiver Vorhersagemethoden zu überwinden. Allerdings ist die Akzeptanz von proaktivem, automatischem Skalieren in der Produktion immer noch sehr gering, da das Risiko hoch ist, sich auf eine einzelne proaktive Methode zu verlassen. Diese Doktorarbeit befasst sich mit der Herausforderung, dieses Risiko zu reduzieren, indem sie unter anderem einen neuen hybriden automatischen Skalierungsmechanismus vorschlägt, der mehrere verschiedene proaktive Methoden kombiniert, die wiederum mit einem reaktiven Rückfallmechanismus gekoppelt sind. Hierbei werden bedarfsgesteuerte, automatisierte Prognoseverfahren zur Vorhersage der ankommenden Lastintensität in Kombination mit einer Servicebedarfsschätzung genutzt, um den erforderlichen Ressourcenverbrauch pro Arbeitseinheit zu berechnen, ohne dass eine Anwendungsinstrumentierung erforderlich ist. Der vorgeschlagene Ansatz wird mit fünf aktuellen proaktiven und reaktiven Auto-Skalierungsmechanismen in drei sowohl privaten wie öffentlichen Cloud-Umgebungen unter eigens entwickelten Wettbewerbsbedingungen fair verglichen. Dabei werden jeweils fünf repräsentative Arbeitslastverläufe generiert, die jeweils aus verschiedenen realen Aufzeichnungen entnommen sind. Insgesamt erreicht der in dieser Arbeit vorgeschlagene Ansatz das beste dynamische Skalierungsverhalten basierend auf Benutzer- und Elastizitätsmetriken von Ergebnissen aus 400 Stunden aggregierter Experimentierzeit.

## 1 Synopsis

Diese Zusammenfassung ist wie folgt strukturiert: Der Abschnitt 2 motiviert das Thema und setzt den Kontext für die hier zusammengefasste Doktorarbeit [He18a]. Im Anschluss fasst der Abschnitt 3 den aktuellen Stand der Technik zusammen und formuliert die Problemstellung. Schließlich hebt der Abschnitt 4 die zwei Leitziele dieser Doktorarbeit hervor, welche jeweils von vier Forschungsfragen zur Definition der Beitragspunkte begleitet werden. Die Dissertation selbst gliedert sich anhand vier separater Beiträge. Darauf aufbauend fasst der Abschnitt 5 die einzelnen Beiträge zusammen, während erste Einblicke in den Entwurf und die Ergebnisse der umfassenden Evaluation gegeben werden. Abschließend soll der Abschnitt 6 einen Ausblick vermitteln.

[1] Englischer Titel der Dissertation: “Methods and Benchmarks for Auto-Scaling Mechanisms in Elastic Cloud Environments”

[2] Universität Würzburg, Fakultät für Mathematik und Informatik, nikolas.herbst@uni-wuerzburg.de

## 2 Motivation und Kontext

Vor etwas mehr als einem Jahrzehnt im Jahr 2006 stellte Amazon Web Services (AWS) als erster kommerzieller Anbieter Cloud-Dienstleistungen der allgemeinen Öffentlichkeit bereit und löste damit einen Hype Rund um Themen des Cloud Computing in Wissenschaft und Industrie aus. Erst zwei Jahre später begann der Wettbewerb im Bereich des Cloud Computing, nachdem Microsoft, Google und IBM mit eigenen Cloud Diensten auf den Markt kamen. Während AWS als Marktführer ein Drittel des Marktes hält, teilen sich Microsoft, Google und IBM ein weiteres Drittel des Marktes. In den folgenden Jahren verzeichnete der Cloud-Computing Markt überwältigende Wachstumsraten, welche sich laut einem Gartner Report[3] in einem Marktvolumen von 247 Milliarden US-Dollar im Jahr 2017 widerspiegeln. Begleitet von einer Reihe neu gegründeter, hochkarätiger Konferenzen (z.B. IEEE Cloud, ACM Symposium on Cloud Computing SoCC) und Fachzeitschriften (z.B. IEEE Transactions on Cloud Computing), hat die Forschergemeinschaft in den letzten zehn Jahren unzählige Publikationen im Bereich des Cloud Computing veröffentlicht.

Nunmehr gehört diese Phase des ausgeprägten Wachstums der Vergangenheit an und befindet sich im Übergang zu einer Stabilisierungs- und Reifephase mit Wachstumsraten von unter 18%, wie von Gartner prognostiziert wird. Dennoch verändert das Cloud Computing Paradigma den Betrieb von Rechenzentren weiter. Chief-Executive-Officer Marc Hurd[4] der Oracle Corporation prognostiziert, dass bis 2025 80% der klassischen Rechenzentren verschwinden werden, da Anwendungen in der Produktion zunehmend in Cloud-Umgebungen betrieben werden. Insbesondere in der anhaltenden Stabilisierungs- und Reifephase von Cloud Computing-Angeboten hängen der wissenschaftliche Fortschritt und das Branchenwachstum von etablierten Messverfahren und einer standardisierten Berichterstattung der Qualitätsmerkmale von Cloud-Systemen ab, wie in einem kürzlich erschienenen Gigaom-Analystenbericht[5] aufgezeigt wird.

Laut einem Gartner-Bericht von 2009[6], ist das wichtigste Verkaufsargument von Cloud Computing-Angeboten ihr Pay-per-Use-Modell ohne langfristige Investitionen und Betriebskosten für den Nutzer. In Kombination mit der Basistechnologie der Hardwarevirtualisierung bietet das Pay-per-Use-Servicemodell die Möglichkeit, die zugewiesenen Rechenressourcen elastisch an die aktuelle Nachfrage anzupassen. Cloud-Betreiber können ihre physischen Ressourcen - zumindest in der Theorie - so verwalten, dass die Effizienz optimiert wird. Dabei geht es in einigen Fällen auch darum, mehr virtuelle Ressourcen zu verkaufen als physisch verfügbar sind - auch bekannt als Überbuchung. Gleichzeitig versucht der Betreiber das Betriebsrisiko für den Kunden auf ein maßgeschneidertes Minimum zu beschränken, indem man ihm die Möglichkeit gibt, Ressourcenprioritäten zu definieren.

[3] Gartner Cloud Report 2017: `https://www.gartner.com/newsroom/id/3616417`
[4] Oracle CEO Marc Hurd: `https://markhurd.com/about-mark-hurd/`
[5] Gigaom Analyst Report: Die Bedeutung von Benchmarking Clouds: `https://gigaom.com/report/the-importance-of-benchmarking-clouds/`
[6] Gartner Highlights Fünf Attribute von Cloud Computing: `https://www.gartner.com/doc/965212/refining-attributes-public-private-cloud`

## 3 Stand der Technik und Problemstellung

Die elastische Skalierung der zugewiesenen Rechenressourcen erfolgt durch so genannte Auto-Skalierungsmechanismen, die überwachten Leistungskennzahlen analysieren. Dabei haben die Mechansimen die Aufgabe die Ressourcenzuweisung dem aktuellen Bedarf derart dynamisch anzupassen, dass im Optimalfall die Leistung stabil bleibt und die Ressourcen effizient genutzt werden. Gängige Praxis ist die Verwendung simpler, Schwellwert-basierter Mechanismen, die aufgrund ihrer reaktiven Natur zu Leistungseinbußen während der Zeiten von Bereitstellungsverzögerungen führen. Im Gegensatz dazu sind die Verantwortlichkeiten von Cloud-Infrastrukturbetreibern hochkomplex um arbeitslastabhängige Ressourcenplatzierung, Wechselwirkungen, Lastverteilung, Dimensionierungs- und Routingfragen kontinuierlich zu optimieren. Als Folge dieser komplexen und miteinander verflochtenen Wechselwirkungen und dynamischer Optimierungspotentiale, erleben die Cloud-Kunden eine hohe Leistungsvariabilität. Das stellt für unternehmenskritische Anwendungen immer noch einen Hinderungsgrund dar, Cloud-Lösungen zu nutzen [IYE11].

Im Laufe des letzten Jahrzehnts wurde in der Literatur eine große Anzahl von Auto-Skalierungsmechanismen vorgeschlagen, die versuchen die Grenzen reaktiver Mechanismen zu überwinden, indem sie proaktive Prognosemethoden anwenden. Lorido-Botran et al. [LBMAL14] untersuchten diese Mechanismen systematisch. Sie schlagen vor, Auto-Skalierungsmechanismen in Ansätze aus der Warteschlangentheorie, der Kontrolltheorie, der Zeitreihenanalyse und dem maschinellen Lernen zu gruppieren.

Proaktive Autoskalierungsverfahren auf Basis der Zeitreihenanalyse schätzen die Ressourcennutzung, die Reaktionszeiten oder die Systemlast unter Verwendung einfacher Regressionsverfahren, Histogrammanalysen oder basierend auf Black-Box-Methoden wie den auto-regressiven integrierten gleitenden Durchschnitten (ARIMA-Modelle). Letztere haben bekannte Defizite in Bezug auf Laufzeit und Genauigkeit in Szenarien mit komplexen saisonalen Mustern und bei feiner als halbstündlich aufgelösten Zeitreihenwerten. Andere Ansätze, z.B. nicht quelloffene Autoskalierer mit Beteiligung von Google oder Entwicklungen von Netflix, nutzen Signalverarbeitungsmethoden, um das Frequenzspektrum über Fourier- oder Wavelet-Transformationen zu charakterisieren, ohne die Fähigkeit zur Erfassung von Trends zu unterstützen. Auto-Skalierungsmechanismen, welche die Theorie der Warteschlangenbildung nutzen, werden in den meisten Fällen mit einem der anderen Ansätze kombiniert. Kontrolltheoretische Ansätze teilen die Einschränkung kurzer Vorhersagehorizonte, während auf maschinellem Lernen basierende Methoden auf Trainingsphasen beziehungsweise bei Systemanpassungen auf Rekalibrierungsperioden angewiesen sind, die in Produktionsumgebungen nicht realisierbar sind. Darüber hinaus teilt die Mehrheit der vorgeschlagenen Auto-Skalierungsmechanismen die Annahme einer linearen und endlosen Skalierbarkeit der von der Cloud zur Verfügung gestellten Ressourcen. In der Praxis ist diese Annahme aufgrund von Kommunikationsaufwänden, Überbuchungspraktiken und zustandsbehafteten Anwendungsdiensten nicht realistisch.

Mit wenigen Ausnahmen bleiben die in der Literatur vorgeschlagenen proaktiven Auto-Skalierungsmechanismen ohne offengelegte Code-Artefakte. Diese Tatsache reduziert die Reproduzierbarkeit der Versuchsergebnisse und die Vergleichbarkeit der Alternativen er-

heblich. Infolgedessen ist der Einsatz von proaktiven Auto-Skalierern in der Produktion immer noch sehr gering, da das Risiko hoch ist, wenn man sich auf eine einzige proaktive Methode verlässt, auf deren Grundlage automatische Skalierungsentscheidungen getroffen werden. Laut einer eigens durchgeführten systematischen Literaturanalyse werden etwa 40% der Publikationen der Cloud-Forschung mittels Simulation evaluiert. Da diese Literaturanalsyse auch Auto-Skalierungsmechanismen umfasst, kann gesagt werden, dass es gängige Praxis ist, Auto-Skalierungsansätze mit simulativen Werkzeugen zu bewerten. Experimentelle Auto-Skalierer-Evaluationen werden in der Regel in mehr oder weniger ähnlicher Art und Weise durchgefuhrt. Es wird in einer Fallsstudie veranschaulicht, dass die vorgeschlagene Methode in der Lage ist, die Einhaltung des Dienstgüte im Vergleich zu einer beliebigen statischen Ressourcenzuweisung zu verbessern. Die Auswertungsszenarien dabei werden oft von synthetischen Lastintensitätsprofilen wie Sinussignalen oder Sägezahnmustern ohne wirkliche Repräsentativität getrieben, zumal sie von einer proaktiven Autoskalierung leicht vorhersehbar sind. Reale Lastprofile weisen eine Mischung aus Trend-, Saison-, Burst- und Rauschkomponenten auf und sind daher komplex zu erfassen, zu teilen, zu modifizieren und skaliert zu erzeugen.

Der Bewertungsprozess von Cloud-Infrastrukturangeboten in Bezug auf die Qualität der realisierten Elastizität bleibt unspezifiziert. Es fehlt an präzise definierten, aussagekräftigen Metriken gibt, welche die Qualität der tatsächlich erzielten Anpassungen elastischer Ressourcen unter Einhaltung von spezifizierten Messablaufregeln erfassen. Daher ist keine klare Anleitung für die Auswahl und Konfiguration eines Auto-Skalierers für einen gegebenen Kontext verfügbar. Die wenigen bestehenden proaktiven Auto-Skalierer werden auf eine sehr anwendungsspezifische Weise optimiert und in der Regel unter Verschluss gehalten, während im Gegensatz dazu viele andere Artefakte von Cloud-Software inzwischen quelloffen sind.

Zusammenfassend lässt sich sagen, dass das in dieser Arbeit behandelte Problem sich wie folgt formulieren lässt: Das Risiko ist nach wie vor hoch, einen proaktiven Auto-Skalierungsalgorithmus in einer Produktionsumgebung einzusetzen und hat eine geringe Akzeptanz zur Folge. Bestehende Lösungen sind entweder nicht offengelegt oder maßgeschneidert. In der Literatur beschriebene Ansätze werden nicht mit Hilfe eines standardisierten Benchmarks getestet, der faire Vergleiche ermöglichen würde.

## 4 Leitziele und Forschungsfragen

Nachdem eine Reihe von Defiziten im aktuellen Stand der Technik und bei der Bewertung von Auto-Skalierern identifiziert wurde, formuliert die Arbeit nun zwei übergeordnete Leitziele. Diese wiederum werden jeweils von vier Forschungsfragen begleitet. Die Doktorarbeit selbst ist in drei Teile gegliedert. Teil I stellt die Hintergründe und Grundlagen vor, die zum Verständnis der Beiträge der Arbeit erforderlich sind, sowie eine umfassende Zusammenfassung des Technikstandes. Teil II konzentriert sich auf das erste Leitziel A mit seinen Forschungsfragen. Schließlich behandelt Teil III das zweite Leitziel mit seinen vier weiteren Forschungsfragen, während Teil IV die Evaluation der einzelnen Beiträge enthält.

**Leitziel A:** Es gilt einen Benchmark für moderne Auto-Skalierer zu entwickeln, um das Vertrauen in neuartige proaktive Mechanismen zu stärken.

Um dieses Leitziel zu erreichen, werden die jeweiligen Herausforderungen in mehrere Teilziele aufgeteilt. Zunächst formulieren zwei Forschungsfragen die Notwendigkeit repräsentative Lastintensitätsprofile flexibel definieren, modifizieren und generieren zu können, um eine realistische Menge an Ressourcenanpassungen auszulösen. Zweitens erfassen zwei weitere Forschungsfragen die Notwendigkeit einer fundierten Definition von Metriken und Messmethodik als Bausteine für einen Elastizitätsbenchmark.

**A.1:** Wie können Lastintensitätsprofile aus realen Aufzeichnungen auf anschauliche, kompakte, flexible und intuitive Weise definiert werden?

**A.2:** Wie lassen sich automatisch Modelle von Lastintensitätsprofilen aus bestehenden Aufzeichnungen mit einer angemessenen Genauigkeit und Rechenzeit extrahieren?

**A.3:** Was sind sinnvolle und intuitive Metriken zur Quantifizierung von Genauigkeit, Timing und Stabilität als Qualitätsaspekte bei der Anpassung elastischer Ressourcen?

**A.4:** Wie können vorgeschlagene Elastizitätsmetriken zuverlässig und wiederholbar gemessen werden, um faire Vergleiche und auch konsistente Rangordnungen über Systeme mit unterschiedlicher Leistungsfähigkeit zu ermöglichen?

**Leitziel B:** Es gilt das Risiko der Verwendung neuartiger Auto-Skalierer im Betrieb zu reduzieren, indem mehrere proaktive Mechanismen genutzt werden, die auch in Kombination mit einem herkömmlichen Reaktionsmechanismus nutzbar sind.

Die Herausforderung dieses Ziels werden adressiert, indem zunächst ein neuartiger hybrider Auto-Skalierungsmechanismus vorschlagen und dann seine Leistung im Detail mit den bestehenden modernen Auto-Skalierern verglichen wird. Die umfassende Auto-Skalierer-Bewertung wird durch die Ergebnisse von Leitziel 1 ermöglicht. Zweitens werden Defizite der derzeitigen Zeitreihenprognosemethoden behoben, indem ein hybrider Prognosemechanismus vorschlagen und seine Vorteile im Kontext der automatischen Skalierung aufzeigt wird.

**B.1:** Wie können widersprüchliche Auto-Skalierungsentscheidungen aus unabhängigen reaktiven und proaktiven Entscheidungsschleifen kombiniert werden, um die Gesamtqualität der Anpassungsentscheidungen zu verbessern?

**B.2:** Wie gut schneidet der vorgeschlagene, hybride Auto-Skalierungsansatz im Vergleich zu modernsten Mechanismen in realistischen Umgebungen und Anwendungsszenarien ab?

**B.3:** Wie kann ein hybrider Prognosemechanismus auf der Grundlage der Zerlegung so konzipiert werden, dass er in der Lage ist, genaue und schnelle Vorhersagen komplexer saisonaler Zeitreihen zu liefern?

**B.4:** Ist ein solcher hybrider Prognoseansatz in der Lage, die Leistung und Zuverlässigkeit von Auto-Skalierungsmechanismen zu verbessern?

## 5 Beiträge und Zusammenfassung der Evaluation

Nachdem zwei Leitziele und je vier Forschungsfragen definiert wurden, werden nun die vier Kernbeiträge dieser Arbeit zusammengefasst. Jeder Beitrag greift zwei der Forschungsfragen und damit je einen Teil von Ziel A oder B auf. Die Beiträge haben gemeinsam, dass sie aufeinander aufbauen und im Endergebnis integriert sind.

**Beitrag I:**


Zur Adressierung von Leitziel A und Beantwortung der Forschungsfragen A.1 und A.2, schlägt die Arbeit ein beschreibendes Lastprofil-Modellierungssystem zusammen mit einer automatisierten Modellextraktion aus Aufzeichnungen vor, um eine reproduzierbare Erzeugung von Arbeitslasten mit realistischen Lastintensitätsschwankungen zu ermöglichen. Der Modellentwurf folgt dem Ansatz der Zerlegung aufgezeichneter Daten in ihre deterministischen Komponenten aus stückweise definierten Trends und wiederkehrenden oder übergreifenden Saisonmustern, wobei gleichzeitig stochastische Rauschverteilungen und explizite Spitzen modelliert werden können. Die Komponenten sind im Prinzip stückweise definierte mathematische Funktionen, die verschachtelt und mit mathematischen Operationen in einem Funktionsbaum kombiniert werden können. Automatisierte Extraktionsprozesse erkennen Frequenzen und zerlegen aufgezeichnete Daten basierend auf einer effizienten Heuristik. Das vorgeschlagene Modell nennt sich “Descartes Load Intensity Model (DLIM)” mit seiner Limbo Werkzeugkette, die wichtige Funktionen zum Benchmarking von Ressourcenmanagementansätzen auf repräsentative und faire Weise bereitstellt.

Die Ausdrucksmächtigkeit des DLIM-Modells wird bewertet, indem unterschiedliche Konfigurationen der Extraktionsprozesse anwendet und auf zehn verschiedenen realen Aufzeichnungen verglichen werden, die zwischen zwei Wochen und sieben Monaten an Anfrageraten umfassen. Automatisch extrahierte DLIM-Modellinstanzen weisen einen durchschnittlichen Modellierungsfehler von 15,2% auf. In Bezug auf Genauigkeit und Verarbeitungsgeschwindigkeit liefern die vorgeschlagenen Extraktionsmethoden, die auf deskriptiven Modellen basieren, bessere oder ähnliche Ergebnisse im Vergleich zu bestehenden nichtdeskriptiven Zeitreihenzerlegungsmethoden. Im Gegensatz zu DLIM-Modellen liefern klassische Zeitreihenzerlegungsansätze drei Reihen von Datenpunkten als Ausgabe im Gegensatz zu einem kompakten und flexiblen deskriptiven Modell. Dieser Beitrag führte zu einem Fachzeitschriftenartikel in den ACM Transactions on Autonomous and Adaptive Systems (TAAS) [Ki17], der 2017 veröffentlicht wurde.

**Beitrag II:**


Um die Forschungsfragen A.3 und A.4 anzugehen, wird zunächst eine klare Definition des Begriffs "**Elastizität**" im Cloud Computing[7] erarbeitet. Die Kernaspekte der Elastizität werden beschrieben und von verwandten Begriffen wie Effizienz und Skalierbarkeit abgegrenzt. Darüber hinaus wird eine Reihe von neuen, intuitiv verständlichen Metriken für die Quantifizierung von Timings-, Stabilitäts- und Genauigkeitsaspekten der Elastizität definiert. Basierend auf diesen Metriken, die auch von der Forschergruppe der Standard Performance Evaluation Corporation SPEC [8] befürwortet wurden, wird einen neuartigen Ansatz für das Benchmarking der Elastizität von Auto-Skalierern vorgeschlagen. Dabei können die praktisch erzielte Elastizität von "Infrastructure-as-a-Service" (IaaS)-Cloud-Plattformen unabhängig von der Leistungsfähigkeit der zugrunde liegenden Ressourcen bewertet und verglichen werden. Das vorgeschlagene Bungee Elastizitätsbenchmarking-Werkzeug nutzt die Modellierungsfunktionen von DLIM, um realistische Lastintensitätsprofile zu erzeugen.

In der zugehörigen Evaluation wird gezeigt, dass für jede der vorgeschlagenen Metriken ein konsistentes Ranking der elastischen Systeme auf einer Ordinalskala geliefert wird. Die Bungee-Messmethodik kann sowohl reproduzierbare Ergebnisse in einer kontrollierten Umgebung als auch Ergebnisse mit einer akzeptablen Variation in unkontrollierten Umgebungen wie öffentlichen Clouds erzielen. Schließlich wird eine umfangreiche Fallstudie von realer Komplexität präsentiert, die zeigt, dass der vorgeschlagene Ansatz in realistischen Szenarien anwendbar ist und mit unterschiedlichen Leistungsniveaus der zugrunde liegenden Ressourcen umgehen kann. Die Definitionen der Elastizitätsmetriken wurden zu einem wesentlichen Bestandteil eines Artikels in den ACM Transactions on Modeling and Performance Evaluation of Computing Systems (ToMPECS) [He18b].

**Beitrag III:**


In diesem Beitrag wird nun das zweite Leitziel angegangen: Das Risiko, sich auf einen einzigen proaktiven Auto-Skalierer zu verlassen, wird reduziert durch einen neuartigen hybriden Auto-Skalierungsmechanismus namens Chameleon. Dabei werden mehrere proaktive Methoden kombiniert, die wiederum mit einem reaktiven Rückfallmechanismus gekoppelt sind. Chameleon verwendet bedarfsgesteuerte, automatisierte, Zeitreihen-basierte Prognoseverfahren, um die ankommende Lastintensität in Kombination mit Schätzverfahren für den Ressourcenbedarf vorherzusagen. Es ist erforderlich den Ressourcenverbrauch pro Arbeitseinheit zu berechnen, ohne dass eine Anwendungsinstrumentierung vorgenommen wurde. Der Ansatz kann auch strukturelles Anwendungswissen nut-

[7] Die in dieser Arbeit vorgeschlagene Definition der Elastizität im Cloud Computing [HKR13] wurde von Wikipedia in einem entsprechenden enzyklopädischen Artikel aufgegriffen (c.f. `https://en.wikipedia.org/wiki/Elasticity_(cloud_computing)`).

[8] Standard Performance Evaluation Corporation SPEC Forschergruppe `http://research.spec.org`

zen, indem es Warteschlangennetzwerke in Produktform löst, aus denen dann optimierte Skalierungsaktionen abgeleitet werden. Der Chameleon-Ansatz löst Konflikte zwischen reaktiven und proaktiven Skalierungsentscheidungen auf intelligente Weise und nutzt als Bausteine die wichtigsten Entwicklungen der Descartes-Forschungsgruppe wie die Descartes Modellierungssprache (DML) [Hu17] zur Erfassung von Anwendungsstrukturen in Rechenzentren sowie die Bibliothek für Schätzverfahren von Ressourcenbedarfen (LibReDE) [Sp15].

In der Arbeit wird ein umfangreicher Auto-Skalierer-Wettbewerb durchgeführt unter Nutzung der Ergebnisse aus den Beiträgen I und II: Der Chameleon Ansatz wird systematisch mit vier verschiedenen modernen proaktiven Auto-Skalierern sowie einem herkömmlichen Schwellenwert-basierten in drei unterschiedlichen Cloud-Umgebungen verglichen: (i) eine private CloudStack-basierte Cloud-Umgebung, (ii) die öffentliche AWS EC2 Cloud, sowie (iii) eine OpenNebula-basierte geteilte IaaS-Cloud. Es werden insgesamt fünf repräsentative Lastprofile generiert, die jeweils aus verschiedenen realen Aufzeichnungen stammen. Die Funktionalität der Werkzeuge Limbo (Beitrag I) und Bungee (Beitrag II) werden genutzt, um eine variierende, CPU-intensive Systemauslastung zu erreichen. Die Beispielanwendung wird dem SPEC Server Efficiency Rating (SERT) Werkzeug entnommen und berechnet Matrixdekompositionen. Insgesamt erreicht Chameleon das beste und stabilste Skalierungsverhalten basierend auf Benutzer- und Elastizitätsmetriken. Dabei werden die Ergebnisse von 400 Stunden aggregierter Experimentierzeit analysiert. Es wird gezeigt, dass durch die Kombination von Skalierungsentscheidungen aus reaktiven und proaktiven Zyklen, basierend auf der vorgeschlagenen Konfliktlösungsheuristik, die Auto-Skalierungsleistung von Chameleon verbessert wird. Dieser Beitrag führte zu einem Artikel zu den IEEE Transactions on Parallel and Distributed Systems (TPDS) [Ba19].

**Beitrag IV:**


Als weiterer Beitrag dieser Arbeit werden Forschungsfragen B.3 und B.4 beantwortet, indem einen ein neuartiges Prognoseverfahren für Zeitreihen namens Telescope vorschlagen wird. Telescope integriert mehrere individuelle Prognosemethoden, indem es die univariaten Zeitreihen in die Komponenten Trend, Saison und Rest zerlegt. Zunächst wird automatisch die Frequenz bestimmt sowie Anomalien erkannt und beseitigt. Danach wird die Art der Zerlegung (multiplikativ oder additiv) basierend auf einer Mehrheitsentscheidung von maßgeschneiderten Tests ermittelt. Nach der Zerlegung wird die ARIMA-Methode (autoregressive integrierte gleitende Durchschnitte) ohne Saisonalität auf das Trendmuster angewendet, wobei die ermittelten saisonalen Muster einfach fortgesetzt werden. Darüber hinaus werden die einzelnen Perioden gruppiert, um kategorische Informationen zu erhalten. Die Cluster-Labels werden durch den Einsatz künstlicher neuronaler Netze prognostiziert. Dies hilft, automatisch zwischen verschiedenen Arten von Tagen zu unterscheiden. Schließlich wird eXtreme Gradient Boosting (XGBoost), eine neuartige und vielversprechende Methode, die 2016 veröffentlicht wurde, verwendet, um die Abhängigkeit zwischen allen zuvor extrahierten Kovariablen zu ermitteln und die Prognosen der einzelnen Komponenten zu kombinieren.

Die Bewertung zeigt anhand von zwei Zeitreihen, dass eine prototypische Implementierung des Telescope-Ansatzes sechs aktuelle Prognosemethoden in Bezug auf die Genauigkeit übertrifft. Telescope verbessert auch die Berechnungszeiten im Vergleich zu den drei wettbewerbsfähigsten Prognosemethoden um das bis zu 19-fache. In einer Fallstudie wird gezeigt, dass Telescope in der Lage ist, die Auto-Skalierungsleistung von Chameleon im Vergleich zu der früher verwendeten Prognosemethode tBATS oder saisonalem ARIMA weiter zu verbessern.

## 6 Ausblick

Die vier Kernbeiträge dieser Arbeit bringen das Potenzial, die Art und Weise zu verändern, wie Cloud-Ressourcenmanagement-Ansätze bewertet werden. Das wiederum dürfte eine Verbesserung der Qualität von autonomen Managementalgorithmen als Ergebnis mit sich bringen. Um eine solche Entwicklung zu unterstützen, wurden Code-Artefakte aller vier Beiträge dieser Arbeit als quelloffene Software-Werkzeuge veröffentlicht, die aktiv gepflegt und von Benutzer- und Entwicklerleitfäden begleitet werden[9].

Über die in den einzelnen Kapiteln explizit genannten Einschränkungen und Annahmen hinaus sehen gibt es eine Reihe von Herausforderungen für die zukünftige Forschung im Cloud-Ressourcenmanagement und dessen Bewertungsmethoden: (I) Die Einführung der Containerisierung auf virtuellen Maschineninstanzen führt zu einer weiteren Ebene der Indirektion. Infolgedessen erhöht die Verschachtelung virtueller Ressourcen die Ressourcenfragmentierung und verursacht unzuverlässige Bereitstellungsverzögerungen. (II) Des Weiteren neigen virtualisierte Rechenressourcen dazu, immer inhomogener zu werden, verbunden mit verschiedenen Prioritäten und Kompromissen. (III) Durch DevOps-Praktiken werden Updates für Cloud-gehostete Dienste mit einer höheren Frequenz veröffentlicht, was sich auf die Dynamik des Nutzerverhaltens auswirkt. Auto-Skalierungsmechanismen müssen sich zunehmend selbst an sich ändernde Serviceanforderungen und Ankunftsmuster anpassen.

## Literaturverzeichnis

[Ba19] Bauer, André; Herbst, Nikolas; Spinner, Simon; Ali-Eldin, Ahmed; Kounev, Samuel: Chameleon: A Hybrid, Proactive Auto-Scaling Mechanism on a Level-Playing Field. IEEE Transactions on Parallel and Distributed Systems, 30(4):800–813, September 2019.

[He18a] Herbst, Nikolas: Methods and Benchmarks for Auto-Scaling Mechanisms in Elastic Cloud Environments. Dissertation, Universität Würzburg, Deutschland, Juli 2018. SPEC Kaivalya Dixit Distinguished Dissertation Award 2018.

[He18b] Herbst, Nikolas; Bauer, André; Kounev, Samuel; Oikonomou, Giorgos; van Eyk, Erwin; Kousiouris, George; Evangelinou, Athanasia; Krebs, Rouven; Brecht, Tim; Abad, Cristina L.; Iosup, Alexandru: Quantifying Cloud Performance and Dependability: Taxonomy, Metric Design, and Emerging Challenges. ACM Transactions on Modeling

[9] Descartes Tools: `https://descartes.tools`

and Performance Evaluation of Computing Systems (ToMPECS), 3(4):19:1–19:36, August 2018.

[HKR13] Herbst, Nikolas; Kounev, Samuel; Reussner, Ralf: Elasticity in Cloud Computing: What it is, and What it is Not. In: Proceedings of the 10th International Conference on Autonomic Computing (ICAC 2013). USENIX, June 2013. Top 1 most cited ICAC paper (according to Google Scholar).

[Hu17] Huber, Nikolaus; Brosig, Fabian; Spinner, Simon; Kounev, Samuel; Bähr, Manuel: Model-Based Self-Aware Performance and Resource Management Using the Descartes Modeling Language. IEEE Transactions on Software Engineering (TSE), 43(5), 2017.

[IYE11] Iosup, A.; Yigitbasi, N.; Epema, D.: On the Performance Variability of Production Cloud Services. In: CCGrid 2011. S. 104–113, 2011.

[Ki17] von Kistowski, Jóakim; Herbst, Nikolas; Kounev, Samuel; Groenda, Henning; Stier, Christian; Lehrig, Sebastian: Modeling and Extracting Load Intensity Profiles. ACM Transactions on Autonomous and Adaptive Systems (TAAS), 11(4):23:1–23:28, Januar 2017.

[LBMAL14] Lorido-Botran, Tania; Miguel-Alonso, Jose; Lozano, Jose A: A Review of Auto-scaling Techniques for Elastic Applications in Cloud Environments. Journal of Grid Computing, 12(4):559–592, 2014.

[Sp15] Spinner, Simon; Casale, Giuliano; Brosig, Fabian; Kounev, Samuel: Evaluating Approaches to Resource Demand Estimation. Elsevier Performance Evaluation, 92:51 – 71, October 2015.

**Nikolas Herbst** leitet die Forschergruppe für prädiktive Datenanalyse am Lehrstuhl für Software Engineering der Universität Würzburg. Er promovierte 2018 an derselben Universität. Bevor er im Jahr 2014 vom Forschungszentrum für Informatik (FZI) am Karlsruher Institut für Technologie (KIT) zusammen mit seinem Doktorvater Samuel Kounev an die Würzburger Universität wechselte, erwarb er 2012 am KIT ein Diplom der Informatik. Nikolas is gewählter, stellvertretender Vorsitzender der SPEC Research Cloud Gruppe. Zu seinen Forschungsthemen gehören neben Elastizität im Cloud Computing, Autoskalierung und Ressourcemanagement auch die Leistungsbewertung von virtualisierten Umgebungen, Techniken des Autonomic und Self-Aware Computing, sowie der Datenanalyse und Modellbildung mittels Kombinationen von maschinellem Lernen und stochastischen Verfahren.

# Ein Verifizierter GDGL-Löser und Smales 14. Problem[1]

Fabian Immler[2]

**Abstract:** Diese Dissertation stellt eine Formalisierung von gewöhnlichen Differentialgleichungen (GDGL) und die Verifikation von rigorosen (mit garantierten Fehlerschranken) numerischen Algorithmen im interaktiven Theorembeweiser Isabelle/HOL vor. Die Formalisierung umfasst Fluss und Poincaré-Abbildung dynamischer Systeme. Die verifizierten Algorithmen basieren auf Runge-Kutta-Verfahren und affiner Arithmetik. Sie zertifizieren numerische Schranken für den Lorenz Attraktor und heben dadurch den numerischen Teil von Tuckers Beweis von Smales 14. Problem auf eine formale Grundlage.

## 1 Einleitung

Gewöhnliche Differentialgleichungen (GDGLen) modellieren eine Vielzahl an Systemen unserer alltäglichen Umgebung. Sei es die Bewegung von Teilchen, Autos, Zügen, Flugzeugen oder Planeten – oft ist eine zuverlässige Analyse sicherheitskritisch. Diese Dissertation zeigt dass derartige Analysen mit höchsten Korrektheitsgarantien ausgeführt werden können – nämlich mit formaler (und maschinenüberprüfbarer) formaler Logik.

Dies ist eine anspruchsvolle Thematik, da sie sich im Spannungsfeld von Mathematik und Informatik bewegt, zwischen der Modellierung kontinuierlicher Phänomene und Ihrer Implementierung mit diskreten Datenstrukturen, zwischen logischer Deduktion und numerischen Berechnungen.

Wir verwenden formale Logik, um zwischen diesen gegensätzlichen Konzepten zu vermitteln und stellen so eine mächtige Verbindung zwischen numerischen Berechnungen und logischer Deduktion her. Zudem schlagen wir die Brücke zwischen Theorie und Praxis: wir stellen realistische Implementierungen mit vernünftiger Performanz vor, welche zeigen, dass alle theoretisch etablierten Korrektheitsgarantien können auch praktisch ausgenutzt werden.

Wir konzentrieren uns auf sogenannte rigorose GDGL-Löser. Neben der Verifikation von bekannten und etablierten Methoden erforscht (und verifiziert formal) diese Arbeit auch neue Algorithmen zur Erreichbarkeitsanalyse kontinuierlicher Systeme. Die praktische Anwendbarkeit wird anhand einschlägiger Probleme aus dem Bereich cyber-physikalischer Systeme (im Rahmen der ARCH-Software-Competition) demonstriert.

Die finale große Anwendung ist Tucker's Beweis von Smale's 14. Problem, welches das volle Potential formal verifizierter Berechnungen aufzeigt: Bisher musste man für Tucker's Beweis den Berechnungen eines komplexen Computerprogramms vertrauen, nun

[1] Englischer Titel der Dissertation: "A Verified ODE Solver and Smale's 14th Problem"

[2] Institut für Informatik, Technische Universität München, immler@in.tum.de

bekommt man formale Garantien, die die Korrektheit des verwendeten Algorithmus auf die sicheren Grundlagen formaler Logik hebt.

**Rigorose GDGL-Löser.** Meist haben GDGLen keine symbolische Lösung, weshalb man auf numerische Verfahren zurückgreift, um Simulationen ihrer Evolution zu studieren. Für sicherheitskritische Anwendungen sind sogenannte *rigorose GDGL-Löser* (engl. rigorous ODE solver) von besonderer Bedeutung: Sie berechnen garantierte Schranken auf alle auftretenden Näherungsfehler.

Allerdings hängt die Korrektheit der berechneten Schranken davon ab, dass die zugrundeliegenden mathematischen Ideen wahrheitsgemäß implementiert werden. Dies ist – auch im Licht der zahllosen Beispiele von fehlerhaften Implementierungen, besonders in sicherheitskritischen Anwendungen – eine wirkliche Sorge. In der Tat existieren Beispiele von Fehlern in vermeintlich rigorosen GDGL-Lösern auf. Es gibt sogar Beispiele von Mängeln in zugrundegelegten mathematischen Ideen.

**Formale Verifikation.** Ein umfassender Ansatz um Korrektheit einer Implementierung sicherzustellen ist *formale Verifikation*. Das bedeutet, über Computerprogramme in einem strengen, logischen Kalkül zu argumentieren. Für Programme mit aufwändiger Spezifikation oder schwierigen Korrektheitsargumenten reichen vollautomatische Verifikationswerkzeuge nicht aus, hier ist formale Verifikation in *interaktiven Theorembeweisern* das Mittel der Wahl. Interaktive formale Verifikation verursacht enormen Aufwand. Dieser Aufwand lohnt, wenn starke Korrektheitsgarantien benötigt werden, insbesondere für grundlegende, etwaig sicherheitskritische Computerprogramme. Bemerkenswerte Beispiele sind der verifizierte Compiler CompCert [Le09] oder der verifizierte Betriebssystemkern seL4 [Kl09].

GDGLen sind allgegenwärtig in vielen wissenschaftlichen und technischen Disziplinen und demnach von ähnlicher grundlegender Bedeutung. Trotzdem wurde bisher kein rigoroser GDGL-Löser formal verifiziert. Es gibt reichlich Arbeit an (rigoroser) Numerik, von IEEE-Gleitkommazahlen bis hin zu formal verifizierten Bibliotheken für rigorose numerische (engl. rigorous numerics oder self-validated numerics) Berechnungen und nichtlinearer Optimierung. Bibliotheken für Mathematik, insbesondere Analysis, sind in vielen Beweisassistenten gut ausgebaut (wie in der Übersicht von Boldo *et al.* [BLM16] dargestellt), jedoch gibt es keine umfassende Formalisierung der Theorie von GDGLen.

- Ein Ziel dieser Arbeit ist die formale Verifikation eines rigorosen GDGL-Lösers – ein *verifizierter GDGL-Löser*.

Dies ist ein herausforderndes Unternehmen, da es sowohl die Formalisierung der zugrundeliegenden Mathematik als auch schwieriger Algorithmen verlangt. Diese Herausforderungen sind es Wert anzunehmen, da uns ein formal verifizierter GDGL-Löser mit bisher nicht gekannten Garantien für die berechneten Schranken belohnt.

Zudem streben wir eine Implementierung an, die nicht nur formal verifiziert ist, sondern auch angemessen effizient um realistische Beispielprobleme zu bearbeiten.

**Smale's 14. Problem.** Ein – neben sicherheitskritischen Systemen – weiteres Gebiet, in dem starke Korrektheitsgarantien wünschenswert sind ist die Mathematik, insbesondere im Kontext von Computerbeweisen (engl. computer-assisted proofs). Computerbeweise sind Beweise, die von der Ausgabe eines Computerprogramms abhängen und demnach entscheidend von einer korrekten Implementierung. Computerbeweise waren schon Ziel formaler Verifikation, wie sich in herausragenden Beispielen wie dem Flyspeck Projekt für einen formalen Beweis der Keplerschen Vermutung [Ha15] sowie der formalen Verifikation des Vier-Farben-Satzes [Go08] zeigt.

- Mit dem verifizierten GDGL-Löser zielen wir auf eine spezielle Anwendung ab: Tucker's Computerbeweis von Smale's 14. Problem (der Lorenz Attraktor).

Dies war ein wichtiges ungelöstes Problem, seine Lösung brachte Tucker z.B. den Preis der European Mathematical Society ein. Das Problem war Teil einer von Fields-Medaillenträger Stephen Smale zusammengestellten Liste mathematischer Probleme für das 21. Jahrhundert, neben der berühmten Riemannschen Vermutung oder der Frage ob $\mathsf{P} = \mathsf{NP}$. Tucker's Beweis [Tu02] basiert auf numerischen Schranken die von einem von ihm speziell für dieses Problem geschriebenen rigorosen GDGL-Löser berechnet wurden. In den ersten Versionen von Tucker's Programm wurden Fehler gefunden (und korrigiert). Aber dies verdeutlicht den Mehrwert formaler Verifikation: Mit einem verifizierten GDGL-Löser können wir sicher sein dass die berechneten Schranken nicht von etwaigen verbleibenden Fehlern kompromittiert sind.

## 2 Beiträge

Der zentrale Beitrag dieser Dissertation ist die formale Verifikation eines rigorosen GDGL-Lösers. Grundlegend für die Verifikation ist die Formalisierung von Mathematik GDGLen, insbesondere die Begriffe von Fluss und Poincaré-Abbildung. Die wesentliche Anwendung ist der Lorenz Attraktor, die Verifikation der Berechnungen von Tuckers Computerbeweis.

Der verifizierte GDGL-Löser ist modular aufgebaut, wesentlich um seine Verifikation handhaben zu können. Die verschiedenen Module sind derart unterschiedlich, dass jedes für sich genommen als unabhängiger Beitrag gesehen werden kann. Nichtsdestotrotz sind alle Module entlang Tuckers Beweis motiviert: Die Lorenz-Gleichungen erzeugen ein dynamisches System und demnach einen Begriff von Fluss. In seinem Beweis verwendet Tucker eine zur Analyse dynamischer Systeme übliche Technik, die sogenannte Poincaré-Abbildung. Tucker beweist, dass der Lorenz-Attraktor chaotisch ist, also insbesondere sensitiv von Anfangsbedingungen abhängt. Diese Abhängigkeit wird mit Ableitungen von Fluss und Poincaré-Abbildung quantifiziert. Neben der Formalisierung der abstrakten

mathematischen Konzepte implementieren und verifizieren wir rigorose numerische Algorithmen, die garantierte Schranken auf diese Quantitäten berechnen. Abschließend wenden wir diese Algorithmen auf Tuckers Computerbeweis an.

Alle in dieser Dissertation genannten Entwicklungen sind formalisiert, das heißt, in einer formalen Sprache geschrieben und mit einem maschinenüberprüfbaren Kalkül bewiesen. Die Formalisierung umfasst etwa 50000 Zeilen an Beweistext und ist im "Archive of Formal Proof" [IH17] verfügbar.

## 3 Formalisierung von Mathematik und GDGLen

Zur Formalisierung verwenden wir den Interaktiven Theorembeweiser Isabelle [Pa89] und seine populärste Instantiierung mit Logik höherer Stufe, Isabelle/HOL [NPW02]. Isabelle folgt der Tradition von Edinburgh LCF und besitzt einen kleinen Kern. Dieser Kern stellt eine abstrakte Schnittstelle zu sogenannten *formalen Theoremen* zur Verfügung und implementiert primitive logische Schlussregeln. Das System stellt sicher, dass formale Theoreme nur durch Anwendung primitiver Schlussregeln auf existierende formale Theoreme konstruiert werden können. Ein formales Theorem liefert demnach höchsten Standard mathematischer Strenge: seine Gültigkeit kann auf die Axiome der zugrundeliegenden Logik zurückgeführt werden.

**Formalisierung von Mathematik in Isabelle/HOL.** Wir verwenden Isabelle/HOL wird als Logik zur Formalisierung von Mathematik. Ein Alleinstellungsmerkmal von Isabelle/HOL unter anderen HOL-basierten Theorembeweisern sind *(axiomatische) Typklassen*. Typklassen sind sehr gut geeignet, um hierarchische Strukturen von Räumen in der Mathematik darzustellen und um polymorphe Spezifikationen zu organisieren. Hier geht es etwa um die Formalisierung von topologischen, metrischen, und Vektorräumen. Die Dissertation arbeitet aus, wie die (existierende) Analysis-Bibliothek entlang einer Hierarchie von Typklassen organisiert ist. Typklassen erlauben es beispielsweise, abstrakt Theoreme über metrische Räume zu beweisen. Konkrete Typen (etwa der Typ der reellen Zahlen, oder endlichdimensionale reellwertige Vektoren) können dann als Instanz einer Typklasse deklariert werden, was es erlaubt, die abstrakten Theoreme für den konkreten Typen zu verwenden.

**GDGLen in Isabelle/HOL.** Dieser Abschnitt liefert einen Überblick über die Formalisierung von GDGLen in Isabelle/HOL und stellt einige der formalisierten Hauptresultate heraus. Wir nehmen eine GDGL als gegeben durch Ihre rechte Seite $f$ an: $\dot{x}(t) = f(t, x(t))$. Für einen Anfangswert $x(t_0) = x_0$ zu Anfangszeit $t_0$ existiert (unter Annahmen an $f$) eine eindeutige Lösung. Insbesondere dann wenn $f$ nicht von $t$ abhängt, wird die Lösung für $t_0 = 0$ als *Fluss* $\phi(x_0, t)$ bezeichnet (um die Abhängigkeit vom Anfangswert $x_0$ zu betonen). Die Poincaré-Abbildung ist ein wichtiges Werkzeug zur Analyse dynamischer Systeme. Ein Poincaré-Schnitt $\Sigma$ ist eine Teilmenge des Zustandsraums. Für einen Punkt

$x \in \Sigma$, ist die Poincaré-Abbildung definiert als der Punkt $P(x)$, an dem der Fluss von $x$, zuerst zum Poincaré-Schnitt rückkehrt.

$$
\begin{aligned}
(\phi \textit{ solves-ode } f)\ T\ X :=& (\phi \textit{ has-vderiv-on } (\lambda t.\ f(t, \phi\ t)))\ T \wedge (\forall t \in T.\ \phi\ t \in X) \\
\textit{lipschitz } S\ g\ L :=& (\forall x, y \in S.\ \textit{dist } (g\ x)\ (g\ y) \leq L \cdot \textit{dist } x\ y) \\
\textit{local-lipschitz } f :=& \forall (t,x) \in T \times X.\ \exists \varepsilon > 0.\ \exists L. \\
& \forall u \in \mathscr{B}_\varepsilon(t) \cap T.\ \textit{lipschitz } (\lambda x.\ f(u,x))\ (\mathscr{B}_\varepsilon(x) \cap X)\ L \\
\textit{ll-open } T\ X\ f :=& \textit{open } T \wedge \textit{open } X \wedge \textit{local-lipschitz } T\ X\ f \wedge \\
& (\forall x \in X.\ \textit{continuous-on } T\ (\lambda t.\ f(t,x))) \\
(\phi \textit{ uniquely-solves-ode } f \textit{ from } t_0)\ T\ X :=& (\phi \textit{ solves-ode } f)\ T\ X \wedge t_0 \in T \wedge \textit{is-interval } T \wedge \\
& (\forall \psi.\ \forall T' \subseteq T.\ (t_0 \in T' \wedge \textit{is-interval } T' \wedge (\psi \textit{ solves-ode } f)\ T'\ X \wedge \psi\ t_0 = \phi\ t_0) \longrightarrow \\
& (\forall t \in T'.\ \psi\ t = \phi\ t)) \\
\phi(x_0, t) :=& \textit{sol } 0\ x_0 \\
\tau(x) :=& (\mathsf{LEAST}\ t > 0.\ \phi(x,t) \in \Sigma) \\
P(x) :=& \phi(x, \tau(x))
\end{aligned}
$$

Abb. 1: Auswahl an formalen Definitionen

Die wesentlichen Theoreme, die formalisiert wurden, betreffen lokale und globale Existenz und Eindeutigkeit, und Abhängigkeiten von Anfangsbedingungen von Fluss und Poincaré-Abbildung. Eine Auswahl an relevanten formalen Definitionen ist in Abbildung 1. Ein erstes grundlegendes Theoreme ist die Existenz einer (auf dem maximalen Existenzintervall *ex-ivl*, dessen Definition wir hier auslassen) eindeutigen Lösung *sol*. Technische Schwierigkeiten treten bei der Formalisierung auf, da Banachräume als Typklasse formalisiert sind. Weitere wesentliche Theoreme betreffen die Stetigkeit und Differenzierbarkeit (in $x_0$ und $t$) des Flusses $\phi$ und der Poincaré-Abbildung $P$.

## 4 Rigorose Numerik

*Rigorose* Numerik bedeutet, mit Mengen (anstatt mit näherungsweisen Werten) zu rechnen, welche garantieren, die anzunähernde Werte einzuschließen. Der klassische Ansatz zu rigoroser Numerik ist Intervallarithmetik. Aber prinzipiell können verschiedenste Datenstrukturen zur Darstellung der Einschließenden Mengen verwendet werden. Beliebte alternative Datenstrukturen sind „centered forms", „affine forms" oder Taylormodelle.

Eine zentrale Idee unseres Ansatzes ist das tiefe Einbetten arithmetische Ausdrücke, was es erlaubt viele Theoreme unabhängig von der später gewählten Datenstruktur zur Mengendarstellung zu beweisen. Für konkrete Berechnungen verwenden wir (beliebig genaue) Gleitkommazahlen mit expliziten Rundungsoperationen zur Effizienzsteigerung. Als Mengendarstellung konzentrieren wir uns auf „affine forms".

**Affine Arithmetik.** Ein Problem von klassischer Intervallarithmetik [MKC09] ist die Tatsache, dass Abhängigkeiten zwischen Variablen nicht berücksichtigt werden. Beispielsweise für $x \in [0;2]$, der Ausdruck $x - x$ wertet zu $[-2;2]$ in Intervallarithmetik aus, wohingegen das exakte Resultat mit dem Intervall $[0;0]$ darstellbar wäre.

Affine Arithmetik [dFS04] ist eine Erweiterung von Intervallarithmetik, die lineare Abhängigkeiten verfolgen kann. Die zugrundeliegende Datenstruktur ist eine formale Summe (eine „affine form") $A_0 + \sum_i \varepsilon_i A_i$ über formalen Parametern $\varepsilon_i$, welche über dem Intervall $[-1;1]$ interpretiert werden. Die Idee ist, dass die $\varepsilon_i$ symbolisch behandelt werden und dadurch lineare Abhängigkeiten vermitteln.

**Spezifikation and Verifikation des GDGL-Lösers.** Eine wichtige Einsicht bei der Verifikation rigoroser numerischer Algorithmen ist die Tatsache, dass *jede* Menge, die den echten Wert enthält zum Beweis von Korrektheit ausreicht.

Das gibt die Möglichkeit, über Implementierungsdetails zu abstrahieren: Einerseits über die konkrete Darstellung der einhüllenden Menge und andererseits der eigentliche Algorithmus der die Einhüllung berechnet.

Wir verwenden Lammichs [La13] *Autoref*-Framework zum automatischen Verfeinern nichtdeterministischer Spezifikationen. Dieses Framework stellt die Infrastruktur zu Verfügung, welche wir benutzen um abstrakte Spezifikationen (wie etwa „eine Einhüllung der Lösung einer GDGL") zu konkreten, ausführbaren Implementierungen (wie etwa ein rigoroses Runge-Kutte-Verfahren) zu Verfeinern.

Der offensichtliche Vorteil eines Frameworks wie Autoref ist, dass man die Korrektheit eines Algorithmus sehr bequem auf einer abstrakten Ebene verifizieren kann und sich nicht um Implementierungsdetails scheren muss.

## 5 Verifikation Algorithmischer Geometrie

Die von einer „affine form" dargestellte Menge ist ein Zonotop. Schnitt von Zonotopen mit Hyperebenen ist eine wichtige Operations, zum Beispiel um Poincaré-Abbildungen zu berechnen. Dieser Abschnitt behandelt die Verifikation eines Algorithmus von le Guernic und Girard [GLG08] um den Schnitt von Zonotopen und Hyperebenen zu überapproximieren.

Formale Verifikation geometrischer Algorithmen ist ein herausforderndes und interessantes Thema. Solche Algorithmen sind leicht mit einer geometrischen Intuition präsentiert, aber diese Intuition muss formal präzise gefasst werden.

### 5.1 Knuths Counterclockwise Prädikate

Der Kern des Schnittalgorithmus den wir verifizieren ist ähnlich zu Berechnungen der konvexen Hülle einer Punktmenge in der Ebene. Für letztere Algorithmen entwickelte

Knuth [Kn92] eine Theorie, welche den Begriff von Orientierung von Punkten in der Ebene axiomatisiert. Die Idee ist, dass wenn drei Punkte $p,q,r$ in der Ebene nacheinander besucht werden, ist eine Drehung entweder im oder gegen den Uhrzeigersinn nötig. Eine Drehung gegen den Uhrzeigersinn wird als ternäres Prädikat $pqr$ geschrieben, im Uhrzeigersinn die Negation $\neg pqr$.

Knuth beobachtete, dass einige wenige Eigenschaften der ternären Prädikate ausreichen um viele Konzepte in algorithmischer Geometrie formal zu fassen. Drei der fünf Eigenschaften sind in Abbildung 2 illustriert.

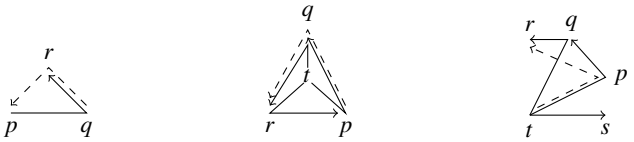
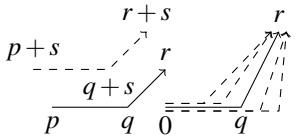



Abb. 2: Zyklische Symmetrie $pqr \longrightarrow qrp$ (links), Enthaltensein $tpq \wedge tqr \wedge trp \longrightarrow pqr$ (mitte), Transitivität $tsp \wedge tsq \wedge tsr \wedge tpq \wedge tqr \longrightarrow tpr$ (rechts); gestrichelte Prädikate werden von durchgezogenen impliziert.




Im Gegensatz zu Knuths Theorie, die auf endliche, diskrete Punktmengen abzielt, benötigt unsere Anwendung Mengen in kontinuierlichen Vektorräumen. Hierfür erweitern wir Knuths Theorie auf kontinuierliche Vektorräume. Zwei der zusätzlichen vier Eigenschaften sind beispielhaft links illustriert.

### 5.2 Verifikation von le Guernic and Girard's Algorithmus

Le Guernic and Girard's Algorithmus arbeitet für beliebig-dimensionale Zonotope. Der erste Schritt des Alorithmus reduziert das Problem auf mehrere zweidimensionale Probleme. Dieser Schritt ist leicht zu verifizieren. Der zweidimensionale Schnittalgorithmus ist aufwändiger: zunächst entledigen wir uns in einer Vorverarbeitung von kollinearen Punkten. Die eigentliche Verifikation ist wesentlich einfacher (da weniger Spezialfälle) wenn man zunächst nur das innere des Zonotops betrachtet und die Kanten vorerst ausschließt. Sie werden dann in einem letzten Stetigkeitsargument wieder mit in die Korrektheitsaussage aufgenommen.

## 6 Ein Verifizierter GDGL-Löser

Im Zentrum des verifizierten GDGL-Lösers steht eine Schleife zur Erreichbarkeitsanalyse. Gestartet auf einer Menge $X_0$ ist das Ziel, einen Poincaré-Schnitt $\Sigma$ zu erreichen. Vergleiche auch Abbildung 3a. Die Schleife verwaltet drei Arten von Mengen: $X$ ist die Menge deren zukünftige erreichbare Mengen noch weiter exploriert werden müssen. $C$ ist die Menge aller bisher explorierten Mengen und $I$ ist die Menge der Punkte für die die Erreichbarkeitsanalyse gestoppt hat, da sie den Poincaré-Schnitt $\Sigma$ erreicht haben.

Die Schleife operiert immer auf einem Teil von $X$ und entweder unterteilt diese Menge (dies erhält die Präzision aufrecht, sollte die Dynamik nicht-konvexe erreichbare Mengen erzeugen) oder wendet einen Schritt eines Runge-Kutta Verfahrens an, um die in einem Zeitschritt erreichbare nächste Menge herauszufinden. Die Runge-Kutta Schritte sind in affiner Arithmetik implementiert. Sollte ein solcher Runge-Kutta-Schritt den Poincaré-Schnitt $\Sigma$ erreichen, muss die zugehörige Poincaré-Abbildung berechnet werden. Dies geschieht mit einer geometrisch berechneten Überapproximation (wie in Abschnitt 5 beschrieben) der exakten Schnittmenge.

Runge-Kutta-Verfahren werden verifizert indem (dies ist Standard für die Herleitung der Konvergenz von Runge-Kutta-Verfahren) die Taylorreihenentwicklungen von Lösung mit der der numerischen Approximation verglichen werden. Für rigorose Methoden wird noch eine explizite Schranke *rk2-remainder* für die Restglieder benötigt.

**Lemma 6.1** (Runge-Kutta-Verfahren mit Fehlerschranke)**.** *Für* $0 < p \leq 1$ *und eine konvexe a-priori Schranke X für den Fluss* $\phi(x_0, [0;h]) \subseteq X$*:*

$$\phi(x_0, h) \in \mathit{rk2}_h(x_0) + \mathit{convex\text{-}hull}(\mathit{rk2\text{-}remainder}_h(x_0, p, [0;1], X))$$

# 7 Smale's 14th Problem

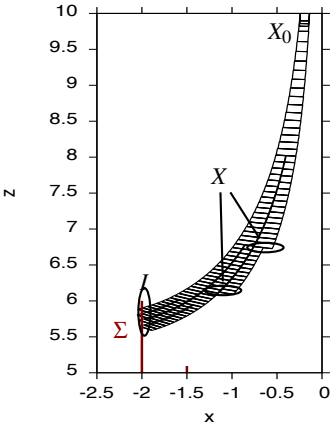
Im Jahr 1963, führte der Meteorologe Edward Lorenz [Lo63] das folgende System von GDGLen als vereinfachtes Modell für atmosphärische Dynamiken ein: $\dot{x} = -\sigma x + \sigma y, \dot{y} = -xz + \rho x - y, \dot{z} = xy - \beta z$ Lorenz stellte fest, dass selbst die kleinste Störung in Anfangsbedingungen zu komplett unterschiedlichem Langzeitverhalten des Systems führen würden. In Bezug auf seine ursprüngliche Motivation machte er den Begriff des Schmetterlingseffekts populär. Lorenz' System entwickelt sich zu einer komplizierten Struktur, dem sogenannten Lorenz-Attraktor (Abbildung 3b), welcher einen Kultstatus als Beispiel deterministischen Chaos genießt.

Trotz seiner Popularität und großem Aufwand der in seine Erforschung gesteckt wurde, konnte lange nicht bewiesen werden, dass der Lorenz-Attraktor in einem streng mathematischen Sinn chaotisch ist. Dies motivierte Fields-Medaillenträger Stephen Smale, den Lorenz-Attraktor auf seine Liste von achtzehn ungelösten mathematischen Problemen für das 21. Jahrhundert zu setzen [Sm98].

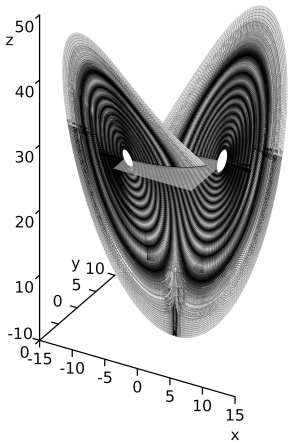
## 7.1 Verifikation von Tucker's Beweis

Tucker löste dieses Problem mit einem Computerbeweis. Er berechnete numerische Überapproximationen für einen Poincaré-Schnitt $\Sigma = [-6;6] \times [6;6] \times \{27\}$ im Lorenz-Attraktor (Abbildung 3b). Tucker identifiziert eine vorwärts invariante Region $N \subseteq \Sigma$, d.h., Lösungen die in $N$ starten, kehren wieder nach $N$ zurück. Zusätzlich berechnet Tucker die Ableitungen der zugehörigen Poincaré-Abbildung und zeigt, dass es ein (unter dem Bild der Ableitung) vorwärts invariantes Kegelfeld gibt, welches ausreichend stark expandiert. Dies

ist eine hinreichende Bedingung für Chaos. Der nicht-computergestützte Teil von Tuckers Beweis behandelt auch die stabile Mannigfaltigkeit $\Gamma$, deren relevanten Eigenschaften wir in unserem formalisierten Beweis annehmen. Zusammenfassend berechnet Tuckers Pro-




(a) Zustand während der Schleife zur Erreichbarkeitsanalyse.




(b) Numerische Schranken für den Lorenz Attraktor, formal verifiziert.

Abb. 3

gramm numerische Schranken für $N$, $P$, $\mathscr{E}$ und $\mathscr{E}^{-1}$, sodass folgende Eigenschaften gelten:

**Theorem 7.1** (Vorwärts Invariente Menge, Ableitungen, Kegel und Expansionen)**.**
*(1)* $\forall x \in N - \Gamma.\ P(x) \in N$
*(2)* $\forall x \in N - \Gamma.\ \forall v \in \mathfrak{C}(x).\ \mathsf{D}P|_x \cdot v \in \mathfrak{C}(P(x))$
*(3)* $\forall x \in N - \Gamma.\ \forall v \in \mathfrak{C}(x).\ \|\mathsf{D}P|_x \cdot v\| \geq \mathscr{E}(x)\,\|v\|$
*(4)* $\forall x \in N - \Gamma.\ \forall v \in \mathfrak{C}(x).\ \|\mathsf{D}P|_x \cdot v\| \geq \mathscr{E}^{-1}(P(x))\,\|v\|$

Die Originaldaten von Tucker sind online verfügbar[3] als eine Unterteilung von $N$, assoziiert mit Informationen über $P, \mathscr{E}, \mathscr{E}^{-1}$. Wir beweisen Theorem 7.1 formal, indem jeder Teil von $N$ mit dem verifizierten GDGL-Löser berechnet wird und überprüft wird, ob die assoziierten Schranken gültig sind.

## Literaturverzeichnis

[BLM16] Boldo, Sylvie; Lelay, Catherine; Melquiond, Guillaume: Formalization of real analysis: a survey of proof assistants and libraries. Mathematical Structures in Computer Science, 26(7):1196–1233, 2016.

[dFS04] de Figueiredo, Luiz Henrique; Stolfi, Jorge: Affine Arithmetic: Concepts and Applications. Numerical Algorithms, 37(1-4):147–158, 2004.

[GLG08] Girard, Antoine; Le Guernic, Colas: Zonotope/Hyperplane Intersection for Hybrid Systems Reachability Analysis. In (Egerstedt, Magnus; Mishra, Bud, Hrsg.): Hybrid Systems: Computation and Control, Jgg. 4981 in LNCS, S. 215–228. Springer, 2008.

[3] `http://www2.math.uu.se/~warwick/main/rodes/ResultFile`

[Go08] Gonthier, Georges: Formal proof–the four-color theorem. Notices of the AMS, 55(11):1382–1393, 2008.

[Ha15] Hales, Thomas; Adams, Mark; Bauer, Gertrud; Dang, Dat Tat; Harrison, John; Hoang, Truong Le; Kaliszyk, Cezary; Magron, Victor; McLaughlin, Sean; Nguyen, Thang Tat et al.: A formal proof of the Kepler conjecture. arXiv preprint arXiv:1501.02155, 2015.

[IH17] Immler, Fabian; Hölzl, Johannes: Ordinary Differential Equations. Archive of Formal Proofs, September 2017. `http://isa-afp.org/entries/Ordinary_Differential_Equations.shtml`, Formal proof development.

[Im18] Immler, Fabian: A Verified ODE Solver and Smale's 14th Problem. Dissertation, Technische Universität München, München, 2018.

[Kl09] Klein, Gerwin; Elphinstone, Kevin; Heiser, Gernot; Andronick, June; Cock, David; Derrin, Philip; Elkaduwe, Dhammika; Engelhardt, Kai; Kolanski, Rafal; Norrish, Michael et al.: seL4: Formal verification of an OS kernel. In: Proceedings of the ACM SIGOPS 22nd symposium on Operating systems principles. ACM, S. 207–220, 2009.

[Kn92] Knuth, Donald: Axioms and Hulls. Springer, Berlin New York, 1992. Number 606 in Lecture Notes in Computer Science.

[La13] Lammich, Peter: Automatic data refinement. In: International Conference on Interactive Theorem Proving. Springer, S. 84–99, 2013.

[Le09] Leroy, Xavier: Formal verification of a realistic compiler. Communications of the ACM, 52(7):107–115, 2009.

[Lo63] Lorenz, Edward N.: Deterministic Nonperiodic Flow. Journal of the Atmospheric Sciences, 20(2):130–141, 1963.

[MKC09] Moore, Ramon E; Kearfott, R Baker; Cloud, Michael J: Introduction to interval analysis. SIAM, 2009.

[NPW02] Nipkow, Tobias; Paulson, Lawrence C.; Wenzel, Markus: Isabelle/HOL: A proof assistant for higher-order logic. LNCS. Springer, 2002.

[Pa89] Paulson, Lawrence C.: The foundation of a generic theorem prover. Journal of Automated Reasoning, 5(3):363–397, 1989.

[Sm98] Smale, Steve: Mathematical problems for the next century. The Mathematical Intelligencer, 20(2):7–15, 1998.

[Tu02] Tucker, Warwick: A Rigorous ODE Solver and Smale's 14th Problem. Foundations of Computational Mathematics, 2(1):53–117, 2002.

**Fabian Immler**, Jahrgang 1987, studierte Informatik (B.Sc. und M.Sc) von 2007 bis 2012 and der Technischen Universität München. Dort promovierte er [Im18] am Lehrstuhl für Logik und Verifikation, mit Forschungsaufenthalten bei Warwick Tucker's Gruppe "Computer Aided Proofs in Analysis" an der Uppsala University in Schweden. Seine Dissertation verteidigte er 2018 und erhielt dafür den „Heinz-Schwärtzel-Dissertationspreis für Grundlagen der Informatik". Die von ihm entwickelten und formal verifizierten Algorithmen erhielten den von Bosch gesponsorten „ARCH 2019 Best Result Award".

# Statistische Appearance-Modelle basierend auf probabilistischen Korrespondenzen für die medizinische Bildanalyse[1]

Julia Krüger[2]

## 1 Einführung

Die Segmentierung medizinischer Bilder ist ein zentrales Problem in der Analyse und Bearbeitung medizinischer Bilddaten. Sie bildet meist die Grundlage für weitergehende Verfahren zur computergestützten Diagnostik und Therapie. Automatische robuste Segmentierungsverfahren werden unter anderem benötig bei der quantitativen Radiologie – automatische Anatomie-/Pathologiedetektion mit anschließender Volumenbestimmung, Strahlentherapie – Identifikation von betroffenem und gesundem Gewebe – oder computergestützten Chirurgie – Planung von chirurgischen Eingriffen mittels 3D Modellen der betroffenen Region/Organe.

Eine automatische Segmentierung der Daten gewährleistet dabei im Gegensatz zur manuellen Segmentierung stabilere und reproduzierbare Ergebnisse. Des Weiteren kann eine genaue manuelle Segmentierung der gesuchten Organe/Strukturen mit einem enormen Zeitaufwand verbunden sein.

Die größten Herausforderungen für automatische Segmentierungsverfahren bilden zum einen die hohe anatomische (und pathologische) Variabilität zwischen Patienten und zum anderen die schwankende Bildqualität der Daten. Bei der Entwicklung solcher Verfahren muss berücksichtigt werden, dass die Form und Erscheinung der gesuchten Objekte (Organe) zwischen Patienten oder sogar zwischen zeitlich auseinanderliegenden Aufnahmen desselben Patienten stark schwanken können. Abgesehen von unterschiedlicher Bildauflösung, verschiedener Positionierung der Patienten oder abweichenden Akquirierungsparametern leiden medizinische Bilddaten außerdem unter Rauschen oder Artefakten. Einfachere Segmentierungsverfahren wie rein signalwertbasierte oder gradientenbasierte Methoden stoßen daher schnell an ihre Grenzen, wenn es sich um schlecht definierte Organgrenzen handelt.

Die Idee der modellbasierten Segmentierungsverfahren ist die Nutzung von Vorwissen über das gesuchte Objekt. Einfache Methoden gehen dabei lediglich von einer z. B. “glatten Konturgrenze” als Modellannahme aus, wohingegen komplexere Modelle die Form oder Erscheinung eines betrachteten Organs modellieren. Die Herausforderung bei der

[1] Titel engl.: Statistical appearance models based on probabilistic correspondences for medical image analysis

[2] Institut für Medizinische Informatik der Universität zu Lübeck, Julia.Krueger.SN@gmail.com

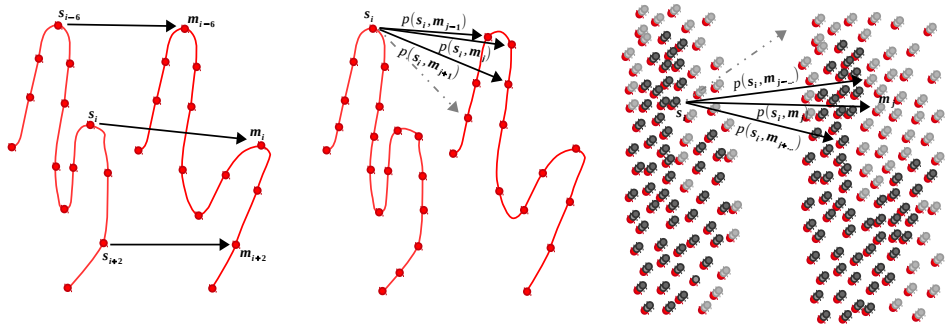



Abb. 1: Klassische Modellansätze gehen von einer Eins-zu-Eins-Korrespondenz zwischen “exakt” lokalisierten Landmarken aus. Probabilistische Formmodelle berechnen stattdessen die Korrespondenzwahrscheinlichkeiten zwischen “potentiell ungenau” bestimmten Merkmalsvektoren. Bei den probabilistischen Appearance-Modellen wird die Korrespondenzschätzung um die Appearance-Merkmale erweitert und des Weiteren wird von einer formunabhängigen Positionierung der Merkmalsvektoren ausgegangen.

Generierung dieses Vorwissens ist die heterogene Natur von medizinischen Strukturen, die oftmals eine hohe Variabilität in Form und Erscheinung aufweisen: Ungeachtet der Tatsache, dass ein bestimmtes Organ einem generellen “Bauplan” folgt, unterscheiden sich die individuelle Form und Erscheinung jedes Patienten voneinander. Um dieses Problem der “Individualität versus Generalität” einer gewissen anatomischen Struktur zu adressieren, können statistische Modelle eingesetzt werden [HM09]. Diese wurden entwickelt, um die statistischen Eigenschaften einer Gruppe von ähnlichen Objekten (Organe, Knochen,...) zu beschreiben. Der Vorteil eines modellbasierten Verfahrens liegt in seiner Robustheit: Wird ein solches Modell, welches das gesuchte Organ mit allen möglichen Variationen beschreibt, auf ein neues Bild adaptiert, ist gewährleistet, dass die Ergebnissegmentierung einer Instanz dieses Objekts entspricht.

Die Grundlage eines jeden statistischen Modells stellt eine Trainingsdatenmenge von $N$ Datensätzen dar, welche das zu modellierende Objekt beschreibt: $\mathbf{S_{tr}} = \{\mathbf{S}_k | k = 1, ..., N\}$. In den klassischen Methoden werden die Datensätze als Sequenz von $N_s$ Landmarken beschrieben: $\mathbf{S}_k = [\mathbf{s}_{k,1}, \mathbf{s}_{k,2} ..., \mathbf{s}_{k,N_s}]$. Jede Landmarke $\mathbf{s}_{k,i} \in \mathbb{R}^D$ repräsentiert dabei eine definierte Position im Objekt. Alle Landmarken $\mathbf{s}_{1,i}, \mathbf{s}_{2,i}, ..., \mathbf{s}_{k,i}, ...\mathbf{s}_{N,i}$ mit demselben Index $i$ in allen Trainingsdaten werden als korrespondierend angenommen. Cootes et al. führte den Begriff der *point distribution models* für diese Repräsentation mittels korrespondierender Landmarken ein [Co92]. Neben der Beschreibung der Form (durch Landmarken) eines Organs kann ebenfalls die Erscheinung (engl. appearance) der betrachteten Struktur mit modelliert werden. Diese “Appearance”-Informationen werden in klassischen Modellen abhängig von der durch Landmarken repräsentierten Objektform abgetastet [CET01].

Da mithilfe dieser Informationen die statistischen Form- bzw. Appearance-Eigenschaften des Objekts bestimmt werden, hängt die Qualität des generierten Modells vollständig von

einer korrekten Positionierung der Landmarken ab. Allerdings stellt sich dabei die Frage, ob die exakte anatomische Landmarke in allen Bildern an gut definierten Positionen existiert – auf Grund von anatomischen (oder pathologischen) Unterschieden zwischen Patienten oder auf Grund von unterschiedlichen Bildakquirierungstechniken. Dieselbe Problematik ergibt sich bei der Adaption des Modells an neue unbekannte Bilder, da hier ebenfalls von einer Eins-zu-Eins-Korrespondenz zwischen Modell und neuem Bild ausgegangen wird.

Um die grundlegende Annahme der Eins-zu-Eins-Korrespondenzen in klassischen Modellen zu ersetzen, können probabilistische Korrespondenzen eingesetzt werden [Hu08]. Diese Korrespondenzwahrscheinlichkeiten bieten Robustheit gegenüber "ungenau" bestimmten "Landmarken" sowie gegenüber fehlenden Strukturen in Bilddaten.

Der zweite wichtige Aspekt – neben den korrekten Landmarken – der die Qualität der modellbasierten Anwendungen ausschlaggebend bestimmt, ist die Art der Anpassung des Modells an neue Bilddaten. Ist die Modellgenerierung in den meisten Fällen durch ein gut definiertes Verfahren abgedeckt, so müssen hingegen für die Modellanpassung unabhängige Methoden entwickelt werden, da zuerst zu definieren ist, wie eine "gute Anpassung" bestimmt ist.

Die hier zusammengefasste Arbeit präsentiert einen neuen vielseitig einsetzbaren probabilistischen Maximum-A-posteriori-Ansatz für statistische Appearance-Modelle [KEH17, KEH15]. Die Grundlage der vorangestellten Problemformulierung bildet eine merkmalsvektorbasierte Repräsentation von Bildern durch eine Menge unabhängiger $D$-dimensionaler Merkmalsvektoren $\mathbf{S} = \{\mathbf{s}_i = (\mathbf{x}_i, \mathbf{f}_i) | i = 1, \ldots, N_s\}$, welche Positions- ($\mathbf{x}_i \in \Omega$) und beliebige Appearance-Informationen ($\mathbf{f}_i$) kombinieren. Im Kontrast zu klassischen statistischen Form- oder Appearance-Modellen nutzt das probabilistische Modell Korrespondenzwahrscheinlichkeiten, wodurch eine vorangehende, potentiell aufwendige Definition von Eins-zu-Eins-Korrespondenzen unnötig wird. Dies eliminiert den Bedarf einer elaborierten Vorverarbeitung der Trainingsdaten und unterstützt außerdem ein arbiträres Abtasten der Test- und Trainingsdaten. Ein abgeleitetes globales Optimierungskriterium wird sowohl für die Modellgenerierung als auch -anpassung genutzt. Für die Optimierung der gesuchten Parameter über das globale Kriterium wird ein iterativer Expectation-Maximization-Algorithmus angewendet, um abwechselnd Modellparameter und probabilistische Korrespondenzen zwischen den Trainingsinstanzen und dem Modell zu bestimmen. Für die Steigerung der Robustheit während der Modelloptimierung wurden verschiedene Multi-Level-Erweiterungen vorgestellt.

## 2 Methode

### 2.1 Korrespondenzwahrscheinlichkeiten zwischen Bildrepräsentationen

Es seien zwei Bildrepräsentationen $\mathbf{S} = \{\mathbf{s}_i | i = 1, \ldots, N_s\}$ und $\mathbf{M} = \{\mathbf{m}_j | j = 1, \ldots, N_m\}$ gegeben. Es sei unbekannt, welche Merkmalsvektoren beider Mengen miteinander korrespondieren. Des Weiteren wird angenommen, dass die Positionierung der Merkmals-

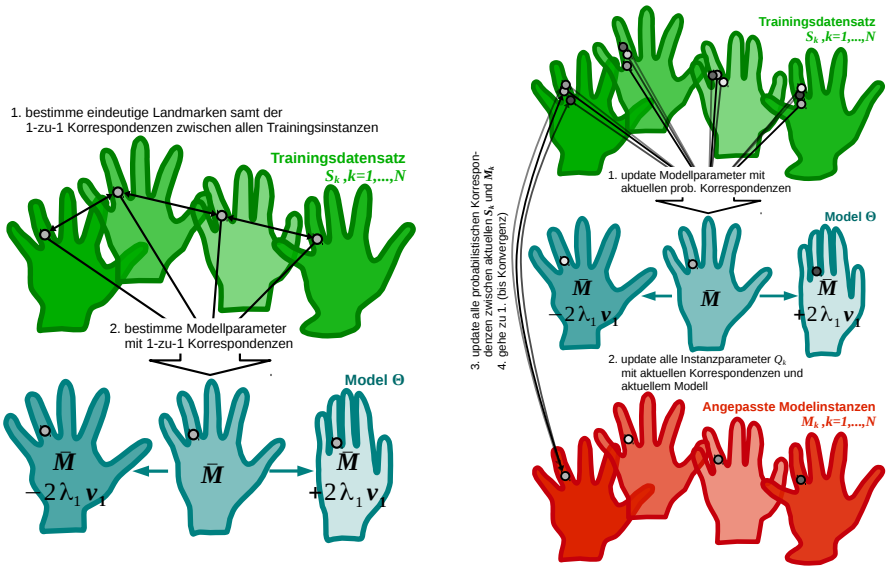



Abb. 2: Links:Modellgenerierung mit bekannten Eins-zu-Eins-Korrespondenzen: In klassischen Modellen sind die eindeutigen Korrespondenzen zwischen den Trainingsinstanzen und somit auch zum resultierenden Modell bekannt. Es können also in einem Schritt das mittlere Modell $\bar{\mathbf{M}}$ und die Variationsmoden $\mathbf{V}$ (mittels PCA) bestimmt werden.
Rechts: Modellgenerierung ohne bekannte Eins-zu-Eins-Korrespondenzen unter Nutzung von Korrespondenzwahrscheinlichkeiten: Das Modell kann nicht in einem Schritt berechnet werden, da die optimalen Modellparameter abhängig von den aktuellen probabilistischen Korrespondenzen sind. Um diese bestimmen zu können, werden nicht nur neue Modellparameter berechnet (1.), sondern ebenfalls das aktuelle Modell auf die Trainingsinstanzen angepasst (2.). Anschließend können die Korrespondenzwahrscheinlichkeiten zwischen $\mathbf{S}_k$ und Modell bzw. $\mathbf{M}_k$ aktualisiert werden (3.). Die Optimierung des Modells wird iterativ fortgesetzt bis zur Konvergenz.

vektoren “ungenau” ist und daher keine Eins-zu-Eins-Korrespondenz zwischen den Mengen vorliegt. Daher sind die Korrespondenzwahrscheinlichkeiten zwischen beiden Repräsentationen über die paarweisen Wahrscheinlichkeiten so definiert, dass Merkmalsvektor $\mathbf{s}_i$ aus $\mathbf{S}$ eine verrauschte Observation des Merkmalsvektors $\mathbf{m}_j$ aus $\mathbf{M}$ ist. Dies wird nach Granger et al. [GP02] als Gauß-verteilt modelliert:

$$p(\mathbf{s}_i|\mathbf{m}_j) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{s}_i-\mathbf{m}_j)^T\Sigma^{-1}(\mathbf{s}_i-\mathbf{m}_j)\right), \tag{1}$$

wobei $\Sigma \in \mathbb{R}^{D\times D}$ die Kovarianzmatrix repräsentiert und sowohl Positions- als auch Appearance-Merkmalsvarianzen beschreibt. Die Korrespondenzwahrscheinlichkeit von $\mathbf{S}$ zu $\mathbf{M}$ kann zusammengefasst werden als: $p(\mathbf{S}|\mathbf{M}) = \prod_{i=1}^{N_s}\frac{1}{N_m}\sum_{j=1}^{N_m} p(\mathbf{s}_i|\mathbf{m}_j)$. $p(\mathbf{S}|\mathbf{M})$ kann demnach als Maß betrachtet werden, welches anzeigt, wie stark beide Bildrepräsentationen miteinander korrespondieren. Abbildung 1 stellt die klassische Eins-zu-Eins-Korrespondenz der probabilistischen Korrespondenzdefinition visuell gegenüber.

### 2.2 Modell auf Basis von zufälligen Observationen

Liegt eine Trainingsmenge von $N$ Bildern samt dazugehörigen Bildrepräsentationen $\mathbf{S_{tr}} = \{\mathbf{S}_k | k = 1,\ldots,N\}$ vor, wird das Modell, welches diese Daten optimal abbildet, gesucht. Die probabilistische Sichtweise betrachtet die gegebenen Daten als Menge von Observationen, auf dessen Basis gewisse Populationsparameter – in diesem Fall "das Modell" – geschätzt werden sollen. Das heißt, folgende Wahrscheinlichkeit wird maximiert: $Modell = \text{argmax}_{Modell} \;\; p(Modell|\mathbf{S_{tr}})$. Die Trainingsobservationen werden als zufällig durch das Modell generiert angenommen. Wie bei klassischen Ansätzen setzt sich das Modell aus einem mittleren Modell $\bar{\mathbf{M}}$ und einer Anzahl an $n$ Variationsmoden $\mathbf{V} = \{\mathbf{V}_p | p = 1,\ldots,n\}$ mit Standardabweichung $\{\lambda_p | p = 1,\ldots,n\}$ zusammen. Laut Modellannahme werden neue Instanzen $\mathbf{M}$ durch ein solches Modell generiert, indem die Moden mittels Gewichte $\omega_p$ manipuliert werden. Größere lineare Anpassungen werden durch eine affine Transformation $T$ durchgeführt: $\mathbf{M} = T^{-1}(\bar{\mathbf{M}} + \sum_{p=1}^{n} \omega_p \mathbf{V}_p)$ mit $\omega_p \sim N(0, \lambda_p)$. $\omega_p$ wird als normalverteilt mit der Standardabweichung $\lambda_p$ angenommen.

Um die optimalen Parameter zu finden, werden nun diejenigen Modellinstanzen $\mathbf{M}_k$ gesucht, welche die gegebene Trainingsdaten $\mathbf{S}_k$ am besten beschreiben $\forall k$: $\mathbf{S}_k \approx \mathbf{M}_k$ mit $\mathbf{M}_k = {T_k}^{-1}(\bar{\mathbf{M}} + \sum_{p=1}^{n} \omega_{k,p} \mathbf{V}_p)$ Um zu bestimmen, ob zwei Merkmalsvektormengen ($\mathbf{S}_k$ und $\mathbf{M}_k$) einander ähnlich sind, müssen Korrespondenzen zwischen beiden bekannt sein. Diese sind zu Beginn der Modellgenerierung nicht bekannt und müssen gemeinsam mit den Parametern optimiert werden. Die probabilistischen Korrespondenzen werden über die Gleichung 1 berechnet. Dabei bestimmen die Merkmalsvarianzen $\Sigma$, welche Vektoren in beiden Mengen wie stark korrespondieren. Die Abbildungen 2 stellen den Unterschied zwischen der klassischen Modellgenerierung in einem Schritt mit bekannten Korrespondenzen (Abbildung 2, links) und der iterativen probabilistischen Modellgenerierung über die gemeinsame Optimierung der Modellinstanzen und Korrespondenzen (Abbildung 2, rechts) optisch dar.

Alle Parameter $\{\bar{\mathbf{M}}, \mathbf{V}_p, \lambda_p, \omega_{k,p}, T_k, n, \Sigma |\; p = 1,\ldots,n, k = 1,\ldots,N\}$ können in *Modellparameter* und *instanzabhängige Parameter* unterteilt werden. Als Modellparameter $\Theta$ werden diejenigen Parameter betrachtet, die für den gesamten Trainingsdatensatz und somit optimalerweise für die gesamte Population gleich sind: $\Theta = \{\bar{\mathbf{M}}, \mathbf{V}_p, \lambda_p, n, \Sigma |\; p = 1,\ldots,n\}$. Als instanzabhängige Parameter $Q = \{Q_k | k = 1,\ldots,N\}$ werden die Parameter definiert, welche die Modellinstanz modifizieren: $Q_k = \{T_k, \omega_{k,p} |\; p = 1,\ldots,n\}$.

### 2.3 Maximum-A-posteriori-Ansatz zur Parameteroptimierung:

Das aufgestellte Problem, diejenigen Modellinstanzen $\mathbf{M}_k$ zu finden, welche die gegebene Trainingsdaten $\mathbf{S}_k$ am besten beschreiben $\forall k$ $\mathbf{S}_k \approx \mathbf{M}_k$ mit $\mathbf{M}_k = {T_k}^{-1}(\bar{\mathbf{M}} + \sum_{p=1}^{n} \omega_{k,p} \mathbf{V}_p)$ kann als Maximum-A-posteriori-Schätzung der Parameter $\Theta$ und $Q = \{Q_k | k = 1,\ldots,N\}$ definiert werden: $p(Q, \Theta|\mathbf{S_{tr}}) \rightarrow \max_{Q,\Theta}$. Unter Nutzung des Bayes-Theorems wird folgendes Kriterium minimiert:

$$C(Q,\Theta) = -\log p(Q,\Theta|\mathbf{S_{tr}}) = -\log\left(\frac{p(\mathbf{S_{tr}}|Q,\Theta)p(Q|\Theta)p(\Theta)}{p(\mathbf{S_{tr}})}\right). \qquad (2)$$

### 2.4 Modellgenerierung und Modellanpassung:

Als *Modellgenerierung* wird die Optimierung des globalen Kriteriums nach $\Theta$ bezeichnet. Durch die unbekannten Korrespondenzen zwischen Trainingsdatensatz und Modell müssen darüber hinaus gleichzeitig die instanzabhängigen Parameter $Q_k$ optimiert werden. Das heißt, die Modellgenerierung kann zusammengefasst werden als:

**- gegeben:** $N$ Trainingsinstanzen $\mathbf{S_{tr}} = \{\mathbf{S}_k | k = 1, \ldots, N\}$

**- gesucht:** das Modell $\Theta$ – mithilfe der Optimierung von $\text{argmin}_{Q,\Theta} C(Q,\Theta) = \text{argmin}_{Q,\Theta} - \log p(Q,\Theta|\mathbf{S_{tr}})$.

Als *Modellanpassung* wird die Optimierung desselben globalen Kriteriums nach $Q_{new}$ mit fixer Parameterbelegung für $\Theta$ bezeichnet. Dies kann zusammengefasst werden als:

**- gegeben:** das Modell $\Theta$ und eine Instanz $\mathbf{S}_{new}$

**- gesucht:** die Modellinstanz $\mathbf{M}_{new}$, die über die instanzabhängigen Parameter $Q_{new} = \{T_{new}, \boldsymbol{\omega}_{new}\}$ definiert ist und $\mathbf{S}_{new}$ "am ähnlichsten" ist – mithilfe der Optimierung von $\text{argmin}_{Q_{new}} C(Q_{new},\Theta) = \text{argmin}_{Q_{new}} - \log p(Q_{new},\Theta|\mathbf{S}_{new})$.

Durch die Optimierung des globalen Kriteriums $C(Q,\Theta)$ sollen Modell- und instanzabhängige Parameter bestimmt werden, welche die gegebene Trainingsmenge abbilden können, $\forall k$: $\mathbf{S}_k \approx \mathbf{M}_k$ mit $\mathbf{M}_k = {T_k}^{-1}(\bar{\mathbf{M}} + \sum_{p=1}^{n} \omega_{k,p} \mathbf{V}_p)$

Die Optimierung des Kriteriums nach den Parametern erfolgt iterative mit Hilfe eines Expectation-Maximization-Algorithmus. Dabei wird zwischen einem E(xpectation)-Schritt, indem die Korrespondenzwahrscheinlichkeiten für die aktuellen Parameter bestimmt werden, und einem M(aximization)-Schritt, indem die Parameter für die aktuellen Korrespondenzen optimiert werden, alterniert. Die Optimierung im M-Schritt erfolgt dabei über die partiellen Ableitungen des Kriteriums nach allen Parametern.

## 3 Zusammenfassung der Vorteile des vorgestellten Frameworks

Die Nutzung der **probabilistischen Korrespondenzen** liefert eine Vielzahl an Vorteilen im Gegensatz zu vorher aufwendig bestimmten Eins-zu-Eins-Korrespondenzen:

- Eine kostenaufwendige Bestimmung von Landmarken und Eins-zu-Eins-Korrespondenzen vor der Modellgenerierung ist nicht notwendig. Außerdem ist die Qualität des resultierenden Modells nicht von potentiell falsch gesetzten Landmarken abhängig.
- Die probabilistischen Korrespondenzen führen zu einer gesteigerten Robustheit gegenüber nicht vorhandener korrespondierender Strukturen in den Bilddaten.
- Außerdem sind zusätzliche Informationen über lokale Unsicherheiten des Modells durch probabilistische Korrespondenzen nach abgeschlossener Modellanpassung gegeben. Es werden Bereiche (mit geringen Korrespondenzwerten) in neuen Bilddaten

indiziert, die nicht oder "schlecht" durch das Modell abgebildet werden konnten. Diese können beispielsweise automatisch auf pathologische Bereiche in den Bilddaten hinweisen.

Im Gegensatz zu klassischen Verfahren, die zu meist auf einer Vielzahl an verschiedenen Algorithmen bestehen, basiert die vorgestellte Methode auf einer **geschlossenen mathematischen Formulierung** eines globalen Optimierungskriteriums:

- Die Qualität des Verfahrens ist nicht abhängig von Ergebnissen von einzelnen unabhängigen aneinandergereihten Verfahren.
- Sowohl die Modellgenerierung und als auch die Modellanpassung werden mit demselben Optimierungsverfahren durchgeführt. Im Gegensatz dazu stellt bei klassischen Modellansätzen die Modellanpassung oftmals ein komplett neues und sogar schwereres Problem als die Modellgenerierung dar.
- Alle Parameter werden entweder über das Kriterium optimiert oder datengetrieben an die aktuelle Situation während der Optimierung angepasst. Es findet kein manuelles Tuning von Parametern statt.

Die Grundlage der Modellformulierung bildet eine Datenrepräsentation über eine **gemeinsame Definition von Positions- und Appearance-Merkmalen** in einer Merkmalsvektormenge:

- Durch die Repräsentation der Merkmalsvektoren als beliebig lange Vektoren, bei denen Positionsmerkmale und Appearance- oder Segmentierungsmerkmale gleich behandelt werden, gibt es keinerlei Einschränkungen die Art oder Anzahl der Merkmale betreffend.
- Es können automatisch "beliebigdimensionale" Bilder verarbeitet werden: Ob 2D, 3D oder 4D (3D+Zeit), ob mehrkanal Bilder (verschiedene MRT-Sequenzen) oder multimodale Bilddaten (MRT und CT Daten), ob eine Segmentierungsmaske oder Multi-Organsegmentierung – es muss keinerlei Anpassung an die vorgestellte Methode erfolgen.
- Durch die zusätzliche Anpassung der Appearance-Werte während der Modellgenerierung ist ebenfalls keine Vorverarbeitung der Signalwerte (Histogrammangleichung, etc.) der Trainingsdaten notwendig.
- Außerdem ist die Anwendung des Ansatzes nicht auf die Segmentierungsproblematik beschränkt.

Der Modellgenerierungsansatz wurde zu einer **probabilistischen Registrierung zweier Bilder** vereinfacht, wobei eine solche Registrierung folgende Vorteile liefert:

- Im Gegensatz zur klassischen Registrierung weist die probabilistische Registrierung durch die Nutzung der Korrespondenzwahrscheinlichkeiten eine hohe Robustheit

gegenüber pathologischen Bereichen, die nicht in beiden Bildern vorhanden sind, auf.

- Solche Bereiche werden darüber hinaus automatisch durch geringe Korrespondenzwerte nach der Registrierung indiziert und können daher zusätzlich segmentiert werden.
- Durch die "beliebigdimensionale Formulierung" der Bildrepräsentationen und der Deformation können im Gegensatz zur klassischen Registrierung nicht nur die Verschiebungsfelder der Bildpositionen bestimmt werden, sondern ebenfalls die Signalwertunterschiede beider Bilddaten ausgeglichen werden.

Im Gegensatz zum klassischen Form- oder Appearance-Modell, das vor allem zur Segmentierung des modellierten Objekts eingesetzt wird, liefert der vorgestellte Modellansatz eine Vielzahl an Anwendungsmöglichkeiten – zum einen durch die flexible Formulierung der Bildrepräsentation mit beliebigdimensionalen Merkmalsvektoren und zum anderen durch die Nutzung der probabilistischen Korrespondenzen. Anwendungsszenarien umfassen unter anderem: Segmentierung von Objektkonturen, Detektion von Landmarken, Multi-Objektsegmentierung bzw. Übertragung von beliebig vielen Labeln auf ein neues Bild, Rekonstruktion von fehlenden Merkmalen (z.B. fehlende Sequenz in MRT-Bildern). robuste Modellanpassung bei fehlenden Korrespondenzen bzw. "fehlerhaften Bereichen" im neuen Bild, Identifikation von lokalen Unsicherheiten des Modells, Klassifikation von pathologischen Regionen oder Rekonstruktion von "fehlerhaften Bereichen" in Bildern.

Neben solchen Modellanwendungen, welche Vorwissen über das betrachtete Objekt in einem Modell speichern und dieses Vorwissen als Grundlage für die Analyse neuer Bilddaten nutzen, wurde außerdem die Anwendung des beschriebenen mathematischen Frameworks zur probabilistischen Registrierung zweier Bilder vorgestellt.

## 4 Diskussion der Ergebnisse

Die Evaluierung des probabilistischen Modellansatzes und der probabilistischen Registrierung in der hier zusammengefassten Arbeit hat vor allem die Intention, die Möglichkeiten der Methode aufzuzeigen und nicht einzelne Problemstellungen so genau wie möglich zu lösen. Aufgrund dessen wurden die genutzten Daten nicht vorverarbeitet – eine Histogrammangleichung würde z. B. die Variabilität der Appearance-Werte in jedem Trainingsdatensatz verringern.

### 4.1 Segmentierung der Hand:

Die Segmentierung der 2D Röntgenbilder der Handdaten demonstriert, dass der Modellansatz in der Lage ist, relativ flexible und stark variierende Handstellungen zu modellieren. Außerdem wird zum einen die Segmentierung einer Objektkontur (Handkontur) und zum anderen eine Landmarkenbestimmung (der Fingerknochen) demonstriert. Da die Genauigkeit der Ergebnisse ($\leq 1mm$) wesentlich kleiner ist als der Abstand zwischen den Merkmalsvektoren ($3mm$), kann geschlussfolgert werden, dass die Modellierung der Kontur

bzw. der Landmarken über die Werte einer Distanzkarte eine ausreichend große Genauigkeit bereitstellt. Die Segmentierung von unbekannten Bildern zeigt, dass das generierte Modell erfolgreich an unbekannte Daten angepasst und dass fehlende Merkmale (Distanzkartenwerte) für die Bilder rekonstruiert werden können. Außerdem wurde gezeigt, dass die Initialisierung der Merkmalsvektoren für ein neues Bild über die Reduktion der Vektoren pro Iteration stabile Ergebnisse liefert. Eine weitere nachgeschaltete probabilistische Registrierung zwischen Bild und angepasstem Modell als Nachbearbeitungsschritt liefert die Möglichkeit, die Ergebnisse zu verbessern.

### 4.2 Klassifikation von 3D Gehirndaten:

Das 3D Gehirnmodell über die 3D MRT-Daten wurde genutzt, um die Modellgenerierung mit partiellen Trainingsdaten zu demonstrieren. Darüber hinaus wurde das generierte "gesunde" Modell eingesetzt, um pathologische (Schlaganfall-) Regionen in neuen Bildern mittels resultierenden Korrespondenzwerten zu identifizieren. Durch die 3D Daten erhöht sich die Anzahl der Merkmalsvektoren stark und durch die Berechnung der Korrespondenzwahrscheinlichkeiten (potentiell zwischen allen möglichen Merkmalsvektorpaaren), welche in jeder Iteration durchgeführt wird, steigt der Rechenaufwand quadratisch mit der Anzahl der Merkmalsvektoren. Das heißt, um den Rechen- und Speicherplatzaufwand für die Evaluierung gering zu halten, wurde ein relativ grober Abstand zwischen den Merkmalsvektoren von $5mm$ gewählt. Dies spiegelt sich dementsprechend in den Ergebnissen wider: große Läsionen ($> 10cm^3$) können gut durch die vorgestellte Methode segmentiert werden, wohingegen kleine Läsionen nicht gefunden werden, da das Modell zu grob ist. Außerdem wurde ein relativ einfaches automatisches Schwellwertverfahren zur Aufteilung in "gute" und "schlechte" Korrespondenzwerte genutzt. Der Vergleich mit den manuell gewählten Schwellwerten zeigt, dass an dieser Stelle eine Verbesserung durch aufwendigere Verfahren möglich ist. Die Evaluierung zeigt jedoch, dass aus den resultierenden Korrespondenzwerten eindeutig Informationen über "nichtmodellierbare" – in diesem Fall pathologische – Regionen extrahiert werden können. Für genauere Ergebnisse (vor allem für kleine Läsionen) muss ein feineres Modell generiert werden. Dies erfordert ebenfalls einen größeren Trainingsdatensatz.

### 4.3 Registrierung von 2D Gehirndaten:

Das probabilistische Appearance-adaptierende Registrierungsverfahren wurde an 2D-MRT-Bildern des Gehirns getestet. Verglichen wurden die Ergebnisse mit einem gut etablierten nichtrigiden Registrierungsverfahren. Die Ergebnisse zeigen, dass die probabilistische Registrierung ohne Appearance-Adaption bereits eine signifikant bessere Performanz zeigt als der klassische Ansatz. Die zusätzliche Anpassung der Appearance-Werte während der Deformationsoptimierung stellt eine weitere eindeutige Verbesserung da. Daraus lässt sich überdies verallgemeinern, dass das vorgestellte Verfahren der iterativen Korrespondenzwahrscheinlichkeitsbestimmung im Vergleich zur registrierungsbasierten Korrespondenz-/Landmarkenbestimmung gute Ergebnisse liefert. Wie beim Modellansatz können hier

ebenfalls beliebig viele Merkmale in der Registrierung genutzt werden. Die Anpassung der Appearance-Werte wird angewendet, um die Korrespondenzen genauer bestimmen zu können, da so "störende" Appearance-Unterschiede ausgeglichen werden. Außerdem kann diese Anpassung eingesetzt werden, um die Signalwerte der Bilder einander anzugleichen (mit oder ohne Positionsdeformation). Bei der evaluierten Anwendung war dies nicht von Bedeutung, da die anhand der Signalwerte optimierte Positionsdeformation auf Labelbilder übertragen wurde.

## Literaturverzeichnis

[CET01] Cootes, T. F.; Edwards, G. J.; Taylor, C. J.: Active Appearance Models. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 23(6):681–685, Juni 2001.

[Co92] Cootes, T. F.; Taylor, C. J.; Cooper, D. H.; Graham, J.: Training Models of Shape from Sets of Examples. In: Proceedings of the British Machine Vision Conference. BMVA Press, S. 2.1–2.10, 1992. doi:10.5244/C.6.2.

[GP02] Granger, Sébastien; Pennec, Xavier: Multi-scale EM-ICP: A Fast and Robust Approach for Surface Registration. In: Computer Vision - ECCV 2002, Jgg. 2353 in Lecture Notes in Computer Science, S. 418–432. Springer, 2002.

[HM09] Heimann, Tobias; Meinzer, Hans-Peter: Statistical Shape Models for 3D Medical Image Segmentation: A Review. Medical Image Analysis, 13(4):543 – 563, 2009.

[Hu08] Hufnagel, Heike; Pennec, Xavier; Ehrhardt, Jan; Ayache, Nicholas; Handels, Heinz: Generation of a Statistical Shape Model with Probabilistic Point Correspondences and the Expectation Maximization-Iterative Closest Point Algorithm. International Journal of Computer Assisted Radiology and Surgery, 2:265 – 273, 3/2008 2008.

[KEH15] Krüger, Julia; Ehrhardt, Jan; Handels, Heinz: Probabilistic Appearance Models for Segmentation and Classification. In: The IEEE International Conference on Computer Vision (ICCV). S. 1698–1706, Dez 2015.

[KEH17] Krüger, Julia; Ehrhardt, Jan; Handels, Heinz: Statistical Appearance Models based on Probabilistic Correspondences. Medical Image Analysis, 37:146 – 159, 2017.

**Julia Krüger** wurde geboren am 6. Juli 1986, in Schwerin. Das Informatik Studium (B. Sc., M. Sc.) mit dem vertiefenden Nebenfach „Medizinische Informatik" absolvierte sie in Lübeck. Bereits mit dem Bachelor spezialisierte sie sich auf die Bildverarbeitung medizinischer Daten. Als wissenschaftliche Mitarbeiterin an der Universität zu Lübeck arbeitet Julia Krüger sieben Jahre (2011-18) in der Forschung und Entwicklung von Methoden und Algorithmen für die medizinische Bildverarbeitung. Die Ergebnisse ihrer Forschungen, die sie auf zahlreichen internationalen Konferenzen sowie in Vorträgen und mehreren wissenschaftlichen Publikationen in renommierten internationalen Journalen präsentieren konnte, führten zu einer Dissertationsschrift, die mit „summa cum laude" bewertet wurde. Seit 2018 arbeitet sie bei einer Hamburger Firma, die durch computergestützte Analysen (mittels künstlicher Intelligenz und Deep Learning Verfahren) von neuro-radiologischen Daten die Versorgung von Patiente unterstützt.

Steffen Hölldobler et. al. (Hrsg.): Ausgezeichnete Informatikdissertationen 2018,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2019 141

# Die Stärke von Lokalität: Welche Rolle spielt Randomisierung in verteilten Graphalgorithmen?

Yannic Maus [1]

**Abstract:** Wir untersuchen verteilte Graphalgorithmen für klassische Probleme, wie zum Beispiel das Kanten- und das Knotenfärbungsproblem. Viele dieser Probleme sind randomisiert effizient lösbar, wohingegen die besten bekannten deterministischen Algorithmen exponentiell langsamer sind. Im ersten Teil der Dissertation nutzen wir einen komplexitätstheoretischen Ansatz und zeigen, dass einige Probleme im folgenden Sinn vollständig sind: Ein effizienter deterministischer Algorithmus für ein vollständiges Problem würde einen effizienten deterministischen Algorithmus für alle randomisiert effizient lösbaren Probleme implizieren. Unter den vollständigen Problemen ist ein scheinbar einfaches natürliches Graphenfärbungsproblem, welches randomisiert trivial und ohne Kommunikation lösbar ist. In den weiteren Teilen der Dissertation entwickeln wir effiziente Kanten- und Knotenfärbungsalgorithmen, und wir beweisen eine untere Schranke für die Laufzeit von Knotenfärbungsalgorithmen in einer abgeschwächten Version des Standardmodells für verteilte Graphalgorithmen.

## 1 Einführung

Netzwerke spielen eine immer größere Rolle in unserer Welt. Viele moderne Systeme, wie das Internet, bauen auf riesigen Netzwerken auf. Auch in der Natur findet man zahlreiche Netzwerke wie das menschliche Gehirn oder unsere Gesellschaft mit ihren sozialen Verknüpfungen. In all diesen Netzwerken gibt es viele Knoten, wie zum Beispiel die Neuronen im Gehirn, die miteinander kommunizieren und obwohl einzelne Knoten nur mit ihren direkten Nachbarn im Netzwerk sprechen können, soll das System eine Art *globale Lösung* berechnen. Eine der Kernfragen in dieser Dissertation lautet:

*Welche globalen Ziele können auf lokalen Informationen basierend erreicht werden?*

Das Gebiet, in welchem Knoten mit ihren Nachbarn Nachrichten austauschen, um ein (globales) Problem zu lösen, heißt *verteilte Graphalgorithmen*. Ursprünglich wurden solche Algorithmen mit der Absicht untersucht, um die Probleme, die beim Routing in Netzwerken auftreten, zu verstehen und zu beheben. Heutzutage spielen Netzwerke in fast jedem Bereich des Lebens eine immense Bedeutung und es einen klaren Trend, der zeigt, dass immer mehr Algorithmen dezentralisiert statt klassisch zentralisiert sind. Daher besteht die Notwendigkeit, dass wir denztralisierte und verteilte Algorithmen besser verstehen.

In all den oben genannten Netzwerken kann jedes Individuum nur mit wenigen anderen Teilnehmern des Netzwerkes kommunizieren, zum Beispiel kann man in einem sozialen

[1] Department of Computer Science, Technion, Haifa, Israel, yannic.maus@cs.uni-freiburg.de

Netzwerk nur mit seinen Freunden und Bekannten kommunizieren. Wegen dieser lokalen Kommunikation und der schieren Größe der Netzwerke[2] beruht die Rolle eines einzelnen Knotens in einer (globalen) Lösung nur auf lokalen Informationen. Dabei nennt man die Entfernung aus welchem ein Knoten auf Informationen im Netzwerk zugreifen kann die *Lokalität* eines Algorithmus.

### 1.1 Das LOCAL-Modell für verteilte Algorithmen

Um das Konzept der Lokalität formal zu untersuchen benutzen wir das gängige *message-passing* Modell für verteilte Graphalgorithmen, das LOCAL-Modell [Li92]: Das Netzwerk wird als ein Graph $G = (V,E)$ mit $n$ Knoten abstrahiert und jeder Knoten hat eine eindeutige ID (wie zum Beispiel seine IP-Adresse). In einem Algorithmus können die Knoten in **synchronen Runden unbeschränkte lokale Berechnungen** durchführen und Nachrichten von **unbeschränkter Größe** zu ihren Nachbarn schicken. Es gibt keine Fehler in der Kommunikation oder in den Berechnungen. Die *Komplexität*, die *Laufzeit* oder auch die *Lokalität* eines Algorithmus ist die Anzahl an synchronen Runden bis jeder Knoten seine Ausgabe, wie beispielsweise seine eigene Farbe in einer Graphenfärbung, berechnet hat. Dieses Modell ermöglicht es die Lokalität von verteilten Algorithmen auf eine präzise und mathetmatische Art und Weise zu untersuchen, da Informationen von einem Knoten $v$ einen anderen Knoten $u$ in Abstand $r$ nur erreichen können, wenn der Algorithmus mindestens $r$ Runden läuft. Umgekehrt kann ein Knoten $v$ wegen der unbegrenzten Nachrichtengröße in $r$ Runden auch alle Informationen in seiner $r$-Nachbarschaft lernen. Ein $r$-Runden Algorithmus ist also nichts Anderes als eine Funktion von der Menge der möglichen $r$-Nachbarschaften. Abbildung 1 zeigt dies an einem Beispiel.

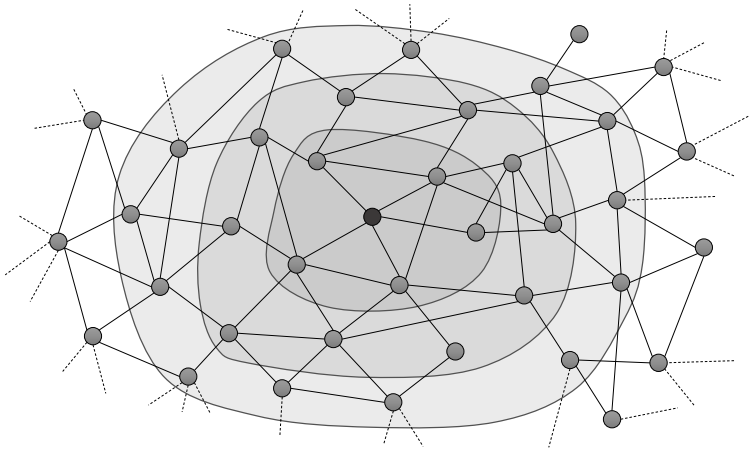


Abb. 1: Lokale Informationen: In einem LOCAL-Algorithmus mit Laufzeit $r$ kann ein Knoten nur Informationen aus seiner $r$-Nachbarschaft erhalten. Die schattierten Bereiche zeigen die Informationen, die der schwarze Knoten in einer, zwei oder drei Runden erhalten kann.

[2] Oft sind Netzwerke so groß, dass sie zusammen mit all ihren Eingabedaten nicht in einem einzelnen Computer gespeichert werden können.

Unbeschränkte lokale Berechnungen werden oft dadurch motiviert, dass die Laufzeit von verteilten Algorithmen in der Realität meist durch die Kommunikationszeit dominiert ist. Obwohl die meisten veröffentlichten Algorithmen keine aufwändigen lokalen Berechnungen benutzen, wird das Modell oft für diese *kostenlosen* lokalen Berechnungen kritisiert. Es ist klar, dass sie jenseits jeglicher Realität liegen — so können Knoten zum Beispiel intern NP-vollständige Probleme lösen, jedoch sind sie unerlässlich für einige Teile dieser Dissertation um das Konzept der Lokalität getrennt von der Berechnungskomplexität eines Problems zu untersuchen. Klassischerweise nennt man einen verteilten Algorithmus im LOCAL-Modell *effizient*, falls seine Rundenkomplexität polylogarithmisch in der Anzahl der Knoten des Netzwerkes ist.

## 1.2 Verteilte Graphenprobleme

Wir untersuchen primär *klassische* Probleme, wie zum Beispiel das Knotenfärbungsproblem. Oft kommen diese Probleme in der Realität zur Anwendung: so lässt sich zum Beispiel das Berechnen eines Ablaufplanes von in Konflikt stehenden Prozessen als Graphenfärbungsproblem modellieren; häufig sind die resultierenden Graphenfärbungsprobleme verteilt gegeben, etwa beim Verteilten von Frequenzen an Mobiltelefone [Sm13]. Formal besteht ein *verteiltes Graphenproblem* $\mathscr{P}$ aus einer Menge von Tupeln $(G,\vec{x},\vec{y})$, wobei $G$ ein einfacher ungerichteter Graph ist; der *Eingabevektor* $\vec{x}$ und der *Ausgabevektor* $\vec{y}$ sind jeweils Vektoren mit Einträgen $x_v$ und $y_v$ für jeden Knoten $v \in V$. Ein Tupel $(G,\vec{x})$ ist eine *Instanz* eines Problems $\mathscr{P}$, falls ein *Ausgabevektor* $\vec{y}$ mit $(G,\vec{x},\vec{y}) \in \mathscr{P}$ existiert. Ein verteilter Algorithmus *löst* ein Problem $\mathscr{P}$, falls er für alle Instanzen $(G,\vec{x})$ einen Ausgabevektor $\vec{y}$ berechnet, so dass $(G,\vec{x},\vec{y}) \in \mathscr{P}$ gilt; hierbei muss jeder Knoten *'seinen Teil der Ausgabe'* berechnen, d.h., Knoten $v$ berechnet nur $y_v$. Meist ist die formelle Definition eines Graphenproblems implizit gegeben, wie zum Beispiel bei den für diese Dissertation zentralen Knoten- und Kantenfärbungsproblemen. Beide sogenannten *symmetry breaking Probleme* werden von Panconesi und Rizzi explizit als zwei der vier wichtigsten Probleme des Gebiets bezeichnet [PR01]. Für einen Graphen $G = (V,E)$ sind sie wie folgt definiert:

- **$C$-Knotenfärbung:** Eine *(gültige) $C$-Knotenfärbung* ist eine Funktion $\phi : V \to \{1,\dots,C\}$, so dass $\phi(v) \neq \phi(w)$ für alle $\{v,w\} \in E$ gilt.
- **$C$-Kantenfärbung:** Eine *(gültige) $C$-Kantenfärbung* ist eine Funktion $\psi : E \to \{1,\dots,C\}$, so dass $\psi(e) \neq \psi(f)$ für alle adjazenten Kanten $e$ und $f$ gilt.

Eine optimale Kantenfärbung, das heißt mit der minimal möglichen Anzahl an Farben, zu berechnen ist eines von Karps 21 NP-vollständigen Problemen [Ka72]. Daher haben verteilte Algorithmen meist das Ziel lediglich eine $(\Delta+1)$-Knotenfärbung oder eine $(2\Delta-1)$-Kantenfärbung zu berechnen, wobei $\Delta$ der Maximalgrad des Netzwerkes ist [BE13]. Wenn man sich auf soviele oder mehr Farben beschränkt, dann können beide Färbungen sehr leicht von sequentiellen *Greedyalgorithmen* gelöst werden: Um zum Beispiel einen Graphen mit $(\Delta+1)$ Farben zu färben, ist es ausreichend in einer beliebigen Reihenfolge durch die Knoten zu iterieren und einen Knoten mit einer beliebigen Farbe

in der Menge $\{1,\dots,\Delta+1\}$ zu färben, die noch keiner seiner bereits gefärbten Nachbarn hat. Weiter kann man bei beiden Problemen eine korrekte Teillösung der Probleme, ergo einen gültigen gefärbten Teilgraphen, immer zu einer Lösung auf dem gesamten Graphen vervollständigen und die Gültigkeit einer Ausgabe im LOCAL-Modell kann in nur einer Runde verifiziert werden.[3]

### 1.3 Inhalte der Dissertation

Viele verteilte Graphenprobleme, wie auch das Knotenfärbungsproblem, lassen sich randomisiert effizient und einfach lösen, wohingegen deterministische Algorithmen exponentiell langsamer sind. Im ersten Teil der vorliegenden Dissertation [Ma18] nutzen wir einen komplexitätstheoretischen Ansatz und zeigen, dass einige Probleme im folgenden Sinn vollständig sind: Ein effizienter deterministischer Algorithmus für ein vollständiges Problem würde einen effizienten deterministischen Algorithmus für alle randomisiert effizient lösbaren Probleme implizieren. Unter den vollständigen Problemen ist ein scheinbar einfaches natürliches Graphenfärbungsproblem, welches randomisiert trivial ohne Kommunikation lösbar ist; deterministisch können wir es (bisher) nicht in polylogarithmischer Zeit lösen. Daher zeigt dieses Problem wie Randomisierung hilft und zeigt auf, was wir deterministisch (bisher) nicht effizient durchführen können.

In den weiteren Teilen der Dissertation verbessern wir die Laufzeit für einige der bekanntesten und wichtigsten verteilten Graphenprobleme. Wir präsentieren den aktuell schnellsten effizienten deterministischen Algorithmus für Kantenfärbungen mit $(2+\varepsilon)\Delta$ Farben, den ersten effizienten deterministischen Kantenfärbungsalgorithmus mit $(1+\varepsilon)\Delta$ Farben und die aktuell schnellsten randomisierten und deterministischen $\Delta$-Knotenfärbungsalgorithmen. Im letzten Kapitel der Dissertation beweisen wir eine untere Schranke für die Lokalität zur Berechnung von Knotenfärbungen in einer abgeschwächten Version des LOCAL-Modells. Die Dissertation beginnt mit einem ausführlichen Überblick über verwandte Forschungsergebnisse und ihrer Einbettung in die aktuelle Forschung; die Dissertation schließt mit einer Vielzahl an konkreten offenen Fragen.

## 2 Welche Rolle spielt Randomisierung in verteilten Graphalgorithmen?

Um die Ergebnisse dieses vor allem konzeptuellen Teiles zu verstehen, schauen wir uns zunächst an, was wir unter einem verteilten randomisierten Algorithmus verstehen. In einem *verteilten randomisierten Algorithmus* stehen jedem Knoten private Zufallsbits zur Verfügung, wobei die Zufallsbits verschiedener Knoten unabhängig sind und es keine globalen von allen Knoten geteilten Zufallsbits gibt. Jedoch können die Zufallsbits an Nach-

[3] Fast alle Probleme, die wir in dieser Dissertation anschauen, haben die Eigenschaft, dass eine Ausgabe im LOCAL-Modell mit polylogarithmischer Lokalität verifiziert werden kann; einige Aussagen in dieser Zusammenfassung gelten nur für Probleme mit dieser Eigenschaft; wir erwähnen diese Voraussetzung im Rahmen dieser Zusammenfassung nicht immer explizit.

barn gesendet werden. Meist wollen wir, dass ein Algorithmus mit hoher Wahrscheinlichkeit funktioniert, d.h., die Wahrscheinlichkeit, dass die Ausgabe das Problem nicht löst ist höchstens $1/n^c$ für ein konstantes $c \geq 2$. Die Welt der Komplexitäten von randomisierten Algorithmen unterscheidet sich drastisch von denen bekannter deterministischer Algorithmen: Einfache randomisierte polylogarithmische Algorithmen für Kanten- sowie Knotenfärbungen und für viele andere eng verwandte Probleme, wie z.B. für das *Maximal Independent Set Problem (MIS)*, kennt man seit den 80er Jahren [Lu86], wohingegen für viele wichtige verteilte Graphenprobleme (unter anderem für das MIS und das Knotenfärbungsproblem) die schnellsten deterministischen Algorithmen bis heute exponentiell langsamer sind.[4] Es gilt als eine der wesentlichsten, bedeutensten und ältesten offenen Fragen des Bereiches, zu Verstehen, ob diese *exponentielle Lücke* inhärent ist [BE13, Li92].[5] Allein die ersten fünf offenen Fragen im viel zitierten und wichtigen Monolog [BE13] über verteiltes Graphenfärben beschäftigen sich alle mit der Frage, ob Randomisierung für effiziente Algorithmen für das Graphenfärbungs- und verwandte Probleme notwendig ist. Selbst als Linial das LOCAL-Modell einführte stellte er ganz explizit die Frage nach einem effizienten deterministischen Algorithmus für das MIS Problem. Intuitiv überrascht es nicht, dass randomisierte Algorithmen für diese Probleme so schnell sind: Im Knotenfärbungsproblem müssen beispielsweise benachbarte Knoten, deren Sicht auf den Graphen symmetrisch ist, unterschiedliche Farben ausgeben. Daher muss es irgendeine *lokale Koordination* bzw. ein sogenanntes *Aufbrechen der Symmetrien/symmetry breaking* zwischen benachbarten Knoten geben. Es ist sehr natürlich, dass Randomisierung hierbei immens helfen kann: Wenn zum Beispiel jeder ungefärbte Knoten einfach zufällig mit gleicher Wahrscheinlichkeit eine Farbe als Kanditat wählt, mit der keiner seiner schon gefärbten Nachbarn gefärbt ist, und diese Farbe behält, falls kein ungefärbter Nachbar die gleiche Kanditatenfarbe gewählt hat, so hat der Knoten eine konstante Wahrscheinlichkeit gefärbt zu werden. Iteriert man diesen sehr einfachen Prozess, so erhält man einen $O(\log n)$-Runden Knotenfärbungsalgorithmus. Es ist nicht klar, wie und ob man diese lokale Koordination überhaupt effizient mit deterministischen Algorithmen erreichen kann.

Der erste Teil dieser Dissertation kann diese (schwierige) Frage zwar nicht beantworten, aber er trägt zu ihrem Verständnis bei, indem er die folgenden informell gestellten Fragen beantwortet: *Welche Rolle spielt Randomisierung beim effizienten Lösen von Problemen, wie Knotenfärbungen oder MIS? Was haben diese Probleme gemeinsam und gibt es noch mehr als die bekannten Probleme mit solch einer exponentiellen Lücke? Was ist die Hauptschwierigkeit für deterministische Algorithmen? Welche Probleme müssen wir uns anschauen, um die Lücke zu schließen oder zumindest um sie zu verkleinern?* Um all diese Fragen zu beantworten, benutzen wir einen komplexitätstheoretischen Ansatz und definieren eine Klasse von Problemen, die wir P-SLOCAL nennen. Die Definition dieser Klasse lässt sich leicht verstehen, wenn wir uns daran erinnern, dass das Knotenfärbungsproblem

[4] Für das $(2\Delta - 1)$-Kantenfärbungsproblem wurde erst im Jahre 2017 von Fischer, Ghaffari und Kuhn der erste effiziente deterministische Algorithmus gefunden [FGK17].

[5] Selbst falls man nur an randomisierten Algorithmen interessiert ist, ist es nötig ihre deterministischen Gegenstücke detailliert zu verstehen, denn Chang, Kopelowitz und Pettie haben gezeigt, dass die Komplexität des besten randomisierten Algorithmus nicht besser sein kann als sein deterministisches Gegenstück auf exponentiell kleineren Instanzen [CKP16]; bisher wurde das Resultat nur für Graphen mit konstantem Grad gezeigt. Dieses Ergebnis zeigt sich in allen aktuell schnellsten randomisierten Algorithmen, die stets einen Term $T(\sqrt{\log n})$ in ihrer Komplexität haben, wobei $T(\cdot)$ die Laufzeit des besten deterministischen Algorithmus ist.

durch einen trivialen sequentiellen Greedyalgorithmus lösbar ist. In diesem iteriert man in beliebiger Reihenfolge durch die Knoten. Dabei hängt die ausgegebene Farbe eines Knoten nur davon ab, wie seine bereits durchlaufenen Nachbarn gefärbt sind. Das SLOCAL-Modell, welches in dieser Dissertation definiert wird, erweitert dieses Konzept. In einem SLOCAL-Algorithmus oder einem *generalisierten sequentiellen Greedyalgorithmus* mit Lokalität $r$ iteriert man in beliebiger Reihenfolge durch die Knoten und die Ausgabe eines Knotens hängt nur vom aktuellen Zustand seiner Nachbarn in seiner $r$-Nachbarschaft ab. Der vorgestellte Greedyalgorithmus für das Knotenfärbungsproblem ist ein SLOCAL-Algorithmus mit Lokalität 1. Die Klasse P-SLOCAL besteht aus den Problemen, die *effizient* durch einen generalisierten sequentiellen Greedyalgorithmus gelöst werden können, d.h., mit polylogarithmischer Lokalität. Zunächst zeigen wir, dass alle Probleme in der Klasse P-SLOCAL effizient im LOCAL-Modell lösbar sind, wenn man Randomisierung erlaubt. Weiter enthält Klasse alle klassischen in der Literatur studierten Probleme und sie enthält insbesondere die genannten Probleme mit der exponentiellen Laufzeitlücke.[6] Daraus ergibt sich eine formale Version der Frage, ob die exponentielle Lücke tatsächlich inhärent ist: *Ist die Klasse der im* SLOCAL*-Modell effizient lösbaren Probleme identisch mit der Klasse der im* LOCAL*-Modell deterministisch effizient lösbaren Probleme?* Auch diese Frage können wir nicht beantworten, jedoch finden wir eine Hand voll im folgenden Sinne P-SLOCAL-*vollständiger* Probleme: Kann man auch nur ein einziges dieser Probleme effizient mit einem deterministischen Algorithmus im LOCAL-Modell lösen, so lassen sich alle Probleme in der Klasse P-SLOCAL deterministisch effizient lösen. Randomisierung wird also nur benötigt, um ein vollständiges Problem zu lösen. Unter den vollständigen Problemen sind zwei eng verwandte Probleme besonders hervorzuheben, das Local Splitting Problem und das Weak Local Splitting Problem: Sei $B = (U \dot\cup V, E)$ ein bipartiter Graph, wobei der Grad von jedem Knoten in $U$ mindestens $\log^c n$ für eine ausreichend große Konstante $c \geq 2$ ist. Ziel ist es, jeden Knoten in $V$ rot oder blau zu färben, so dass ungefähr die Hälfte der Nachbarn von jedem Knoten in $U$ rot bzw. blau gefärbt ist. Genauer, für $\lambda \in (0, 1/2]$, ist eine 2-Färbung der Knoten in $V$ ein $\lambda$*-Local Splitting*, falls jeder Knoten $u \in U$ mindestens $\lfloor \lambda \cdot d(u) \rfloor$ Nachbarn von jeder Farbe hat, wobei $d(u)$ der Grad von $u$ ist. Eine 2-Färbung der Knoten in $V$ ist ein *Weak Local Splitting*, falls jeder Knoten $u \in U$ mindestens 1 Nachbarn von jeder Farbe hat. Wir zeigen:

**Theorem 1** ([GKM17])**.** *Sei $\lambda = \frac{1}{\operatorname{poly}\log n}$. In bipartiten Graphen $H = (U \dot\cup V, E)$, in denen alle Knoten in $U$ mindestens Grad $c \ln^2 n$ für eine ausreichend große Konstante $c$ haben, ist das $\lambda$-Local Splitting Problem* P-SLOCAL*-vollständig. Falls sich zusätzlich der Grad von allen Knoten in $U$ maximal um einen konstanten Faktor unterscheidet, so ist auch das Weak Local Splitting Problem* P-SLOCAL*-vollständig.*

Mit Randomisierung kann man beide Probleme durch einen trivialen 0-Runden Algorithmus lösen, indem man jeden Knoten in $V$ mit Wahrscheinlichkeit $1/2$ rot oder blau färbt. Falls wir jeden Knoten auch mit mehreren Farben jeweils anteilig färben dürften, so wären beide Probleme gelöst, wenn jeder Knoten in $V$ zur Hälfte rot und zur Hälfte

[6] Die Klasse P-SLOCAL wurde weiter von Ghaffari, Harris und Kuhn untersucht und sie zeigten, dass die Klasse P-SLOCAL genau mit der Klasse der randomisiert effizient im LOCAL-Modell lösbaren Problemen übereinstimmt, sofern man nur *locally-checkable* Probleme betrachtet [GHK18].

blau gefärbt wäre. Von dieser Färbung mit anteiligen Farben zu einer richtigen Färbung zu kommen, kann als ein einfaches Rundungsproblem von rationalen Zahlen zu ganzen Zahlen unter der Beachtung von schwachen Randbedingungen interpretiert werden. Da dieses *Rundungsproblem* randomisiert ohne Kommunikation gelöst werden kann, aber deterministisch bisher nicht effizient gelöst werden kann, zeigt es genau auf, welche Schritte wir ohne Randomisierung nicht effizient lösen können. Wir fassen zusammen:

> *Das Einzige, was wir deterministisch im* LOCAL*-Modell nicht mit* $\operatorname{poly}\log n$ *Lokalität können, ist das grobe Runden (unter der Beachtung von Randbedingungen) von rationalen Zahlen zu ganzen Zahlen. Falls* $\operatorname{poly}\log n$ *Lokalität ausreicht, um deterministisch runden zu können, so können wir alle klassischen Probleme im* LOCAL*-Modell mit* $\operatorname{poly}\log n$ *Lokalität lösen.*

Da wir in diesem Teil auch zeigen, dass man SLOCAL-Algorithmen in LOCAL Algorithem übersetzen kann — mit geringer Vergrößerung der Lokalität in randomisierte LOCAL-Algorithmen und unter erheblicher Vergrößerung der Lokalität auch in deterministische LOCAL-Algorithmen —, dient das Modell auch als Werkzeug, um neue Algorithmen im LOCAL-Modell zu entwickeln. In der Dissertation wird dies genutzt, um Approximationsalgorithmen für gewisse verteilt gegebene lineare Programme zu entwickeln. Mittlerweile haben auch andere Autoren das Modell aufgegriffen, was zu zahlreichen verbesserten Algorithmen in verschiedenen Bereichen geführt hat, unter anderem die Kantenfärbungsalgorithmen und die Algorithmen für das Lovász Local Lemma in [GHK18].

## 3 Knoten- und Kantenfärbungen

In den weiteren Teilen der Dissertation untersuchen wir die Komplexität verschiedener Varianten des verteilten Graphenfärbungsproblems.

### 3.1 Kantenfärbungsalgorithmen

Alle unsere Kantenfärbungsalgorithmen nutzen als Subroutine schnelle Methoden zum Degree Splitting. Das Ziel im *Degree Splitting Problem* ist es, die Kanten eines Graphen in zwei Mengen zu unterteilen, so dass der Grad jedes Knotens in beiden Mengen ungefähr die Hälfte seines Originalgrades ist. Dabei nennt man die Differenz zwischen der Anzahl inzidenter Kanten eines Knotens in den beiden Mengen die *Diskrepanz* des Knotens im Degree Splitting; Ziel ist es also für jeden Knoten eine möglichst kleine Diskrepanz zu erhalten. Um die Kanten eines Graphen zu färben, benutzen wir diese Degree Splittings für eine Divide-and-Conquer Lösung. Wir teilen den Graphen rekursiv in $\Delta/d$ Graphen, wobei jeder Graph Maximalgrad $\approx d$ für ein $d$ in polylogarithmischer Größenordnung hat. Danach färben wir jeden dieser Teilgraphen mit einer unterschiedlichen Farbmenge, der Größe $(1+o(1))d$ falls unser Ziel $(1+o(1))\Delta$ Farben sind, und mit $(2d-1)$ Farben falls wir mit $(2+o(1))\Delta$ Farben färben wollen. So erhalten wir effiziente Laufzeiten für allgemeine Graphen, obwohl die Laufzeit der Algorithmen zum Färben mit $1+(o(1))d$ oder

$(2d-1)$ Farben polynomiell im Maximalgrad des jeweiligen Graphen sind. Im folgenden Theorem zeigen wir, dass wir Degree Splittings effizient berechnen können.

**Theorem 2** ([Gh17]). *Für jedes $\varepsilon > 0$ existiert ein deterministischer Algorithmus mit Laufzeit $O\big(\varepsilon^{-1} \cdot \log \varepsilon^{-1} \cdot \big(\log\log \varepsilon^{-1}\big)^{1.71} \cdot \log n\big)$, um ein Degree Splitting zu berechnen, in dem die Diskrepanz von Knoten v höchstens $\varepsilon \cdot d(v) + 4$ ist.*

Durch rekursives Anwenden von Theorem 2 und Färben der Kanten der resultierenden Teilgraphen erhalten wir den folgenden effizienten Färbungsalgorithmus.

**Corollary 3** ([Gh17]). *Für jedes $\varepsilon > 1/\log\Delta$, existiert ein deterministischer Algorithmus, der eine $(2+\varepsilon)\Delta$-Kantenfärbung in $O\big(\log^2\Delta \cdot \varepsilon^{-1} \cdot \log\log\Delta \cdot (\log\log\log\Delta)^{1.71} \cdot \log n\big)$ Runden berechnet.*

Klassisch färben verteilte Algorithmen mit $2\Delta-1$ oder mehr Farben. Corollary 3 ist der schnellste Algorithmus, wenn man leicht über dieser Schranke bleibt. Jedoch weiß man schon seit 1964 durch Vizings berühmtes Theorem, dass Kantenfärbungen mit $\Delta+1$ Farben immer existieren [Vi64]. In [FGK17] wurde das $(2\Delta-1)$-Kantenfärbungsproblem erstmals effizient deterministisch gelöst und die Frage aufgeworfen, *wie nah man mit einem deterministischen und effizienten Algorithmus an Vizings Schranke kommen (könne)?* Als Antwort geben wir den ersten effizienten deterministischen verteilten Algorithmus überhaupt, der deutlich weniger als $(2\Delta-1)$ Farben benutzt.

**Theorem 4** ([Gh18b]). *Es existiert eine Konstante $c > 0$, so dass für jedes $\varepsilon > 0$ deterministische Algorithmen existieren, die die Kanten jedes Graphen mit n Knoten und Maximalgrad $\Delta$ mit*

*a)* $(1+\varepsilon)\Delta$ *Farben in $O\big(\varepsilon^{-9}\log^3 n \log^4\Delta\log\varepsilon^{-1}\big)$ Runden färbt, falls $\Delta \geq c\cdot\varepsilon^{-1}\cdot\log\varepsilon^{-1}\cdot\log n$ gilt, oder mit*

*b)* $3\Delta/2$ *Farben in $O\big(\Delta^9\log^8 n\log^5\Delta\big)$ Runden färbt (für alle $\Delta$).*

Der Beweis von Vizings Theorem benutzt globale Argumente und es existieren keine trivialen Greedyalgorithmen, um die Kanten eines Graphen mit weniger als $2\Delta-1$ Farben zu färben. Sprich, das $C$-Kantenfärbungsproblem mit $C \ll 2\Delta-1$ hat eine völlig andere Natur als das Färbungsproblem mit $2\Delta-1$ oder mehr Farben. Daher nutzt Theorem 4 andere Techniken als bisherige Kantenfärbungsalgorithmen.

### 3.2 Knotenfärbungen

Im vorletzten Kapitel der Dissertation präsentieren wir $\Delta$-Knotenfärbungsalgorithmen. Die Existenz solcher Färbungen für zusammenhängende Graphen (solange der Graph nicht vollständig und kein Ring mit ungerader Knotenanzahl ist) geht auf ein sehr bekanntes Resultat von Brooks [Br41] zurück. Wir zeigen das folgende Theorem.

**Theorem 5** ([Gh18a]). *Es existiert ein randomisierter Algorithmus, der jeden nicht vollständigen Graphen mit Maximalgrad $\Delta \geq 3$, mit hoher Wahrscheinlichkeit mit $\Delta$ Farben färbt und $O\left(\sqrt{\Delta \log \Delta} \cdot \log^* \Delta \cdot \log^2 \log n\right)$ Runden benötigt. Gilt $\Delta \geq 4$ ist die Laufzeit $O(\log \Delta) + 2^{O(\sqrt{\log \log n})}$.*

Man beachte, dass die Bedingung $\Delta \geq 3$ notwendig ist, da eine 2-Färbung eines Graphen mit $\Delta = 2$ ein globales Problem ist und $\Omega(n)$ Runden benötigt. Neben den randomisierten Algorithmen aus Theorem 5 enthält dieser Teil auch neue deterministische $\Delta$-Knotenfärbungsalgorithmen. Die in dieser Dissertation enthaltenen $\Delta$-Knotenfärbungsalgorithmen verbessern 25 Jahre alte Resulate von Panconesi und Srinivasan [PS95].

Im letzten Kapitel, das auf der Publikation [He16] basiert, präsentieren wir für das verteilte $(\Delta+1)$-Knotenfärbungsproblem eine untere Schranke von $\Omega(\Delta^{\frac{1}{3}})$ Runden in einer abgeschwächten Version des LOCAL-Modells. Dieses Resultat ist, trotz intensiver Forschung, die einzige untere Schranke für die Laufzeit des verteilten Graphenfärbungsproblems seit über 30 Jahren.

### 3.3 Die Zentralität von Splittingproblemen

Wie in Abschnitt 3.1 beschrieben, nutzen unsere schnellen Kantenfärbungsalgorithmen Degree Splitting Methoden, die deterministisch effizient lösbar sind. Diese Degree Splitting Methoden sind auf dem Kantengraphen definiert. Dahingegen zeigt der erste Teil der Arbeit, dass ähnliche Splitting Probleme auf allgemeinen Graphen P-SLOCAL-vollständig sind und wir nicht wissen, wie wir diese effizient lösen können. Daher entält die Dissertation auch einen kurzen Teil, der die zentrale Bedeutung von Splittingalgorithmen für verteilte Graphalgorithmen herausstellt und kurz zusammenfasst, in welchen anderen wichtigen Resultaten Splitting Probleme eine zentrale Rolle spielen. Die Dissertation schließt mit konkreten offenen Fragen.

## Literaturverzeichnis

[BE13] Barenboim, L.; Elkin, M.: Distributed Graph Coloring: Fundamentals and Recent Developments. Morgan & Claypool Publishers, 2013.

[Br41] Brooks, R. L.: On Colouring the Nodes of a Network. Mathematical Proceedings of the Cambridge Philosophical Society, 37(2):194–197, 1941.

[CKP16] Chang, Y. J.; Kopelowitz, T.; Pettie, S.: An Exponential Separation between Randomized and Deterministic Complexity in the LOCAL Model. In: Proceedings of the Symposium on Foundations of Computer Science (FOCS). S. 615–624, 2016.

[FGK17] Fischer, M.; Ghaffari, M.; Kuhn, F.: Deterministic Distributed Edge-Coloring via Hypergraph Maximal Matching. In: Proceedings of the Symposium on Foundations of Computer Science (FOCS). IEEE Computer Society, S. 180–191, 2017.

[Gh17] Ghaffari, M.; Hirvonen, J.; Kuhn, F.; Maus, Y.; Suomela, J.; Uitto, J.: Improved Distributed Degree Splitting and Edge Coloring. In: Proc. Int. Symp. on Distributed Computing (DISC). S. 19:1–19:15, 2017.

[Gh18a] Ghaffari, M.; Hirvonen, J.; Kuhn, F.; Maus, Y.: Improved Distributed Δ-Coloring. In: Proc. ACM Symp. on Principles of Distributed Computing (PODC). 2018.

[Gh18b] Ghaffari, M.; Kuhn, F.; Maus, Y.; Uitto, J.: Deterministic Distributed Edge-Coloring with Fewer Colors. In: Proc. ACM Symp. on Theory of Computing (STOC). ACM, 2018.

[GHK18] Ghaffari, M.; Harris, D. G.; Kuhn, F.: On Derandomizing Local Distributed Algorithms. In: Proc. Symp. on Foundations of Computer Science (FOCS). S. 662–673, 2018.

[GKM17] Ghaffari, M.; Kuhn, F.; Maus, Y.: On the Complexity of Local Distributed Graph Problems. In: Proc. ACM Symp. on Theory of Computing (STOC). ACM, S. 784–797, 2017.

[He16] Hefetz, D.; Maus, Y.; Kuhn, F.; Steger, A.: A Polynomial Lower Bound for Distributed Graph Coloring in a Weak LOCAL Model. In: Proc. Int. Symp. on Distributed Computing (DISC). S. 99–113, 2016.

[Ka72] Karp, R. M.: Reducibility among Combinatorial Problems. In: Symposium on Complexity of Computer Computations. S. 85–103, 1972.

[Li92] Linial, N.: Locality in Distributed Graph Algorithms. SIAM Journal on Computing, 21(1):193–201, 1992.

[Lu86] Luby, M.: A Simple Parallel Algorithm for the Maximal Independent Set Problem. SIAM Journal on Computing, 15(4):1036–1053, 1986.

[Ma18] Maus, Y.: The Power of Locality: Exploring the Limits of Randomness in Distributed Computing. Dissertation, University of Freiburg, Freiburg im Breisgau, Germany, 2018.

[PR01] Panconesi, A.; Rizzi, R.: Some Simple Distributed Algorithms for Sparse Networks. Distributed Computing, 14(2):97–100, 2001.

[PS95] Panconesi, A.; Srinivasan, A.: The Local Nature of Δ-Coloring and its Algorithmic Applications. Combinatorica, 15(2):255–280, Jun 1995.

[Sm13] Smorodinsky, S.: Conflict-Free Coloring and its Applications. In (Bárány, Imre; Böröczky, Károly J.; Tóth, Gábor Fejes; Pach, János, Hrsg.): Geometry — Intuitive, Discrete, and Convex: A Tribute to László Fejes Tóth. Springer Berlin Heidelberg, Berlin, Heidelberg, S. 331–389, 2013.

[Vi64] Vizing, V.: On an Estimate of the Chromatic Class of a p-Graph. Diskret analiz, 3:25–30, 1964.

**Yannic Maus** studierte an der *RWTH Aachen* und der *National University of Singapore* sowohl Mathematik (Bachelor und Master) als auch Informatik (Bachelor). Für seine mit Auszeichnung abgeschlossenen Studiengänge wurde er unter anderem mit der Springorum-Denkmünze der RWTH Aachen geehrt. Im Anschluss promovierte er am Lehrstuhl für Algorithmen & Komplexität der *Albert-Ludwigs-Universität Freiburg* unter der Betreuung von Professor Dr. Fabian Kuhn. Seine Promotion über verteilte Algorithmen schloss er im Oktober 2018 mit dem Gesamtprädikat *summa cum laude* ab. Danach zog es ihn als Postdoktorand in das Land mit der höchsten Dichte an Forschern im Bereich der verteilten Algorithmen, genauer gesagt ans Technion (הטכניון) in Haifa, Israel. In seiner Freizeit fährt er leidenschaftlich (Renn-)rad oder ist im Winter auf der Loipe beim Langlaufen anzutreffen.

# Effizientes Lernen aus Vergleichen

Lucas Maystre[1]

## 1 Motivation

In Auswahlentscheidungen manifestieren sich unsere Meinungen und Präferenzen. Wir treffen eine Auswahl der Musik, die wir hören, und der Filme, die wir sehen. Wir wählen den Ort aus, an dem wir leben, und den politischen Kandidaten, dem wir unsere Stimme geben. Laufend vergleichen wir Alternativen miteinander, um diejenige zu erkennen, die für uns die richtige ist. So überrascht es nicht, dass sich durch Observieren der Ergebnisse solcher Vergleiche ein umfassendes Verständnis für kollektive und persönliche Meinungen erlangen lässt.

Die Idee, menschliche Auswahlentscheidungen zu analysieren, zieht schon seit geraumer Zeit Forscher und Praktiker aus zahlreichen Disziplinen, wie der Psychologie, der Soziologie und den Wirtschaftswissenschaften, in ihren Bann. Um nur ein Beispiel von vielen zu nennen: In der Ökonometrie zählt die Discrete-Choice-Analyse (DCA) inzwischen zum Standardrepertoire. Die DCA hat wichtige Anwendungen. So konnten mit ihrer Hilfe beispielsweise die Auswirkungen einer neuen Stadtbahnlinie in der San Franscisco Bay Area auf die Nutzung unterschiedlicher Verkehrsmittel genau vorhergesagt werden [Mc77]. Die in diesem Zusammenhang entwickelten Theorien und Methoden brachten ihrem hauptsächlichen Erfinder einen Nobelpreis ein [Mc01].

Die vorliegende Arbeit reiht sich ein in die Bemühungen um bessere Methoden zur Analyse von menschlichen Auswahlentscheidungen. Konkret interessieren wir uns für das Problem, wie aus rohen *Auswahldaten* knappe und aussagekräftige Informationen (etwa über unsere Präferenzen) gewonnen werden können. Unter Auswahldaten verstehen wir dabei empirische Daten, die eine von mehreren Alternativen auszeichnen. Konkret könnte eine typische Aufgabe lauten, alle Alternativen in eine Rangfolge von der am meisten zu der am wenigsten bevorzugten Alternative zu gliedern. Oft erfolgt dies anhand von numerischen Wertungen, die den Nutzen der einzelnen Alternativen beschreiben und denen Vorhersagekraft für künftige Entscheidungen zukommt. Obwohl die Forschung zu Auswahlmodellen bereits eine Reihe bewährter Methoden hervorgebracht hat, machen moderne Online-Anwendungen (für die im Weiteren Beispiele angeführt werden) neue Ansätze zum Handhaben großer Datenmengen erforderlich. Tatsächlich werden neue Herausforderungen sowohl durch die

[1] École Polytechnique Fédérale de Lausanne, Schweiz. Korrespondenz: lucas.maystre@epfl.ch. Englischer Originaltitel: *Efficient Learning from Comparisons*.

große Anzahl an *Observationen* als auch durch die große Anzahl an *Alternativen* geschaffen, die für moderne Anwendungen typisch sind. Es wird wichtig, Methoden zu entwickeln, die effizient sind – nicht nur, um schnell alle Observationen zu verarbeiten, sondern auch, um ausreichende Informationen über jede einzelne Alternative zu erhalten. Der *Effizienzgedanke* zieht sich als roter Faden durch die vorliegende Arbeit und wird im Abschnitt 3 weiter ausgeführt.

**Wozu Auswahldaten studieren?** Wenn es also darum geht, eine numerische Wertung des Nutzens einer jeweiligen Alternative zu erhalten, so stellt sich die berechtigte Frage: Warum nicht einfach *direkt* nach einer solchen Wertung fragen? Dafür gibt es zwei wichtige Gründe:

1. Für Menschen ist es besonders einfach und natürlich, Vergleiche anzustellen. Eine weit verbreitete Theorie in der sozialen Psychologie sagt sogar aus, dass wir uns selbst, unsere Überzeugungen und unsere Meinungen dadurch erkennen und definieren, dass wir uns mit anderen vergleichen [Fe54]. Demgegenüber fällt uns das Abgeben angemessener und konsistenter numerischer Wertungen eher schwerer. Was bedeutet eine Wertung von „3,5 Sternen“ bei einem Restaurant wirklich? In einer Welt, in der alles relativ ist, ist eine absolute Wertung vielleicht einfach die falsche Abstraktion.
2. Manchmal ist es möglich, Auswahlentscheidungen *implizit* zu beobachten, indem wir einfach Handlungen protokollieren sowie den Kontext, in dem diese stattfinden. So lassen sich Auswahldaten auf eine viel unaufdringlichere Art und Weise erlangen als durch *explizites* Fragen nach Feedback. In der Praxis kann so oft Zugriff auf viel größere Datensätze erlangt werden, was potentiell zu genaueren Modellen führt.

**Umgang mit inkonsistenten Daten** Auf den ersten Blick mag es einfach erscheinen, aus Vergleichsdaten ein Verständnis für die zugrunde liegenden Meinungen zu entwickeln. Tatsächlich wäre das auch so, wenn die beobachteten Auswahlentscheidungen auf perfekte Weise einen einzige Menge von Meinungen widerspiegeln würden. Betrachten wir jedoch „in freier Wildbahn“ gesammelte Daten, so erkennen wir schnell, dass die Ergebnisse von Vergleichen nicht immer miteinander konsistent sind: Anscheinend treffen wir auch bei identischen Alternativen manchmal unterschiedliche Auswahlentscheidungen. Hierfür sind zahlreiche Faktoren verantwortlich, beispielsweise: (*a*) Ein Teil des Kontexts, in dem die Auswahl getroffen wird, wurde eventuell nicht beobachtet, hat jedoch möglicherweise einen wesentlichen Einfluss auf das Ergebnis. (*b*) Bei dem Versuch, kollektive Präferenzen auf der Grundlage individueller Auswahlentscheidungen zusammenzufassen, ist offensichtlich ein gewisses Maß an Nichtübereinstimmung zwischen den Individuen zu erwarten, selbst wenn es auch gemeinsame Trends gibt. (*c*) Manchmal schleichen sich auch Fehler in die Daten ein, die auf fehlerhafte Messungen oder unvollkommene Interpretationen zurückzuführen sind. Die vorliegende Arbeit geht davon aus, dass solche Inkonsistenzen unvermeidbar sind,

es jedoch möglich ist, mit Hilfe eines Wahrscheinlichkeitsmodells systematisch mit ihnen umzugehen. Kurz gefasst beruht der Ansatz darauf, dass bei einem gegebenen Satz von Alternativen *jedes* Vergleichsergebnis möglich ist, aber manche Ergebnisse wahrscheinlicher sind als andere – abhängig von den zugrunde liegenden Präferenzen. Die Aufgabe reduziert sich dann darauf, diejenigen Präferenzen zu ermitteln, die die Observationen gut erklären. Dieser Ansatz dominiert in der Fachwelt und wurde auch für die vorliegende Arbeit gewählt. Er wird im Abschnitt 2 näher erläutert.

**Moderne Anwendungen** Auswahlmodelle haben eine lange und reichhaltige Tradition, doch im Zusammenhang mit massenhafter Online-Datenerfassung ist das Interesse an ihnen neu aufgelebt. Das Web macht es für Unternehmen einfach, Kunden auf der ganzen Welt zu erreichen und dabei ihre Interaktionen mit dem Leistungsangebot des Unternehmens zu protokollieren. Betrachten wir hierzu drei Beispiele.

- Anbieter kommerzieller Online-Dienste verlassen sich in zunehmendem Maße auf Empfehlungssysteme (d. h. Systeme, die die Präferenzen von Kunden erlernen), um die Kundenbindung zu erhöhen und den Absatz zu steigern. Spotify und Netflix, zwei beliebte Musik- und Videostreaming-Dienste, erlernen Präferenzen anhand von impliziten Observationen der Auswahlentscheidungen der Benutzer (d. h., welche Titel sie hören bzw. welche Filme sie sehen). Der E-Commerce-Riese Amazon unterbreitet seinen Kunden personalisierte Kaufempfehlungen, die auf früheren Einkäufen basieren.
- Wissenschaftler haben Online-Plattformen erstellt, mit denen sie große Mengen an Vergleichsdaten sammeln können, um schwierige Fragen aus der psychologischen und soziologischen Forschung zu beantworten. So hat sich zum Beispiel das GIFGIF-Projekt[2] zum Ziel gesetzt, den emotionalen Inhalt animierter GIF-Bilder zu verstehen, indem Benutzern Bildpaare gezeigt und die Benutzer gefragt werden, welches Bild eine Emotion wie Freude, Scham usw. besser ausdrückt. Das Place Pulse-Projekt[3] möchte verstehen, wie Stadtviertel wahrgenommen werden, und benutzt dazu ebenfalls paarweise Vergleichsfragen. In beiden Fällen sind Vergleiche ein natürliches Mittel, um Feedback von Benutzern zu erhalten. Beide Projekte haben Millionen von Datenpunkten über Tausende von Objekte gesammelt und faszinierende Erkenntnisse geliefert, die früher mit herkömmlichen Methoden nicht zu erreichen waren.
- Paarweise Vergleiche bilden die Grundlage von *Wiki-Surveys*, einer neuartigen Befragungsmethodik, die von Salganik; Levy [SL15] entwickelt wurde. Wiki-Surveys sind der Versuch, die Lücke zwischen Fragebögen auf der einen und mündlichen Befragungen auf der anderen Seite zu schließen – Fragebögen skalieren gut, bieten aber keinen Raum für das Hervortreten neuer Informationen. Mündliche Befragungen

[2] Siehe: `http://www.gif.gf/`.
[3] Siehe: `http://pulse.media.mit.edu/`.

können unverhoffte neue Erkenntnisse liefern, sind aber kostspielig in der Durchführung. Beispielsweise hat die Stadtverwaltung von New York City mit Wiki-Surveys Feedback zu einem Nachhaltigkeitsplan eingeholt. Die Respondenten konnten entweder neue Ideen vorschlagen oder eine Vergleichsfrage der Art „Welche der folgenden beiden Ideen ist Ihrer Meinung nach besser geeignet, eine grünere und bessere New York City zu schaffen?“ beantworten. Der Wiki-Survey-Dienst macht es möglich, gleichzeitig sowohl neue Vorschläge zu erhalten als auch bekannte Vorschläge nach ihrer Priorität zu ordnen. Zum Zeitpunkt der Entstehung der vorliegenden Arbeit gab es auf `http://www.allourideas.org/` bereits 11 739 Wiki-Surveys mit insgesamt 17.8 Millionen Stimmen zu 631 682 Ideen.

**Mehr als Präferenzen: Anwendungen im Sport** Abschließend sei darauf hingewiesen, dass dieselben Methoden, mit denen menschliche Auswahlentscheidungen modelliert werden, auch für Probleme benutzt werden können, die auf den ersten Blick konzeptuell völlig andersartig zu sein scheinen. So behandelt die vorliegende Arbeit auch das Problem der Vorhersage von Fußballergebnissen auf der Grundlage historischer Daten. Bei einem Fußballspiel werden zwei Mannschaften miteinander verglichen, und am Ende gewinnt eine davon. In unserer Terminologie können wir die Mannschaften als Alternativen betrachten, die miteinander verglichen werden, und den Sieger als das Ergebnis des Vergleichs. Interessanterweise haben sich die wesentlichen Modelle und Ideen, die in der vorliegenden Arbeit verwendet werden, gleichzeitig sowohl im Kontext der Analyse menschlicher Auswahlentscheidungen als auch im Kontext der Vorhersage von Wettkampfergebnissen entwickelt, wie wir im nächsten Abschnitt sehen werden.

## 2 Ausgewählte Wahrscheinlichkeitsmodelle

In diesem Abschnitt werden die statistischen Modelle und zugehörigen Methoden vorgestellt, die im Rahmen der vorliegenden Arbeit benutzt werden oder auf die Bezug genommen wird. Wir nehmen einen historischen Blickwinkel ein: Der Kontext, in dem diese Modelle und Verfahren entstanden sind, ist faszinierend. Dieser Abschnitt liefert nur einen kurzen Überblick über die Entwicklungen, enthält jedoch Verweise auf umfassendere Informationen für den interessierten Leser.

### 2.1 Thurstones Modell

Im Jahre 1927 veröffentlichte Thurstone einen Artikel, der weithin als grundlegend für das Gebiet der Wahrscheinlichkeitsmodelle für Vergleichsergebnissen angesehen wird [Th27]. Thurstone interessierte sich für das Problem der Messung in der Psychologie und entwickelte ein Verfahren, das die Antworten von menschlichen Probanden auf Vergleiche zwischen zwei von $N$ Stimuli erklärt. Um die Tatsache zu berücksichtigen, dass die Reaktion

auf einen Stimulus variieren kann, schlug Thurstone vor, den wahrgenommenen Wert eines Stimulus $i$ während eines Experiments mittels einer *zufälligen* Variablen $x_i \in \mathbf{R}$ zu modellieren. Das Ergebnis des Vergleichs zwischen den Stimuli $i$ und $j$ ist durch eine Realisierung der entsprechenden zwei Zufallsvariablen gegeben, d. h. durch das Ereignis $x_i > x_j$. Thurstone postulierte ferner, dass diese Zufallsvariablen einer multivariaten Gauss-Verteilung $\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ folgen. Hierbei bezeichnet $i \succ j$ das Ereignis „$i$ wird gegenüber über $j$ bevorzugt“.

$$\mathbf{P}\left[i \succ j\right] = \mathbf{P}\left[x_i > x_j\right] = \Phi\left(\frac{\theta_i - \theta_j}{\sqrt{\Sigma_{ii} + \Sigma_{jj} - 2\Sigma_{ij}}}\right).$$

Hierbei ist $\Phi(\cdot)$ die kumulative Dichtefunktion der Standardnormalverteilung. Die letzte Gleichung ergibt sich daraus, dass $x_i - x_j \sim \mathrm{N}(\theta_i - \theta_j, \Sigma_{ii} + \Sigma_{jj} - 2\Sigma ij)$. Thurstone betrachtete mehrere Varianten des Modells, die Schritt für Schritt restriktivere Annahmen über die Kovarianzmatrix $\boldsymbol{\Sigma}$ machen. Die heutzutage am häufigsten benutzte Variante erhält man durch Einsetzen von $\boldsymbol{\Sigma} = \frac{1}{2}\boldsymbol{I}$. In diesem Fall ist

$$\mathbf{P}\left[i \succ j\right] = \Phi(\theta_i - \theta_j). \tag{1}$$

Der Vektor der $N$ Parameter $\boldsymbol{\theta} = [\theta_1 \ \cdots \ \theta_N]^\top \in \mathbf{R}^N$ bestimmt die Wahrscheinlichkeiten aller $\binom{N}{2}$ möglichen paarweisen Vergleiche. Intuitiv lässt sich $\theta_i$ als die *Wertung* des Stimulus $i$ verstehen, und die Wahrscheinlichkeit eines Vergleichsergebnisses für $i$ und $j$, das mit der tatsächlichen Reihenfolge konsistent ist, steigt mit der Distanz $\theta_i - \theta_j$. Man beachte, dass, weil (1) nur paarweise Distanzen enthält, die Parameter $\boldsymbol{\theta}$ nur bis auf eine Konstante bestimmt werden können. Um diese Unbestimmtheit aufzulösen, werden die Parameter oft derart gewählt, dass $\sum_i \theta_i = 0$.

Die vielleicht erste Anwendung, die Thurstone im Sinn hatte, betrifft das Gebiet der Psychophysik. Man stelle sich vor, dass man zwei Bälle erhält und gefragt wird: „Welcher dieser beiden Bälle ist schwerer?“ Hat man eine Sammlung von Observationen dieser Art (von denen wohl einige inkonsistent sein werden), könnte das Modell (1) benutzt werden, um durch Schätzen der Parameter $\boldsymbol{\theta}$ die Stimuli auf einer reellwertigen Skala einzuordnen (wodurch alle Daten kompakt zusammengefasst werden).

### 2.2 Bradley-Terry-Modell

Fast zeitgleich mit Thurstone schlug Zermelo (in deutscher Sprache) eine Methode zum Bewerten von Schachspielern basierend auf Spielergebnissen vor [Ze28]. Er betrachtete zwei Probleme: (*a*) Den Umgang mit *unausgeglichenen* Tournieren, bei denen Spieler eine ungerade Anzahl von Spielen gegen verschiedene Gegnergruppen spielen. (*b*) Das Schätzen der *relativen Stärke* der Spieler derart, dass die Schätzung Vorhersagekraft für künftige Spielergebnisse hat. Dazu führte er ein Wahrscheinlichkeitsmodell für die Spielergebnisse ein. In seinem Modell ist jeder Spieler $i \in [N]$ durch einen latenten Stärkeparameter

$\gamma_i \in \mathbf{R}_{>0}$ charakterisiert. Die Wahrscheinlichkeit, dass Spieler $i$ gegen Spieler $j$ gewinnt, ist eine Funktion der relativen Stärken der Spieler:

$$\mathbf{P}[i > j] = \frac{\gamma_i}{\gamma_i + \gamma_j}. \tag{2}$$

Es sei angemerkt, dass die Parameter $\boldsymbol{\gamma}$ nur bis auf einen multiplikativen Faktor bestimmt werden können. Aus diesem Grund wird oft angenommen, dass $\sum_i \gamma_i = 1$. Zermelo schlug vor, die Parameter $\boldsymbol{\gamma}$ durch Maximieren ihrer Likelihood angesichts der beobachteten Daten zu finden, eine für die damalige Zeit sehr fortschrittliche Idee. Er formulierte eine notwendige und hinreichende Bedingung[4] für die Existenz einer eindeutigen Maximum-Likelihood-Schätzung, entwickelte einen iterativen Algorithmus zu ihrer Ermittlung und bewies, dass der Algorithmus konvergiert. Insgesamt behandelt er das Modell sehr gründlich und vollständig; leider wurde es anscheinend für ungefähr 50 Jahre größtenteils ignoriert [Da88]. Eine spannende Einführung in die Arbeit von Zermelo findet sich bei Glickman [Gl13]. Abschließend sei angemerkt, dass das aktuell vom Weltschachbund eingesetzte Wertungssystem direkt auf Zermelos Modell basiert [El78].

**Verhältnis zu Thurstones Modell** Fast zwei Jahrzehnte später entdeckten Bradley; Terry [BT52], die Zermelos Arbeit offensichtlich nicht kannten, das Modell im Kontext der Ranganalyse von Experimenten auf der Grundlage paarweiser Vergleiche wieder, und verbanden somit das Modell wieder mit der Analyse von menschlichen Meinungen. Die Verbindung zu Thurstones Modell wurde in Bradley [Br53] deutlich, wo Bradley zeigt, dass sich durch Einsetzen von $\theta_i = \log \gamma_i$ für alle $i$ die Wahrscheinlichkeit (2) schreiben lässt als

$$\mathbf{P}[i > j] = \frac{1}{1 + \exp[-(\theta_i - \theta_j)]}. \tag{3}$$

Das Bradley-Terry-Modell (wie es in der Regel bezeichnet wird) ist somit ein weiterer Fall eines generalisierten linearen Modells [Ag15] für paarweise Vergleiche: Die Wahrscheinlichkeit eines Ergebnisses hängt von der Distanz $\theta_i - \theta_j$ zwischen zwei Parametern ab, die den Wertungen der Alternativen entsprechen. Yellot [Ye77] arbeitete die Verbindung weiter heraus, indem er zeigte, dass $\mathbf{P}[i > j]$ in (3) als $\mathbf{P}\left[x_i > x_j\right]$ für unabhängige zufällige Variablen $\{x_k : k \in [N]\}$ umgeschrieben werden kann, so dass $x_k \sim \text{Gumbel}(\theta_k, 1)$, das heißt, $\mathbf{P}[x_k \leq y] = \exp\{-\exp[-(y - \theta_k)]\}$. Ergebnisse kann man sich daher auch als den Vergleich der Realisierungen zweier zufälliger Variablen vorstellen, die um die Wertungen der Alternativen zentriert sind, woraus eine Interpretation im Rahmen von *Random Utility* folgte. Schließlich zeigte Stern [St92], dass sich das Thurstone-Modell und das Bradley-Terry-Modell zu einem einheitlichen Modell verallgemeinern lassen, wobei die Gamma-Verteilung benutzt wird. In der Praxis liefern beide Modelle in den meisten Fällen quantitativ ähnliche Ergebnisse [TG11].

[4] Die Maximum-Likelihood-Schätzung existiert, wenn und nur wenn es keinen Weg gibt, alle Spieler derart in zwei disjunkte, nicht leere Teilmengen $A, B \subset [N]$ zu unterteilen, dass es keinen Spieler in $A$ gibt, der ein Spiel gegen einen Spieler in $B$ gewonnen hat.

### 2.3 Das Auswahl-Axiom von Luce

Die zwei vorstehend besprochenen Modelle sind auf Vergleiche zwischen Elemente-*Paaren* beschränkt. Wie lassen sich diese Modelle auf *multivariate* Vergleiche verallgemeinern? Gegeben sei eine Menge von Alternativen $\mathcal{A} \subseteq [N]$ und ein Element $i \in \mathcal{A}$, und dabei bezeichne $i \succeq \mathcal{A}$ das Ereignis „$i$ wird unter den Alternativen $\mathcal{A}$ ausgewählt“. Eine natürliche Art und Weise, das Bradley-Terry-Modell (2) auf die Auswahl unter beliebig vielen Alternativen zu erweitern, lautet dann wie folgt:

$$\mathbf{P}\left[i \succeq \mathcal{A}\right] = \frac{\gamma_i}{\sum_{j \in \mathcal{A}} \gamma_j}. \tag{4}$$

Einfach ausgedrückt ist die Wahrscheinlichkeit für eine Auswahl immer proportional zu der Stärke $\gamma_i$ des Elements $i$, egal welche Menge an Alternativen vorliegt. Dieses Auswahlmodell geht auf Luce [Lu59] zurück, der zeigte, dass es eng mit der nachstehenden Eigenschaft zusammenhängt.

**Definition** (Unabhängigkeit irrelevanter Alternativen)**.** Ein probabilistisches Auswahlmodell erfüllt die Eigenschaft der *Unabhängigkeit von irrelevanten Alternativen* (UIA), wenn für jedes $\mathcal{A} \subseteq [N]$ und jedes $i, j \in \mathcal{A}$ gilt:

$$\frac{\mathbf{P}\left[j \succeq \mathcal{A}\right]}{\mathbf{P}\left[i \succeq \mathcal{A}\right]} = \frac{\mathbf{P}\left[j \succ i\right]}{\mathbf{P}\left[i \succ j\right]}.$$

Die UIA-Eigenschaft ist im Wesentlichen äquivalent[5] zum *Auswahl-Axiom* von Luce (1959, S. 6), und im Rahmen der vorliegenden Arbeit nehmen wir in austauschbarer Weise auf diese beiden Konzepte Bezug. Der wesentliche Beitrag von Luce bestand darin, zu zeigen, dass die UIA-Eigenschaft eine axiomatische Charakterisierung der Auswahl-Wahrscheinlichkeiten erlaubt.

**Satz 1** ([Lu59])**.** *Ein Auswahlmodell erfüllt die UIA-Eigenschaft, wenn und nur wenn ein Vektor $\boldsymbol{\gamma} \in \mathbf{R}_{>0}$ existiert, so dass die Auswahl-Wahrscheinlichkeiten durch* (4) *gegeben sind.*

Die Unabhängigkeit von irrelevanten Alternativen ist eine mächtige Eigenschaft, da sie zu einem Auswahlmodell führt, das *kombinatorisch* viele Auswahlwahrscheinlichkeiten mit Hilfe von lediglich $N$ Parametern repräsentiert. Dies ermöglicht die Ermittlung von Auswahlwahrscheinlichkeiten anhand einer möglicherweise kleinen Anzahl von Observationen. Es schränkt jedoch unvermeidlich die Aussagekraft des Modells ein. In Fällen, in denen einige Alternativen sehr ähnlich sind, kann UIA eine unrealistische Annahme sein, wie Debreu [De60] anhand eines einfachen Beispiels zeigt. Im Kontext moderner Anwendungen mit einer großen Anzahl von Elementen, worauf der Schwerpunkt der vorliegenden Arbeit liegt, halten wir diesen Kompromiss für akzeptabel (und vielleicht sogar für notwendig).

[5] Luce [Lu59] führt die UIA-Eigenschaft als Konsequenz des Auswahl-Axioms ein, was etwas allgemeiner ist: Seine Formulierung lässt auch $\mathbf{P}\left[i \succeq \mathcal{A}\right] = 0$ zu, ein Detail, das in der vorliegenden Arbeit unberücksichtigt bleibt.

## 3 Überblick und Beiträge

Die vorliegende Arbeit behandelt das Problem, auf *effiziente* Weise eine Rangfolge für eine Menge von Elementen zu ermitteln (was in der Regel durch das Schätzen von Auswahlmodell-Parametern erfolgt). Effizienz ist dabei der rote Faden.

- Je weiter die Größe der Datensätze ansteigt, desto wichtiger wird es, Inferenz-Methoden zu entwickeln, die *rechnerisch* effizient sind, ohne dabei Abstriche bei der *statistischen* Effizienz, d. h. der Genauigkeit, zu machen.
- Bei hohen Anzahlen unterschiedlicher Elemente wird es wichtig, Observationen mit Bedacht derart zu nehmen, dass die Observationen so viel Informationen wie möglich beitragen. Dies bezeichnen wir als *Daten*-Effizienz.

In Kapitel 2 konzentrieren wir uns auf Algorithmen für Parameter-Inferenz und entwickeln zwei Verfahren für Modelle auf der Grundlage des Auswahl-Axioms von Luce. Dazu wird das Inferenz-Problem als das Problem gefasst, die stationäre Verteilung einer Markow-Kette zu finden – ein Ansatz, der von Negahban et al. [NOS12] bereits im Kontext paarweiser Vergleiche vorgeschlagen wurde. Die Ermittlung der stationären Verteilung einer Markow-Kette ist ein gut erforschtes Problem, und es stehen schnelle Löser zur Verfügung. Zunächst wird gezeigt, wie die Markow-Kette aus der Likelihood-Funktion abgeleitet werden kann – eine wesentliche Erkenntnis, welche die Verallgemeinerung der Ideen von Negahban et al. auf andere auf dem Auswahl-Axiom von Luce basierende Modelle ermöglicht. Der erste Algorithmus, LSR, ermittelt eine *spektrale* Schätzung der Modellparameter durch Lösen einer homogenen Markow-Kette: Er ist rechnerisch sehr effizient, und die Schätzung erweist sich als akkurater als Schätzungen alternativer Methoden mit vergleichbarer Laufzeit. Der zweite Algorithmus, I-LSR, ermittelt die Maximum-Likelihood-Schätzung (MLE) durch Lösen einer nichthomogenen Markow-Kette. Die MLE ist statistisch effizienter als die spektrale Schätzung, jedoch auch rechenintensiver. Doch selbst dann erweist sich I-LSR als deutlich schneller als andere oft verwendete Algorithmen zum Ermitteln der MLE.

In Kapitel 3 wenden wir unsere Aufmerksamkeit der Aufgabe zu, auf „intelligente“ Weise Ergebnisse von paarweisen Vergleichen zu sammeln, und zwar basierend auf den observierten Ergebnissen vorheriger Vergleiche. Unter der Annahme, dass wir adaptiv auswählen können, welches Paar von Elementen zu jedem Zeitpunkt abgefragt wird, streben wir danach, die über das Modell (insbesondere über die Rangfolge der $N$ Elemente) erhaltenen Informationen zu maximieren und gleichzeitig die Anzahl zu minimieren. In der Literatur über maschinelles Lernen ist dies als das Problem des *aktiven Lernens* bekannt [Se12]. Wir starten mit einer Analyse von Quicksort [Ho62], einem bekannten Sortieralgorithmus, der eine Rangfolge berechnet, wobei Vergleiche stets mit der tatsächlichen Reihenfolge konsistent sind. Unter einigen natürlichen Annahmen über die Verteilung von Bradley-Terry-Modellparametern (welche die Schwierigkeit von Rangfolgen charakterisieren) zeigen wir, dass Quicksort erstaunlich resilient gegenüber inkonsistenten Vergleichsergebnissen ist.

Dies führt uns zu einer praktischen und dateneffizienten Abfragestrategie, die wiederholt einen Sortieralgorithmus ausführt, bis ein vorgegebenes Vergleichsbudget verbraucht ist. Im Bezug auf andere Strategien des aktiven Lernens erreicht die vorgeschlagene Methode eine vergleichbare Dateneffizienz, ist aber wesentlich weniger rechenintensiv.

In Kapitel 4 betrachten wir ein Szenario, bei dem Auswahlentscheidungen in einem Netzwerk stattfinden, was durch die Arbeit von Kumar et al. [Ku15] inspiriert ist. Es geht darum zu verstehen, wie Benutzer in einem Netzwerk navigieren (z. B. auf welche Links sie im Web klicken), und zwar unter der Annahme, dass wir Zugriff auf den aggregierten Datenverkehr an jedem Knoten im Netzwerk haben, jedoch nicht auf die individuellen Entscheidungen (d. h. die eigentlichen Übergänge). Wenn die Übergänge das Auswahl-Axiom von Luce erfüllen, können wir zeigen, dass der aggregierte Datenverkehr eine ausreichende statistische Größe für die Übergangswahrscheinlichkeiten ist. Als Nächstes entwickeln wir einen Inferenz-Algorithmus, der (*a*) robust gegenüber verschiedenen mangelhaft gestellten Szenarien ist und (*b*) effizient implementiert werden kann. Zum Beispiel skaliert der Algorithmus erfolgreich bis zu einem Schnappschuss eines WWW-Hyperlink-Graphen mit Milliarden Knoten. Schließlich zeigen wir anhand von realen Klickstream-Daten, dass die vorgeschlagene Methode Übergangswahrscheinlichkeiten gut schätzen kann, und zwar trotz der starken Annahmen, die das Axiom von Luce impliziert.

Schließlich verlassen wir in Kapitel 5 das Gebiet der menschlichen Meinungen und betrachten eine Anwendung in der Welt des Sports. Konkret beschäftigen wir uns mit der Vorhersage der Ergebnisse von Fußballspielen zwischen Nationalmannschaften. Dies ist ein schwieriges Problem, weil Nationalmannschaften nur wenige Spiele pro Jahr absolvieren, so dass ihre Stärke sich nur schwer allein aus den Ergebnissen der von ihnen gespielten Spiele schätzen lässt. Doch berücksichtigen wir, dass die meisten Nationalspieler in Spielen ihrer Clubs gegeneinander antreten, und versuchen, die (vergleichsweise) große Anzahl an Spielen zwischen den Clubs zu nutzen, um die Vorhersagen zu verbessern. Dazu stellen wir alle Spiele in einem *Spieler-Raum* dar und sorgen mit einer Kernel-Methode dafür, dass die Modellinferenz rechnerisch handhabbar wird. Wir evaluieren die erhaltene Vorhersage anhand von Daten der letzten drei Europameisterschaften. Dabei stellen wir fest, dass die Vorhersagen auf Grundlage des kombinierten Modells exakter sind als die Vorhersagen, die lediglich auf den Ergebnissen der Spiele zwischen den Nationalmannschaften beruhen.

## Literaturverzeichnis

[Ag15] Agresti, A.: Foundations of Linear and Generalized Linear Models. Wiley, 2015.

[Br53] Bradley, R. A.: Some Statistical Methods in Taste Testing and Quality Evaluation. Biometrics 9/1, S. 22–38, 1953.

[BT52] Bradley, R. A.; Terry, M. E.: Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. Biometrika 39/3/4, S. 324–345, 1952.

[Da88] David, H. A.: The Method of Paired Comparisons. Charles Griffin & Company, 1988.

[De60] Debreu, G.: Review of Individual Choice Behavior: A Theoretical Analysis. The American Economic Review 50/1, S. 186–188, 1960.

[El78] Elo, A.: The Rating Of Chess Players, Past & Present. Arco Publishing, 1978.

[Fe54] Festinger, L.: A Theory of Social Comparison Processes. Human Relations 7/2, S. 117–140, 1954.

[Gl13] Glickman, M. E.: Introductory note to 1928 (= 1929). In: Ernst Zermelo - Collected Works II. Springer, S. 616–671, 2013.

[Ho62] Hoare, C. A. R.: Quicksort. The Computer Journal 5/1, S. 10–16, 1962.

[Ku15] Kumar, R.; Tomkins, A.; Vassilvitskii, S.; Vee, E.: Inverting a Steady-State. In: Proceedings of WSDM'15. Shanghai, China, Feb. 2015.

[Lu59] Luce, R. D.: Individual Choice Behavior: A Theoretical Analysis. Wiley, 1959.

[Mc01] McFadden, D.: Economic Choices. American Economic Review 91/3, S. 351–378, 2001.

[Mc77] McFadden, D.; Talvitie, A.; Cosslett, S.; Hasan, I.; Johnson, M.; Reid, F.; Train, K.: Demand Model Estimation and Validation, Techn. Ber., Institute of Transportation Studies, University of California, Berkeley, 1977.

[NOS12] Negahban, S.; Oh, S.; Shah, D.: Iterative Ranking from Pair-wise Comparisons. In: Advances in Neural Information Processing Systems 25. Lake Tahoe, CA, Dez. 2012.

[Se12] Settles, B.: Active Learning. Morgan & Claypool Publishers, 2012.

[SL15] Salganik, M. J.; Levy, K. E. C.: Wiki Surveys: Open and Quantifiable Social Data Collection. PLoS ONE 10/5, S. 1–17, 2015.

[St92] Stern, H.: Are all linear paired comparison models empirically equivalent? Mathematical Social Sciences 23/1, S. 103–117, 1992.

[TG11] Tsukida, K.; Gupta, M. R.: How to Analyze Paired Comparison Data, Techn. Ber., Seattle, WA, USA: University of Washington, Mai 2011.

[Th27] Thurstone, L. L.: A Law of Comparative Judgment. Psychological Review 34/4, S. 273–286, 1927.

[Ye77] Yellot Jr., J. I.: The Relationship between Luce's Choice Axiom, Thurstone's Theory of Comparative Judgment, and the Double Exponential Distribution. Journal of Mathematical Psychology 15/2, S. 109–144, 1977.

[Ze28] Zermelo, E.: Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. Mathematische Zeitschrift 29/1, S. 436–460, 1928.

**Lucas Maystre** erhielt M.Sc. und Ph.D. Abschlüsse der Eidgenössische Technische Hochschule Lausanne (EPFL), Schweiz bzw. 2012 und 2018. Derzeit ist er wissenschaftlicher Mitarbeiter und Mitglied der Gruppe *Satisfaction, Interaction and Algorithms* bei Spotify, London, Vereinigtes Königreich. Seine Forschungsinteressen liegen im Bereich des maschinellen Lernens, insbesondere der probabilistischen Modellierung und des effizienten Algorithmusdesigns. Dr. Maystre wurde 2016 mit dem Google Fellowship in Machine Learning ausgezeichnet und erhielt 2018 die EPFL Auszeichnung für Abschlussarbeiten. Zusammen mit einem Kollegen der EPFL betreibt er bei `https://kickoff.ai` eine Fußball-Vorhersageplattform.

# Verhaltensbasierte Architekturkonformitätsüberprüfung[1]

Ana Nicolaescu[2]

**Abstract:** Architekturkonformitätsüberprüfung ist wichtig, um die unvermeidliche Abweichung zwischen der vorgesehenen Architekturbeschreibung und der eigentlich implementierten Architektur eines Softwaresystems zu kontrollieren. In den letzten Jahren wurden einige Ansätze für eine statisch-basierte Architekturkonformitätsüberprüfung vorgeschlagen. Diese haben jedoch erhebliche Nachteile, insbesondere dann, wenn damit moderne, technologisch heterogene und verteilte Systeme analysiert werden sollen. Eine Prüfung der Architekturkonformität solcher Systeme kann häufig nicht mit statisch-basierten Verfahren durchgeführt werden, da diese nicht immer feststellen können, ob sich ein System so verhält, wie dies vorgesehen ist. In dieser Arbeit stellen wir ARAMIS vor, einen verhaltensbasierten Ansatz zur Prüfung der Architekturkonformität, der dieses Problem löst.

## 1 Einleitung

Änderungen an Softwaresystemen werden normalerweise unter Zeit- und Kostendruck durchgeführt und sind daher nicht oder nur unzureichend dokumentiert. Sehr häufig verletzen diese auch die ursprünglichen Architekturentscheidungen zugunsten von weniger adäquaten, aber kurzfristig einfacher zu implementierenden Lösungen [Ga13], [dSB12]. Diese Tatsache mag kurzfristig akzeptabel sein. Wenn aber mittel- und langfristig keine korrigierenden Maßnahmen ergriffen werden, entwickelt sich das System sehr schnell anders als in der vorgesehenen Architekturbeschreibung [PW92], [LV95] vorgegeben. Dadurch wird die vorgesehene Architekturbeschreibung allmählich unbrauchbar oder sogar schädlich für die Unterstützung des Systemverständnisses auf einer abstrakteren, konzeptionelleren Ebene. Infolgedessen wurde eine Vielzahl von Ansätzen zum Extrahieren implementierter Architekturbeschreibungen vorgeschlagen ([DP09], [KP07]). Die meisten basieren jedoch auf Strukturanalysen der Artefakten eines Systems (Quellcode, Konfigurationsdateien usw.) und geben daher nur die statische Sicht des Systems wider. Die Komplexität moderner Systeme beruht jedoch häufig auf dem Zusammenspiel verschiedener Subsysteme und nicht nur auf ihrer Struktur [Si10].

ARAMIS ist unser verhaltensbasierter Ansatz zur Lösung des oben genannten Problems [Ni10].In einem ersten Schritt wird dabei die vorgesehene Architekturbeschreibung des zu analysierenden Softwaresystems mittels eines ARAMIS-spezifischen Metamodells erfasst. Diese Beschreibung besteht im Wesentlichen aus den Architektureinheiten des Systems und ihren Kommunikationsregeln. Um die Akzeptanz des Ansatzes zu erhöhen, werden darüber hinaus Techniken des Modell-Engineerings genutzt. Dabei wird das Ziel verfolgt, schon existierende Architekturbeschreibungen verarbeiten zu können, die nicht dem

[1] Behavior-Based Architecture Conformance Checking

[2] Research Group Software Construction, nicolaescu@swc.rwth-aachen.de

ARAMIS-Metamodell entsprechen. Im Anschluss daran werden die Interaktionen innerhalb einer ausgeführten Software aufgezeichnet, wobei vorhandene Monitoring-Systeme verwendet werden. Im Gegensatz zu den statisch-basierten Ansätzen, ist eine vollständige Analyse eines nicht trivialen Softwaresystems jedoch unmöglich. Um dem entgegenzuwirken, werden eine Reihe von Indikatoren eingeführt, die Hinweise bezüglich der Angemessenheit des analysierten Verhaltens, im Verhältnis zu einer systemweiten Konformitätsüberprüfung, liefern. Die Interaktionen innerhalb des Systems werden im nächsten Schritt analysiert. Dabei wird die Kommunikation den definierten Architektureinheiten zugewiesen und anschließend gegen die definierten Kommunikationsregeln geprüft. Dieses Ergebnis liefert die implementierte Architektur des Softwaresystems, die insbesondere auch die Abweichungen von der vorgesehenen Architekturbeschreibung definiert. Die implementierte Architektur kann vielfältig analysiert werden, so können zum Beispiel benutzerdefinierte Architektursichten und Perspektiven definiert oder dedizierte Visualisierungen genutzt werden. Abschließend werden Prozesse vorgeschlagen, die einen Leitfaden für eine verhaltensbasierte Architekturkonformitätsüberprüfung darstellen.

In Anbetracht der Beschränkungen hinsichtlich des Umfangs dieser Veröffentlichung werden wir in den nächsten Abschnitten nur eine Teilmenge unserer Ergebnisse präsentieren und den Leser für weitere Details auf die vollständige Dissertation verweisen cite nicolaescuDis. Zunächst geben wir einen Überblick über das Metamodell der ARAMIS-Architekturbeschreibungen und die Regeln, die mit der ARAMIS-Regelsprache ausgedrückt werden können. Wir fahren mit einem Überblick über die Indikatoren fort, die entwickelt wurden, um die Angemessenheit eines extrahierten Verhaltens zu analysieren, eine verhaltensbasierte Architekturkonformitätsprüfung zu unterstützen. Zuletzt stellen wir das im Rahmen unserer Forschung identifizierte Metamodell-Inkompatibilitätsproblem vor, geben einen kurzen Überblick über unsere Lösung und schließen diese Arbeit mit einer Diskussion über unsere Ergebnisse und die identifizierten Validitätsbedenken ab.

## 2 ARAMIS Architekturbeschreibungen

Da ARAMIS eine verhaltensbasierte Analyse einsetzt, muss das System überwacht und die auftretenden Interaktionen aufgenommen werden, da diese die Beziehungen zwischen den Codebausteinen des Systems abbilden. Dem hierarchischen Reflexionsmodell [KS03] entsprechend, werden diese extrahierten Beziehungen folglich auf die Elemente der beabsichtigten Architekturbeschreibung angewendet. Folglich werden damit die Konvergenzen, Divergenzen und Abwesenheiten identifiziert. In den nächsten Abschnitten stellen wir das Metamodell der ARAMIS Architekturbeschreibungen vor.

Wie im Metamodell in Abbildung 1 dargestellt, hat eine Architekturbeschreibung je nach ihrer Beziehung zum betreffenden Softwaresystem entweder eine präskriptive oder deskriptive Rolle. In ihrer präskriptiven Rolle wird die Beschreibung als die vorgesehene Architekturbeschreibung des Systems (*intended architecture description*) bezeichnet. Umgekehrt wird die Beschreibung in ihrer deskriptiven Rolle - in der sie hauptsächlich darstellt, wie das System tatsächlich aufgebaut ist - als die Beschreibung der implementierten Architektur des Systems (*implemented architecture description*) bezeichnet.

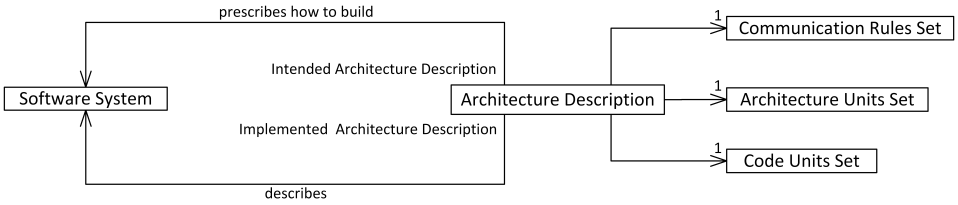



Abb. 1: ARAMIS Architekturbeschreibungen

Unabhängig von ihrer Rolle bestehen Architekturbeschreibungen aus drei Mengen: die Menge der Architektureinheiten, die Menge der Codeeinheiten und die Menge der Kommunikationsregeln. Die Menge der Codeeinheiten (*code units set*) beinhaltet alle im Rahmen einer Architekturbeschreibung definierten Codeeinheiten. Eine Codeeinheit (*code unit*) ist eine nicht-typisierte, programmiersprachenunabhängige Abstraktion eines konkreten Codebausteins.

Wie im Metamodellausschnitt in Abbildung 2 dargestellt, haben wir Codeeinheiten als eigenständige atomare Elemente modelliert. Auch wenn die Codeeinheit einen Codebaustein (z.B. Paket) darstellt, der selbst aus weiteren Codebausteinen (z.B. Klassen) besteht, enthält die Codeeinheit keine weiteren Codeeinheiten, sondern gilt als Platzhalter für alle betroffenen Bausteine. Dadurch wurde das Meta-Modell einfach gehalten und der Modellierungsaufwand wurde reduziert.

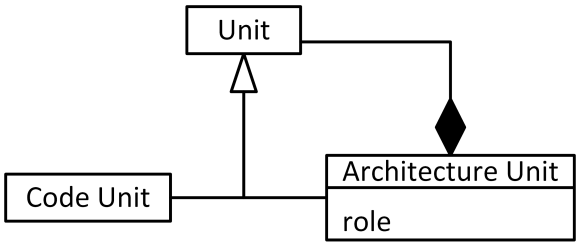



Abb. 2: Architecture Units in ARAMIS

Die Menge der Architektureinheiten (*architecture units set*) beinhaltet alle im Kontext der Architekturbeschreibung eines Systems definierten Architektureinheiten. Eine Architektureinheit (*architecture unit*) fasst Teile eines Softwaresystems zusammen, die gemeinsam eine architektonischen Bedeutung aufweisen. Eine Architektureinheit spiegelt sich nicht unbedingt explizit im Code wider. Eine Architektureinheit kann aus Codeeinheiten und weiteren Architektureinheiten bestehen. Folglich, wie in Abbildung 2 dargestellt, definiert eine Architektureinheit ein hierarchisches Konstrukt mit einem eindeutigen Identifikator. Das Architektureinheitskonzept hat eine lose Semantik. Dies steht im Gegensatz zu den meisten anderen Architekturbeschreibungssprachen, die in verwandten Arbeiten definiert sind. Der Grund dafür war, Flexibilität zu ermöglichen, da Architekten kein Standard-Meta-Modell zur Beschreibung von Architekturen verwenden. Z.B. bestehen in einigen Beschreibungen Schichten aus Komponenten, in anderen wiederum sind Komponenten

selbst geschichtet etc. Nicht zuletzt unterstützt eine lose Semantik auch die Möglichkeit, mehrere Ansichten der Architekturbeschreibung zu modellieren. Mit Blick auf das 4+1-Modell von Kruchten [Kr95], können die Architektureinheiten somit flexibel eingesetzt werden, um z.B. die Konstrukte der logischen Sicht des Systems zu modellieren, aber auch um die Prozesse in ihrer Prozesssicht darzustellen, indem Architektureinheiten für die während der Laufzeit existierenden Prozesse definiert werden.

Die Menge der Kommunikationsregeln (*communication rules set*) fasst alle Kommunikationsregeln zusammen, die die Kommunikation von Architektureinheiten regeln und werden im folgenden Abschnitt näher erläutert.

## 3 Kommunikationsregeln

Konformitätsprüfungsansätze sollten über triviale strukturelle Untersuchungen hinausgehen. Architekten äußerten oft die Notwendigkeit, Kommunikationsregeln für komplexe Systeme formulieren zu können. Diese Systeme weisen z.B. komplexe Interaktionsmuster und Kommunikationsmechanismen auf, die vor der Laufzeit nicht analysierbar sind. Bei der Gestaltung der Regelspezifikationssprache von ARAMIS war es unser Ziel, sicherzustellen, dass auch Regeln formuliert werden können, die z.B. Einschränkungen der verwendeten Kommunikationsprotokolle sowie indirekte Kopplungen und asynchrone Interaktionen ausdrücken können. Wie in [Ni10] ausführlich beschrieben, baut ARAMIS auf bestehenden Softwaremonitoren auf, um Interaktionen aus einem laufenden System zu extrahieren. Eine Interaktion wird durch ihren Aufrufer (*caller*), Aufgerufenen (*callee*) und eine Menge von Kommunikationsparametern (*communication parameters*) gekennzeichnet. Die Kommunikationsparameter stellen weitere technische Details der Kommunikation dar (z.B. das verwendete Kommunikationsprotokoll). ARAMIS verwendet dann einen auf Regulärausdrücken basierten Ansatz, um den Aufrufer und den Aufgerufenen auf entsprechenden Architektureinheiten abzubilden. Auf diese Weise entstehen mehrstufige Abstraktionen des analysierten Verhaltens. Die architektonisch abgebildeten Interaktionen können dann mit Hilfe von Kommunikationsregeln validiert werden.

Die Kommunikationsregeln stehen im Mittelpunkt der ARAMIS-Analyse. Eine taxonomische Übersicht der ARAMIS-Kommunikationsregeln ist in Abbildung 3 dargestellt. Erstens kann eine Regel je nach Berechtigungstyp eine bestimmte Kommunikation erlauben (*allow*), verweigern (*deny*) oder erzwingen (*enforce*). Darüber hinaus können die Regeln je nach Herkunft entweder spezifiziert (*specified*) (z.B. Einheit A kann Einheit B aufrufen), abgeleitet (*derived*) (Einheit A kann Einheit B aufrufen, da Einheit A in Einheit X enthalten ist und Einheit B in Einheit Y enthalten ist und es spezifiziert wurde, dass X Y aufrufen darf) oder default sein. Default Regeln regeln die Kommunikation für die Fälle, in denen keine spezifizierten oder abgeleiteten Regeln zutreffen. Darüber hinaus unterscheiden wir je nach Kommunikationsart zwischen (1) Caller-Callee Regeln, die die gerichtete Kommunikation zwischen einem Paar spezifizierter Caller- und Callee-Architektureinheiten betreffen (z.B. Schicht A darf nicht auf Schicht B zugreifen), (2) Caller-Regeln, die die von einer Einheit zu allen anderen ausgehende Kommunikation betreffen (z.B., die *utilities* Architektureinheit darf keine anderen Architektureinheiten auf-

rufen) und (3) Callee-Regeln, die sich auf die Kommunikation beziehen, die auf eine bestimmte Callee-Architektureinheit abzielt (z.B. die Fassade Schicht kann von allen anderen Architektureinheiten aufgerufen werden). Darüber hinaus können Regeln mit einem

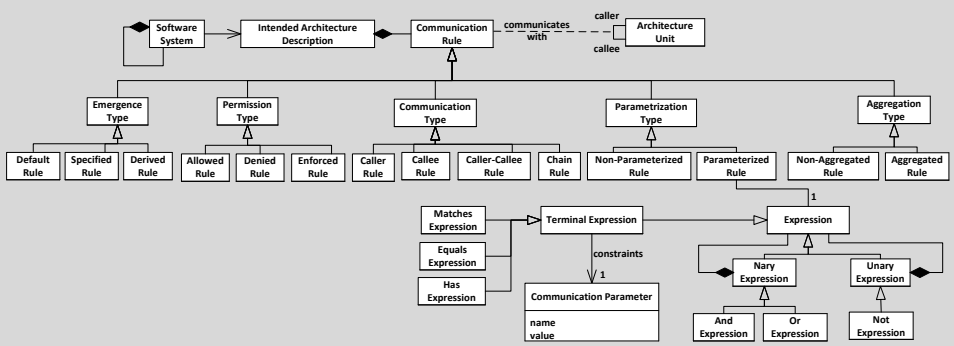



Abb. 3: Taxonomie der ARAMIS Regeln

Parameter versehen werden, um Beschränkungen für beliebigen Kommunikationsparameter zu spezifizieren. Das zugrunde liegende Modell der Kommunikationsparameter ist von dem für die Extraktion verwendeten Überwachungswerkzeug abhängig. In einem ersten Schritt werden die Interaktionen durch ARAMIS-Adapter extrahiert. Die dazugehörenden Kommunikationsparameter werden in entsprechende Schlüssel-Wert-Paare transformiert. Solche Schlüssel können z.B. die Art der Kommunikation darstellen (z.B. Queue, SOAP usw.). Folglich können Regeln eine Kommunikation erlauben oder verweigern, indem sie auf mehrere Parameter verweisen, die durch logische Operatoren verbunden sind, wie z.B. oder (*or*), und (*and*) oder nicht (*not*) (z.B. die Kommunikation zwischen zwei Architektureinheiten ist nur erlaubt, wenn die erste auf die zweite über einen restful Web-Service zugreift und wenn der Name des aufgerufenen Endpunktes mit einem bestimmten regulären Ausdruck validiert werden kann). Um die Spezifikation solcher Regeln zu ermöglichen, haben wir eine ausdruckbasierte Sprache definiert, die aus drei Arten von Ausdrücken besteht: Terminalausdrücke (*terminal expressions*) schränken die Kommunikationsparameter direkt ein. Z.B. können die Ausdrücke *matches* und *equals* verwendet werden, um eine Kommunikation zuzulassen/zu verweigern, wenn der Wert eines Kommunikationsparameters (z.B. "Queue-Name") mit einem bestimmten regulären Ausdruck übereinstimmt (z.B. <key: Warteschlangenname, Wert: matches (queueData*)>) oder einen bestimmten Wert hat (z.B. <key: Warteschlangenname, Wert: equals (queueDataTransfer)>).

N-äre (*n-ary*) und unäre (*unary*) Ausdrücke können verwendet werden, um Bedingungen durch logische Operatoren auszudrücken. Der unäre nicht (*not*) kann auf einen Ausdruck angewendet werden, um seinen booleschen Wert zu invertieren. Darüber hinaus kann man unter Verwendung eines n-ären Ausdrucks eine Kommunikation einschränken, wenn (1) mehrere ausdruckbasierte Bedingungen gleichzeitig gelten (*und* Ausdrücke—z.B. der Kommunikationstyp sollte SOAP Webservice sein und der Endpunktname sollte mit dem regulären Ausdruck "getData*"übereinstimmen) oder (2) wenn einige von ihnen gelten sollen(*oder* Ausdrücke—z.B., die Kommunikation ist erlaubt, wenn die Art der Kommunikation SOAP Webservice oder restful Webservice ist).

Darüber hinaus können wir je nach Komplexität der Kommunikationsregel zwischen aggregierenden (*aggregating*) und nicht-aggregierenden (*non-aggregating*) Regeln unterscheiden. Die nicht-aggregierenden Regeln entsprechen Konformitätsüberprüfungen, die für einzelne Interaktionen durchgeführt werden. So kann beispielsweise eine nicht aggregierende Regel voraussehen, dass Einheit A nicht auf Einheit B zugreifen sollte. Folglich kann im Falle einer architektonisch abgebildeten Interaktion diese sofort gegen diese Regel validiert werden, da keine zusätzlichen Informationen über andere Interaktionen benötigt werden. Bei einer aggregierenden Regel müssen mehrere Interaktionen berücksichtigt werden. Zwei Beispiele für solche Regeln sind (1) Einheit A sollte mit Einheit B über die Datenbank gekoppelt werden und (2) Einheit A sollte mit Einheit B über einen einzigen Kommunikationsmechanismus interagieren.

Dadurch ermöglicht ARAMIS die Definition sehr spezifischer Regeln, die die vorgesehene Kommunikation innerhalb eines Systems einschränken. Zu diesem Zweck wurde die XML-basierte Aramis-Regelsprache eingeführt, wie in [Ni10] beschrieben.

## 4 Untersuchung der Angemessenheit

Um fundierte Ergebnisse einer verhaltensbasierten Architektur-Konformitätsprüfung zu erhalten, muss auch sichergestellt werden, dass die Überwachung auf einer angemessenen Grundlage durchgeführt wird. Zu diesem Zweck haben wir eine Reihe von White-Box und Black-Box-Indikatoren vorgeschlagen. White-Box-Indikatoren geben eine Einschätzung darüber ab, inwieweit ein System aus der Sicht seiner vorgesehenen und implementierten Architekturen überwacht wurde. Im Gegensatz dazu analysieren die Black-Box-Indikatoren das Ausmaß, in dem ein System hinsichtlich seiner Spezifikation untersucht wurde.

### 4.1 White-Box Indikatoren

Auf einer ersten Analyse-Ebene kann mit jedem bekannten Codeabdeckung Metrik abgeschätzt werden, inwieweit ein System untersucht wurde. Eine hohe Codeabdeckung kann jedoch relativ einfach auf der Unit Tests-Ebene erreicht werden. Nichtdestotrotz argumentieren wir, dass man das System als Ganzes betrachten muss, wenn das Verhalten bezüglich Architekturkonformitätsüberprüfungen untersucht wird. Auf dieser Integrationsebene ist es oft schwierig, eine hohe Abdeckung zu erreichen. Zu diesem Zweck haben wir ([Ni10]) eine Reihe von Indikatoren definiert, um die folgenden Fragen zu beantworten:
**Q1.** Inwieweit wurde das System als Ganzes abgedeckt?

**Q2.** Gibt es Einheiten, die nicht abgedeckt wurden, obwohl sie in der vorgesehenen Architekturbeschreibung definiert wurden?

**Q3.** Inwieweit wurde eine bestimmte Einheit abgedeckt?

**Q4.** Inwieweit wurde die vorgesehene Kommunikation während der Überwachung ausgelöst?

Die vorgeschlagenen Indikatoren, die die oben genannten Fragen unterstützen, können dann top-down oder bottom-up analysiert werden. Wenn eine niedrige bekannte Codeabdeckung (z.B. Anweisungsabdeckung) erreicht wird, ist der Architekt möglichst skeptisch, was die Angemessenheit der durchgeführten Überwachung betrifft. Einerseits könnte er durch die top-down-Untersuchung der vorgesehenen Architektur feststellen, ob es Architektureinheiten gibt, die bei der Überwachung nicht abgedeckt wurden. Eine top-down-Traversierung der Architektureinheitenhierarchie könnte zunächst hochrangige nicht abgedeckte Architektureinheiten aufdecken. Nach der Identifizierung einer nicht oder schlecht abgedeckten Architektureinheit kann der Architekt Annahmen darüber treffen, warum diese Einheit nicht (richtig) in das überwachte Verhalten des Systems involviert war. Z.B. kann er erkennen, welches Verhalten ausgelöst werden muss, um die Beteiligung der identifizierten Architektureinheit zu erhöhen. Der Architekt kann dann seine Annahmen verfeinern, indem er die Abdeckung der tieferen Architektureinheiten genauer analysiert. Zu diesem Zweck kann er z.B. untersuchen, ob alle darin enthaltenen Codeeinheiten während der Überwachung verwendet wurden. Bei Bedarf kann er durch die Untersuchung der bekannten Codeabdeckungsmetriken der Codeeinheiten ein technisch besseres Verständnis über den Umfang der durchgeführten Überwachung erlangen. Zu jedem Zeitpunkt der Analyse kann der Architekt auch untersuchen, welche Kommunikationsregeln bei der Architekturkonformitätsprüfung verwendet wurden. Diese könnten zusätzlich die Identifizierung von Verhaltensvarianten unterstützen, die bei der Überwachung nicht ausgelöst wurden. Diese können wiederum Hinweise zur Verbesserung der Angemessenheit geben. Ergibt die zuvor vorgestellte Analyse keine wichtigen Abwesenheiten, kann der Architekt das erfasste Verhalten und die durchgeführte Konformitätsprüfung trotz des möglicherweise niedrigen Wertes der bekannten Codeabdeckungsmetriken als angemessen darstellen.

### 4.2 Black-Box Indikator

Unsere durch Fallstudien bestätigte Annahme ist, dass das extrahierte Verhalten eines Systems eine adäquate Grundlage für eine verhaltensbasierte Architekturkonformitätsprüfung darstellt, wenn es so viele relevante Szenarien des Systems wie möglich umfasst und wenn diese in ihren unterschiedlichsten Kontexten durchgeführt werden. In diesem Zusammenhang ist ein Szenario relevant, wenn es einer Schlüsselanforderung entspricht oder wenn es vom Architekten als besonders geeignet erachtet wird, das Verhalten des Systems über seine Architektureinheiten hinweg darzustellen. In [Ni10] haben wir die *Scenario Coverage Metric* definiert, um eine Abschätzung über die Angemessenheit eines extrahierten Verhaltens im Hinblick auf die Szenarien des Systems zu geben. Diese Metrik hat sich in unseren durchgeführten Fallstudien als nützlich erwiesen und basiert hauptsächlich auf den folgenden wichtigen Annahmen:

1. Die in nicht relevanten Szenarien auftretenden Architekturabweichungen haben weniger Einfluss auf die architektonische Gesamtqualität eines Systems. Die Wahrscheinlichkeit, den entsprechenden Code zu entwickeln, ist geringer: es ist wahrscheinlicher, dass die Systemteile, die den relevanten Szenarien entsprechen, weiterentwickelt werden, da neue oder sich ändernde Anforderungen oft durch ständige Nutzung ausgelöst werden.

2. Es gibt verschiedene Kontexte, in denen ein Szenario ausgeführt werden kann. Im Rahmen einer Architekturkonformitätsprüfung können diese unterschiedliche Ergebnisse liefern.

## 5 Das Meta-Modell-Inkompatibilitätsproblem

Es gibt keinen Konsens bezüglich einer universellen Sprache zur Formulierung vorgesehener Architekturbeschreibungen. Wie Malavolta betonte, würde eine solche universelle Sprache wahrscheinlich nicht einmal an Popularität gewinnen [Ma13]. Die Vielfalt der Möglichkeiten, Architekturbeschreibungen zu formulieren, spiegelt sich auch in den verfügbaren Werkzeugen zur Konformitätsprüfung wider. Diese verwenden proprietäre Metamodelle, die oft nicht erweiterbar sind und eine spezifische Semantik aufweisen. Das *Meta-Modell-Inkompatibilitätsproblem* drückt die syntaktische und/oder semantische Uneinigkeit zwischen den Sprachen aus, die einerseits von den Architekten bei der Erstellung von Architekturbeschreibungen verwendet werden und die andererseits von den verschiedenen Tools zur Architekturkonformitätsüberprüfung eingesetzt werden, um vorgesehene und/oder implementierte Architekturen darzustellen. Während das ARAMIS Meta-Modell für Architekturbeschreibungen nur sehr wenige semantische Einschränkungen aufweist, bleibt das Problem der Meta-Modell-Inkompatibilität bestehen. Wenn ein beliebiges System mit ARAMIS analysiert wird, stehen die Chancen gut, dass seine vorgesehene Architekturbeschreibung mit einem anderen Metamodell als ARAMIS definiert wurde.

Um die Einschränkung durch das Meta-Modell-Inkompatibilitätsproblem im Rahmen von ARAMIS zu verringern, haben wir den ARAMIS Architecture Description Transformation Process (AADT-Proc) entwickelt. Der AADT-Proc unterstützt die Transformation von ursprünglich vorgesehenen Beschreibungen zu ARAMIS-spezifischen Äquivalenten. Der AADT-Proc kann dazu dienen, die Architekten im Falle einer manuellen Transformation anzuleiten oder die Entwicklung von automatischen Transformationen zu unterstützen. Das Ziel von automatischen Transformationen ist, flexible Formate für die vorgesehene und implementierte Architekturbeschreibung zu ermöglichen. So könnte dann ein Architekt eine non-ARAMIS vorgesehene Architekturbeschreibung als Input für die ARAMIS-Architekturkonformitätsüberprüfung liefern und als Output das gleiche Diagramm erhalten, das aber durch zusätzliche Ergebnisse erweitert wird. Der AADT-Proc wird ausführlich in [Ni10] vorgestellt.

## 6 Diskussion

Die verhaltensbasierte Architekturkonformitätsprüfung ist in der Regel teurer als die statische Variante, da sie die Extraktion von Interaktionen aus laufenden Systemen und deren anschließende Analyse voraussetzt. Trotz dieser Einschränkung erweist sich eine verhaltensbasierte Architekturkonformitätsprüfung als nützlich, wenn man moderne Systeme betrachtet. Dieser Aspekt wurde in [Ni10] ausführlich diskutiert, sowohl auf theoretischer als auch auf praktischer Ebene. Darüber hinaus kann die Komplexität der verhaltensbasierten Architekturkonformitätsprüfung reduziert werden, wenn der verwendete Ansatz

auf existierende Performance Monitoring-Systemen aufbaut, die es bereits ermöglichen, Interaktionen aus einer Vielzahl von Systemen zu extrahieren.

Des Weiteren haben wir untersucht, welche Arten von Regeln ausdrückbar sein sollten, um das vorgesehene Verhalten eines bestimmten Systems auf einer architektonischen Ebene zu charakterisieren. Die vorgeschlagene Taxonomie war ausreichend, um die Definition aller Kommunikationsregeln zu unterstützen, die in den durchgeführten Fallstudien identifiziert wurden. Einer der wichtigsten in diesem Zusammenhang gelernten Aspekte war die Vermeidung unflexibler Lösungen. Folglich können mit Hilfe unserer Regeln Beschränkungen für jeden während der Überwachung extrahierten Interaktionsparameter festgelegt werden.

Eine weitere wichtige Erkenntnis ist, dass Kommunikation eine wichtige Rolle bei der Akzeptanz der Konformitätsergebnisse spielt. Mit unserer Model-Engineering-basierten Lösung des Meta-Modell-Inkompatibilitätsproblems erzielten wir eine Erhöhung der Akzeptanz von ARAMIS. In einer der drei durchgeführten Fallstudien lehnten aber die Architekten unsere Model-Engineering Lösung ab. Sie argumentierten, dass eine augmentierte Version der vorgesehen Architektur unübersichtlich wäre. Anstatt dessen sei eine einfache tabellarische Darstellung der Ergebnisse vorteilhafter. Die Kommunikation mit den Stakeholdern und das Verständnis ihrer Bedürfnisse können also den investierten Aufwand erheblich reduzieren.

Nicht zuletzt haben wir gelernt, dass die Codeabdeckung nicht immer eine gute Schätzung der Angemessenheit des erfassten Verhaltens darstellt. In den durchgeführten Fallstudien haben wir gezeigt, dass selbst bei geringer Codeabdeckung das extrahierte Verhalten ausreichend sein kann, um eine verhaltensbasierte Architekturkonformitätsprüfung zu unterstützen. Die Angemessenheit des Verhaltens kann stattdessen effektiver untersucht werden, indem man eine Reihe von vorgeschlagenen White-Box- und Black-Box-Indikatoren entsprechend einsetzt.

### Einschränkungen

Dieser Abschnitt gibt einen Überblick der wichtigsten Einschränkungen unserer Arbeit.

**ARAMIS baut auf bestehenden Monitoren auf, aber diese sind nicht für ihren Zweck optimiert.** Eines der Ziele von ARAMIS war es, bereits bestehende Arbeiten im Bereich Systemüberwachung wiederzuverwenden. Dies stellt aber gleichzeitig eine Einschränkung dar. Die meisten Monitore sind für die Identifizierung von Performance-Problemen optimiert. Dadurch werden erhebliche Datenmengen extrahiert, die zu Schwierigkeiten in der ARAMIS Analyse führen. Stattdessen könnten kundenspezifische, auf den Architekturbeschreibungen basierende Instrumentierungen dieses Problem weitgehend lindern.

**ARAMIS identifiziert Architekturverletzungen, kann diese aber nicht nach Priorität einordnen.** Die Identifizierung auftretender Verletzungen ist eine wichtige Aktivität,

die die Bewertung und Weiterentwicklung der Architektur unterstützt. Eine große Anzahl von festgestellten Verstößen kann jedoch zu Demoralisierung und Zurückhaltung bei der Verbesserung des Systems führen. Derzeit werden alle durch ARAMIS identifizierten Verstöße gleich behandelt. Eine automatische Priorisierung nach flexiblen Kriterien, wie z.B. Performance-Auswirkungen oder betroffenes architektonisches Niveau, könnte die beteiligten Architekten weiter motivieren, die identifizierte architektonische Abweichung schrittweise zu reduzieren.

## References

[DP09] Ducasse, Stephane; Pollet, Damien: Software Architecture Reconstruction: A Process-Oriented Taxonomy. IEEE Transactions on Software Engineering, 35(4):573–591, 2009.

[dSB12] de Silva, Lakshitha; Balasubramaniam, Dharini: Controlling Software Architecture Erosion: A Survey. Journal of Systems and Software, 85(1):132–151, January 2012.

[Ga13] Garcia, Joshua; Krka, Ivo; Mattmann, Chris; Medvidovic, Nenad: Obtaining Ground-truth Software Architectures. In: Proc.of the International Conference on Software Engineering (ICSE). IEEE Press, Piscataway, NJ, USA, S. 901–910, May 2013.

[KP07] Knodel, Jens; Popescu, Daniel: A Comparison of Static Architecture Compliance Checking Approaches. In: 6th Working IEEE/IFIP Conference on Software Architecture (WICSA), Mumbai, Maharashtra, India. S. 12–21, January 2007.

[Kr95] Kruchten, Philippe: The 4+1 View Model of Architecture. IEEE Software, 12(6):42–50, November 1995.

[KS03] Koschke, Rainer; Simon, Daniel: Hierarchical Reflexion Models. In: Proceedings of the 10th Working Conference on Reverse Engineering (WCRE). IEEE, S. 36–45, 2003.

[LV95] Luckham, David C.; Vera, James: An Event-Based Architecture Definition Language. IEEE Transactions on Software Engineering, 21(9):717–734, September 1995.

[Ma13] Malavolta, Ivano; Lago, Patricia; Muccini, Henry; Pelliccione, Patrizio; Tang, Antony: What Industry Needs from Architectural Languages: A Survey. IEEE Transactions on Software Engineering, 39(6):869–891, jun 2013.

[Ni10] Nicolaescu, Ana: Behavior-Based Architecture Conformance Checking. Shaker, 1st. Auflage, 2010.

[PW92] Perry, Dewayne E.; Wolf, Alexander L.: Foundations for the Study of Software Architecture. SIGSOFT Softw. Eng. Notes, 17(4):40–52, Oktober 1992.

[Si10] Sigelman, Benjamin H.; Barroso, Luiz Andr; Burrows, Mike; Stephenson, Pat; Plakal, Manoj; Beaver, Donald; Jaspan, Saul; Shanbhag, Chandan: Dapper, a Large-Scale Distributed Systems Tracing Infrastructure. Bericht, Google, Inc., 2010.

**Ana Nicolaescu** verteidigte ihre Doktorarbeit im September 2018 an der RWTH Aachen University. Ihre Kernkompetenzen liegen im Software-Engineering-Bereich mit dem Schwerpunkt Software-Architektur. Sie hat zuvor ihr Master-Studium an der RWTH Aachen und ihr Bachelor-Studium an der Politehnica University of Bukarest abgeschlossen.

# Durch Orientierungsschätzung unterstütztes Multitarget Tracking für optische Bandsortierer[1]

Florian Pfaff[2]

**Abstract:** Um bei optischen Bandsortierern eine hohe Zuverlässigkeit der Separation zu erzielen, sind präzise Vorhersagen der Teilchenbewegung unerlässlich. Ersetzt man die bisher übliche Zeilenkamera durch eine Flächenkamera, wird es möglich, die Teilchen mithilfe auf das Problem zurechtgeschnittener Verfahren über die Zeit hinweg zu tracken. Aus den Informationen über die Teilchenbewegung können unter Verwendung neu eingeführter Bewegungsmodelle präzise Vorhersagen abgeleitet werden. Neben den Positionen werden auch die Orientierungen der Teilchen geschätzt. Hierfür werden echtzeitfähige Schätzer beschrieben, die bisherige Verfahren in ihrer Schätzqualität übertreffen. Die vorgestellten Verfahren verwenden orthogonale Basisfunktionen. Der Einsatz variabler Anzahlen von berücksichtigten Funktionen erlaubt einen flexiblen Trade-off zwischen Laufzeit und Genauigkeit. Durch Integration der geschätzten Orientierungen in das Tracking der Teilchen kann dessen Zuverlässigkeit weiter erhöht werden. Insgesamt zeigt die Arbeit, dass der Einsatz einer Flächenkamera in Kombination mit maßgeschneiderten Algorithmen ermöglicht, die Vorhersagen und somit auch die erwartete Sortierqualität von optischen Bandsortierern signifikant zu verbessern.

## 1 Einleitung

Für viele Industriezweige sind Schüttgüter von großer Bedeutung. Schätzungen gehen davon aus, dass rund 10 % des weltweiten Energiebedarfs auf den Transport und die Verarbeitung von Schüttgütern entfällt. Ein zentraler Prozess ist das Auftrennen eines heterogenen Schüttgutstroms in unterschiedliche Klassen, beispielsweise um diesen von Verunreinigungen zu trennen. Einige Sortieraufgaben lassen sich nicht hinreichend mittels klassischer Verfahren wie dem Sieben trennen. Optische Bandsortierer bieten die Möglichkeit, Materialströme auf Basis bildgebender Verfahren aufzuspalten. Als Merkmale zur Unterscheidung von Schüttgutteilchen lassen sich nicht nur die Form und Farbe der Teilchen nutzen, sondern beispielsweise auch Unterschiede in der Infrarotstrahlung oder der radioaktiven Strahlung.

Der schier unendlichen Vielfalt unterscheidbarer Schüttgüter steht lediglich der aufwendige Separationsprozess gegenüber. In einem optischen Bandsortierer, wie ihn Abb. 1 zeigt, werden die Teilchen des Schüttguts zunächst auf ein Förderband aufgebracht. Dieses dient dazu, die Relativbewegung der Teilchen zueinander zu reduzieren und sie so weit wie möglich auf eine identische Geschwindigkeit zu beschleunigen. Nachdem die Teilchen am Ende des Bandes in eine Flugphase übergegangen sind, erreichen sie den Separationsmechanismus. Dieser besteht aus einem Druckluftdüsenbalken, der parallel zur Bandkante

[1] Englischer Titel der Dissertation: „Multitarget Tracking Using Orientation Estimation for Optical Belt Sorting“

[2] Karlsruher Institut für Technologie, pfaff@kit.edu

ausgerichtet ist. Der Düsenbalken verfügt über mehrere Ventile. Abhängig vom angesteuerten Ventil können Druckluftstöße an unterschiedlichen Stellen entlang des Düsenbalkens generiert werden, um so gezielt einzelne Teilchen auszuschleusen.

Bisher nutzten industriell eingesetzte optische Bandsortierer zur Klassifikation und Lokalisierung der Schüttgutteilchen eine Zeilenkamera, mittels derer eine Linie parallel zum Düsenbalken beobachtet werden kann. Aufgrund von Verzögerungen in der Datenverarbeitung und der Düsenansteuerung muss die Beobachtung des jeweiligen Teilchens zeitlich hinreichend vor der Separation erfolgen. Die Zeitspanne zwischen Lokalisierung und Separation muss durch eine Vorhersage überbrückt werden. Bei klassischen Systemen wird dabei typischerweise angenommen, dass die Teilchen sich ausschließlich in Transportrichtung bewegen und dabei die gleiche Geschwindigkeit aufweisen. Passt die Teilchenbewegung, wie in Abb. 2 dargestellt, nicht zu diesen Annahmen, können Druckluftstöße generiert werden, die das Teilchen nicht treffen.

Diesem Problem lässt sich durch Verwendung einer Flächenkamera begegnen. Im nächsten Abschnitt werden auf das Problem zurechtgeschnittene Verfahren erläutert, mittels derer die Trajektorien der Schüttgutteilchen erfasst und vorhersagt werden können. Um das

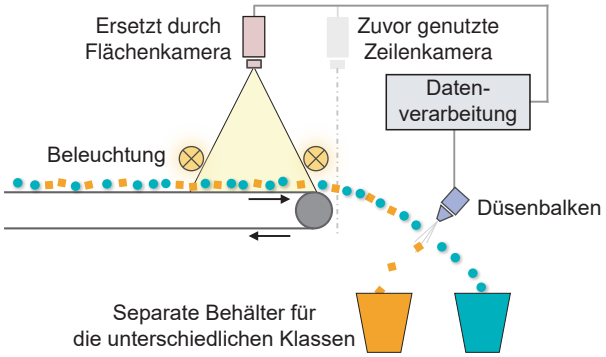



Abb. 1: Schematische Skizze eines optischen Bandsortierers. Die zuvor eingesetzte Zeilenkamera ist semitransparent dargestellt und die neu eingeführte Flächenkamera ist hellrot gefärbt.

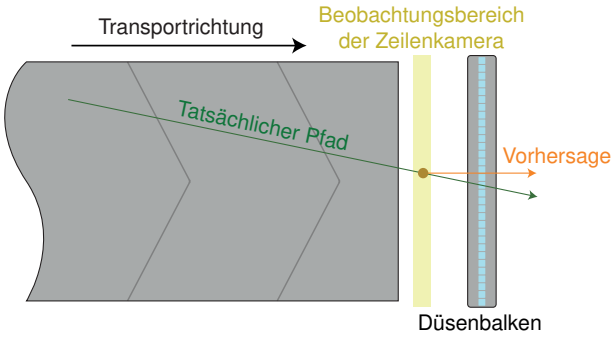



Abb. 2: Illustration der falschen Aktivierung einer Düse bei Verwendung einer Zeilenkamera.

Tracking robuster zu gestalten, wird in Abschn. 3 auf Filter zum Schätzen der Orientierungen der Schüttgutteilchen eingegangen. Eine Zusammenfassung und ein Ausblick finden sich im letzten Abschnitt.

## 2 Tracking der Schüttgutteilchen zur Verbesserung der Separation

Die Strecke von der Aufgabe der Schüttgutteilchen bis zum Erreichen des Düsenbalkens kann in unterschiedliche Phasen unterteilt werden, die in Abb. 3 illustriert sind. Ist ein Teilchen noch nicht im Sichtbereich der Kamera, so wird es lediglich durch das Band beruhigt. Hat es den Sichtbereich betreten, wird das Teilchen basierend auf regelmäßigen Beobachtungen getrackt. Verbleibt weniger als eine vorgegebene Zeit bis zum Erreichen des Düsenbalkens, muss die Separationsentscheidung getroffen werden. Die sich anschließende Prädiktionsphase muss zur korrekten Ansteuerung der Düsen mithilfe einer Vorhersage überbrückt werden. Da die Separation während des Flugs durchgeführt wird, überschneiden sich Prädiktions- und Flugphase. Im Folgenden wird zunächst betrachtet, wie Schüttgutteilchen in der Trackingphase getrackt werden können. Anschließend werden Modelle zur Überbrückung der Prädiktionsphase vorgestellt.

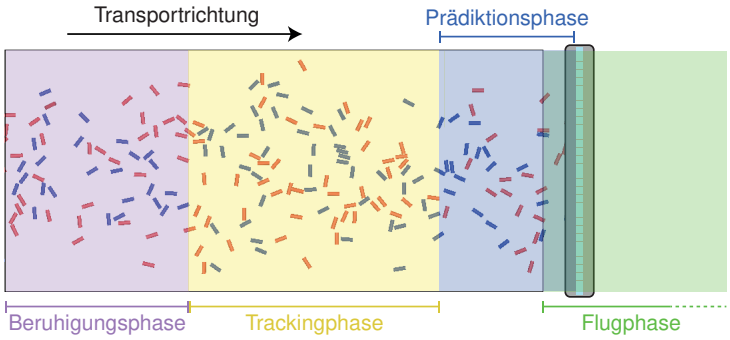



Abb. 3: Phasen des Teilchentransports.

### 2.1 Tracking der Schüttgutteilchen

Die Herausforderung, mehrere Ziele gleichzeitig zu tracken, ist in der Literatur als Multitarget-Tracking-Problem bekannt [BP99]. Verfahren, die direkt auf den Bilddaten aufbauen, sind nicht für Anwendungsfälle ausgelegt, bei denen hunderte bis tausende Objekte gleichzeitig getrackt werden müssen. Deshalb wurden die Bilddaten mithilfe geeigneter Bildverarbeitungsalgorithmen zunächst auf Punktmessungen in Form der Zentroide der Teilchen reduziert. Eine beispielhafte Pipeline besteht aus einer Background Subtraction, dem Erkennen von Einzelteilchen mithilfe von Connected-Component Labeling und der anschließenden Zentroidbestimmung. Letztere kann beispielsweise durch Berechnung des ersten Image Moments erfolgen.

Eine Herausforderung beim Tracking von Schüttgutteilchen ist, dass diese in der Regel nur mit nicht vertretbarem Aufwand eindeutig voneinander unterschieden werden können.

Somit ist bei jeder vorliegenden Messung unklar, von welchem Teilchen sie stammt. Zur Lösung dieses Problems sind Vorhersagen der Teilchenpositionen im nächsten Zeitschritt essenziell. In dieser Arbeit wurde ein Global Nearest Neighbor (GNN) eingesetzt, der eine Eins-zu-eins-Zuordnung zwischen den Vorhersagen und Messungen festlegt. Der GNN kann auch bei hohen Anzahlen an Teilchen angewendet werden und die eindeutigen Zuordnungen erlauben es, visuelle Merkmale aus unterschiedlichen Zeitschritten eindeutig einem Teilchen zuzuordnen. Somit können Informationen über das Teilchen gesammelt werden, um beispielsweise Beschädigungen zuverlässiger zu erkennen.

Bei additiven normalverteilten Rauschtermen minimiert die wahrscheinlichste Zuordnung die Summe der Mahalanobis-Distanzen der zueinander zugeordneten Vorhersagen und Messungen [Ma07, Abschn. 10.3]. Bei Mahalanobis-Distanzen werden Abweichungen entlang Richtungen hoher Unsicherheit weniger bestraft als Abweichungen entlang Richtungen geringer Unsicherheit. Um die optimale Zuordnung effizient zu finden, erstellt man eine Assoziationsmatrix, die die Mahalanobis-Distanzen aller Paare enthält. Für $N$ Teilchen und $O(N)$ Messungen lässt sich die beste Zuordnung in $O(N^3)$ finden.

Im realen Sortierbetrieb kommt es regelmäßig vor, dass Teilchen neu in den Sichtbereich eintreten oder diesen verlassen. Dies wird berücksichtigt, indem der Assoziationsmatrix zusätzliche Zeilen und Spalten hinzugefügt werden. Um passende Werte für diese zu finden, wird Wissen darüber eingebracht, wo die Teilchen in den Sichtbereich eintreten und wo sie diesen verlassen. Besagt die aktuelle Bewegungsvorhersage eines Teilchens, dass sich dieses nun außerhalb des Sichtbereichs befindet, so ist die Wahrscheinlichkeit hoch, dass es nicht erneut beobachtet wird (siehe Abb. 2). Die Wahrscheinlichkeit, dass eine Messung von einem neu in den Sichtbereich eingetretenen Teilchen stammt, ist maßgeblich von den Geschwindigkeiten der Teilchen beim Eintreten in den Sichtbereich abhängig. Eine grobe Schätzung für diese Geschwindigkeit reicht aus, um das Tracking zu initialisieren. Indem man Wissen über bereits beobachtete Teilchen nutzt, kann die Schätzung der anfänglichen Geschwindigkeit mit der Zeit verbessert werden.

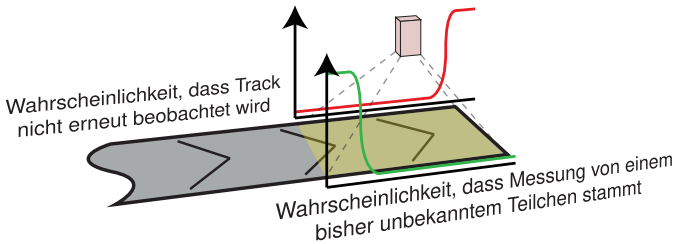



Abb. 4: Positionsabhängige Wahrscheinlichkeit, dass ein Teilchen nicht erneut beobachtet wird oder dass eine Beobachtung zu einem neuen Teilchen gehört. Der Sichtbereich der Kamera (hellrot) ist gelb dargestellt.

## 2.2 Bewegungsmodelle für die Überbrückung der Prädiktionsphase

Um die Separation zu verbessern, wird basierend auf Bewegungsmodellen vorhergesagt, wo und wann auszuschleusende Teilchen den Düsenbalken passieren. Da hierfür die Teil-

chenbewegung mehrere Zeitschritte in die Zukunft vorhergesagt werden muss, sind akkurate Bewegungsmodelle von großer Bedeutung. In der Tracking-Literatur [LJ03] finden sich Modelle wie das Constant Velocity Model und das Constant Acceleration Model. Bei ersterem wird angenommen, dass die Geschwindigkeit des bewegten Objekts gleichbleibend ist, wohingegen bei letzterem eine konstante Beschleunigung angenommen wird. Diese Modelle eignen sich nur eingeschränkt zur Modellierung der Bewegung der Schüttgutteilchen, da deren Bewegungsverhalten maßgeblich von der Differenzgeschwindigkeit zu dem Förderband abhängig ist.

Bei der Entwicklung neuer Bewegungsmodelle wurde auf Daten einer physikalischen Simulation eines Bandsortierers [Pi16] zurückgegriffen. In den Simulationsdaten liegen die Positionen der Schüttgutteilchen ohne Fehler in der Bildakquise und Bildverarbeitung vor, sodass ein Fokus auf die Teilchenbewegungen möglich ist. Im Folgenden werden unterschiedliche Modelle für die zeitliche und örtliche Vorhersage vorgestellt. Ihnen liegt die Annahme zugrunde, dass sich künftige Teilchen ähnlich wie zuvor beobachtete Teilchen verhalten.

Bei dem neuen Modell für die zeitliche Vorhersage wird angenommen, dass auf alle Teilchen in der Prädiktionsphase eine ähnliche mittlere Beschleunigung wirkt. Diese wird basierend auf den Informationen über zuvor beobachtete Teilchen approximiert. Bei geeigneter Wahl des Sichtbereichs kann der Zeitpunkt $t^{\text{Wahr}}$, an dem ein Teilchen die Koordinate $\mathrm{x}^{\text{Düsen}}$ des Düsenbalkens in Transportrichtung erreicht hat, präzise ermittelt werden. Darauf basierend löst man für alle Teilchen abhängig von der letzten nutzbaren Schätzung für die Position $\mathrm{x}^{\text{Letzte}}$ und Geschwindigkeit $\dot{\mathrm{x}}^{\text{Letzte}}$ zum Zeitpunkt $t^{\text{Letzte}}$ die Gleichung

$$\mathrm{x}^{\text{Düsen}} = \mathrm{x}^{\text{Letzte}} + (t^{\text{Wahr}} - t^{\text{Letzte}})\dot{\mathrm{x}}^{\text{Letzte}} + \frac{1}{2}(t^{\text{Wahr}} - t^{\text{Letzte}})^2\ddot{\mathrm{x}}^{\text{Optimal}} \tag{1}$$

nach $\ddot{\mathrm{x}}^{\text{Optimal}}$. Von allen $\ddot{\mathrm{x}}^{\text{Optimal}}$ der betrachteten Teilchen wird dann der Median $\ddot{\mathrm{x}}^{\text{Median}}$ gebildet. Beobachtet man nun ein neues Teilchen, so stellt man eine Bewegungsgleichung auf, indem man in (1) $\ddot{\mathrm{x}}^{\text{Optimal}}$ durch $\ddot{\mathrm{x}}^{\text{Median}}$ und $t^{\text{Wahr}}$ durch $t$ ersetzt. Durch Lösen dieser Bewegungsgleichung nach $t$ erhält man die zeitliche Vorhersage $t^{\text{Präd}}$.

Bei der örtlichen Vorhersage wird die Bewegung der Teilchen entlang der y-Achse, die parallel zur Bandkante verläuft, betrachtet. Zuerst wird für jedes der bereits beobachteten Teilchen das Verhältnis der Geschwindigkeit $\dot{\mathrm{y}}^{\text{Letzte}}$ bei der letzten nutzbaren Schätzung zu der Geschwindigkeit $\dot{\mathrm{y}}^{\text{Düsen}}$ beim Erreichen des Düsenbalkens ermittelt. Für künftige Teilchen wird angenommen, dass ein ähnlicher Anteil der Geschwindigkeit verbleibt. Indiziert man bereits beobachtete Teilchen mit einem hochgestellten Index $i$, so lässt sich eine passende Beschleunigung $\ddot{\mathrm{y}}^{\text{Verhältnis}}$ für neu beobachtete Teilchen durch die Gleichungen

$$r = \underset{i}{\text{median}}\left(\frac{\dot{\mathrm{y}}^{i,\text{Düsen}}}{\dot{\mathrm{y}}^{i,\text{Letzte}}}\right), \qquad \ddot{\mathrm{y}}^{\text{Verhältnis}} = \frac{-(1-r)\dot{\mathrm{y}}^{\text{Letzte}}}{t^{\text{Präd}} - t^{\text{Letzte}}}$$

ermitteln. Um die prädizierte Position für neue Teilchen abhängig von den zeitlichen Vorhersagen zu berechnen, verwendet man die Formel

$$\mathrm{y}^{\text{Präd}} = \mathrm{y}^{\text{Letzte}} + (t^{\text{Präd}} - t^{\text{Letzte}})\dot{\mathrm{y}}^{\text{Letzte}} + \frac{1}{2}(t^{\text{Präd}} - t^{\text{Letzte}})^2\ddot{\mathrm{y}}^{\text{Verhältnis}}.$$

Die neu hergeleiteten Modelle für die Bewegung der Schüttgutteilchen wurden mit den klassischen Modellen verglichen. Von letzteren erzielte das Constant Acceleration Model die besten Ergebnisse. Außerdem wurde ein Modell umgesetzt, das die Annahmen von Systemen basierend auf Zeilenkameras zugrunde legt. Ein Vergleich der Modelle bei einer Prädiktionsphase von 15 cm Länge ist für drei unterschiedliche Schüttgüter in Abb. 5 dargestellt. Das neue Modell für die zeitliche Vorhersage übertrifft alle anderen Modelle deutlich. Das neue Modell für die örtliche Vorhersage ist für zwei der drei Schüttgüter dem Constant Acceleration Model überlegen. Bei den Zylindern wird die mangelnde Verbesserung darauf zurückgeführt, dass deren Bewegungsverhalten orthogonal zur Transportrichtung stark von ihrer Ausrichtung abhängt. Unter anderem deshalb ist es hilfreich, die Orientierung der Schüttgutteilchen zu schätzen.

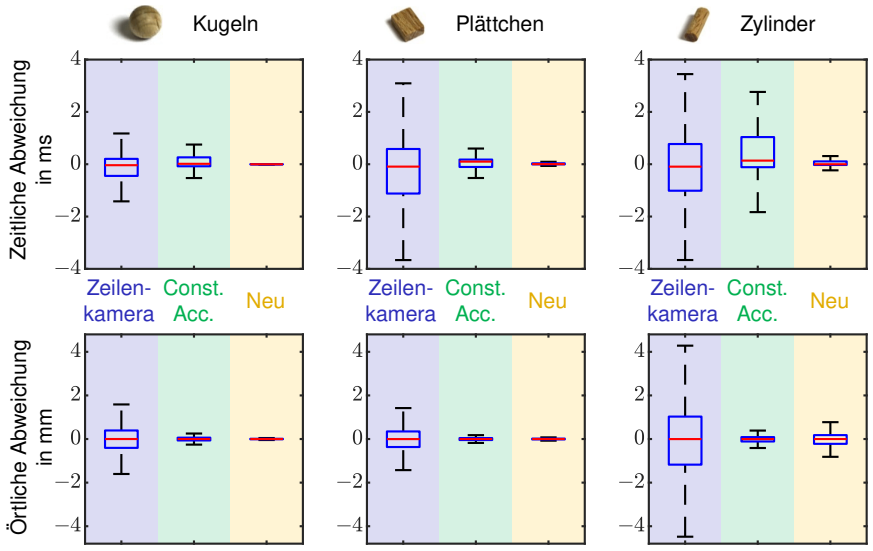



Abb. 5: Vergleich von Bewegungsmodellen bei Datensätzen mit Kugeln, Plättchen und Zylindern.

## 3 Orientierungsschätzung mittels orthogonaler Basisfunktionen

Beim Schätzen von Richtungen kann die Verwendung klassischer Verfahren, wie beispielsweise des Kalman Filters, zu schlechten Schätzergebnissen führen, da diese Wahrscheinlichkeitsmasse, die über die Periodizitätsgrenzen hinausgeht, nicht berücksichtigen. Um diesem Problem zu begegnen, wurden in der Literatur [KGH16] Filter vorgestellt, die auf periodischen Dichten basieren. Solche Filter sind jedoch dadurch eingeschränkt, dass sie sich auf eine Klasse von unimodalen Dichten fokussieren. Dies ist besonders in der Anwendung der Schüttgutsortierung limitierend, da es, wie in Abb. 6 gezeigt, zu Mehrdeutigkeiten in den Bilddaten kommen kann. Diese können in Multimodalitäten der involvierten Dichten resultieren. In diesem Abschnitt werden neuartige Filter für das Schätzen von Größen auf periodischen Mannigfaltigkeiten beschrieben, die orthogonale Basisfunktionen verwenden. Zunächst wird die Topologie des Kreises und des Hypertorus betrachtet.

Anschließend wird dargelegt, wie die Verfahren auf der Einheitssphäre angewendet werden können.

Bei der Topologie des Kreises werden trigonometrische Polynome (Fourierreihen mit endlich vielen Termen) eingesetzt. Für viele gängige Dichten wurde eine schnelle Konvergenz der Approximation beobachtet. Approximiert man eine Dichte durch ein trigonometrisches Polynom, kann es, wie in Abb. 7 dargestellt, dazu kommen, dass die Approximation negative Funktionswerte annimmt. Da dies bei Wahrscheinlichkeitsdichten unerwünscht ist, wurde auch die Approximation der Quadratwurzel der Dichten durch trigonometrische Polynome betrachtet. Quadriert man die Werte der Approximation, erhält man stets nichtnegative Funktionswerte, was sich auch im Beispiel in Abb. 7 zeigt. Zur Unterscheidung der darauf basierenden Filter wird die Variante, in der die Dichte direkt approximiert wird, als Fourier Identity Filter (IFF) bezeichnet und die Variante, bei der die Wurzel approximiert wird, als Fourier Square Root Filter (SqFF).

Rekursive Bayes-Schätzer, die auch dem Multitarget Tracking zugrunde liegen, bauen auf einem Prädiktions- und Updateschritt auf. Bei ersterem wird Wissen über den nächsten Zeitschritt $t+1$ basierend auf dem aktuellen Wissen abgeleitet, bei letzterem werden neu erhaltene Messungen in die Schätzung integriert. Verfügt man über ein Messmodell in Form einer Likelihood, lässt sich der Updateschritt mithilfe des Bayes-Theorems umsetzen. Dessen Anwendung zeigt, dass die posteriore Dichte $f_t^{\mathrm{e}}$ (die das Wissen der neuen Messung miteinbezieht) als normiertes Produkt der Likelihood $f_t^{\mathrm{L}}$ und der prioren Dichte $f_t^{\mathrm{p}}$ (die nur vorherige Messungen einbezieht) geschrieben werden kann.

Die Multiplikation der Funktionen sowie die anschließende Normierung lassen sich effizient basierend auf Fourierkoeffizienten umsetzen. Die Multiplikation zweier Funktionen entspricht einer diskreten Faltung der Koeffizientenvektoren. Eine Normierung des Ergebnisses kann durch Division aller Koeffizienten durch das Produkt des nullten Koeffizienten mit $2\pi$ erreicht werden. Somit ist die Umsetzung des Updateschritts des IFFs durch Verkettung dieser beiden Operationen darstellbar. Da die Multiplikation der Wurzeln zweier Funktionen der Wurzel der Multiplikation entspricht, kann auch beim SqFF die erste Operation durch eine diskrete Faltung umgesetzt werden. Bei der Normierung muss die rekonstruierte Dichte, die sich aus den quadrierten Funktionswerten des trigonometrischen Polynoms berechnet, normiert werden. Mithilfe des Satzes von Parseval lässt sich herleiten, dass die Normierung durchgeführt werden kann, indem jeder Eintrag des Vektors durch die Norm des Vektors multipliziert mit $\sqrt{2\pi}$ dividiert wird.

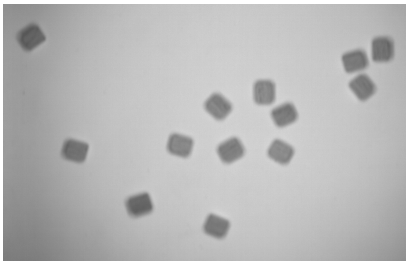


Abb. 6: Beispielbild mit Holzplättchen.

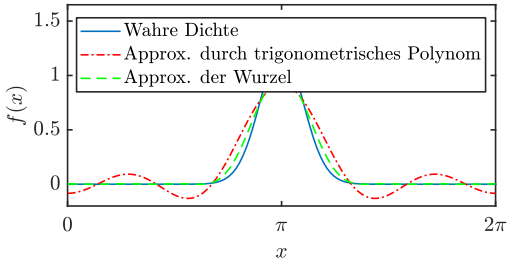



Abb. 7: Dichte und deren Approximationen.

Wird durch das Systemmodell bei dem Prädiktionsschritt lediglich ein Rauschterm $w_t$ mit Dichte $f_t^w$ auf den Zustand addiert, so lässt sich die priore Dichte für den nächsten Zeitschritt $f_{t+1}^{\mathrm{p}}$ als Faltung von $f_t^{\mathrm{e}}$ und $f_t^w$ berechnen. Diese Operation kann im IFF durch ein elementweises Produkt der Koeffizientenvektoren und einer anschließenden Multiplikation mit $2\pi$ erreicht werden. Anders als im Updateschritt sind für das SqFF weitreichende Änderungen notwendig, da die Faltung zweier Wurzeln nicht der Wurzel der Faltung entspricht. Aus den Koeffizientenvektoren, die die Wurzeln beschreiben, werden zunächst die Koeffizientenvektoren, welche die Dichten direkt beschreiben, abgeleitet. Hierfür werden die Koeffizientenvektoren für $\sqrt{f_t^{\mathrm{e}}}$ und $\sqrt{f_t^w}$ jeweils mit sich selbst gefaltet. Anschließend kann analog zum IFF ein Koeffizientenvektor für $f_{t+1}^{\mathrm{p}}$ berechnet werden. Die Fourierkoeffizienten für die Wurzel von $f_{t+1}^{\mathrm{p}}$ lassen sich jedoch nicht auf einfache Art und Weise aus den Koeffizienten für $f_{t+1}^{\mathrm{p}}$ ableiten. Um in die Repräsentation des SqFFs zurückzukehren, werden zunächst mittels einer inversen Fast Fourier Transform (FFT) die Funktionswerte von $f_{t+1}^{\mathrm{p}}$ auf einem äquidistanten Grid berechnet. Anschließend werden die Wurzeln der Funktionswerte berechnet und basierend auf diesen die Fourierkoeffizienten der Wurzel durch eine FFT approximiert.

Für $n$ Koeffizienten ergibt sich durch die diskrete Faltung für den Updateschritt eine Laufzeitkomplexität von $O(n \log n)$. Bei dem Prädiktionsschritt mit additivem Rauschen ergibt sich eine Komplexität von $O(n)$ für das IFF und, aufgrund der zusätzlichen FFT, von $O(n \log n)$ für das SqFF. Indem man die Chapman-Kolmogorov-Gleichung basierend auf Fourierkoeffizienten umsetzt, lässt sich auch ein Prädiktionsschritt für nicht additives Rauschen realisieren. Die dafür notwendigen Operationen lassen sich in $O(n^2 \log n)$ umsetzen. Alle Verfahren lassen sich durch Verwendung multidimensionaler Fourierreihen auch auf Dichten auf Hypertori anwenden. Die Laufzeitkomplexität abhängig von der Anzahl an Koeffizienten ändert sich nicht, jedoch sollten bei steigender Dimension auch signifikant mehr Koeffizienten verwendet werden. Wie viele Koeffizienten benutzt werden, kann abhängig von der verfügbaren Laufzeit und der gewünschten Genauigkeit festgelegt werden. Beim Tracking von Schüttgutteilchen ist es möglich, lastabhängig zwischen zwei Zeitschritten die Anzahl an Koeffizienten an die verfügbare Rechenzeit anzupassen.

Zur Anwendung der Verfahren auf Schätzprobleme auf der Einheitssphäre können, wie in Abb. 8 illustriert, Kugelflächenfunktionen genutzt werden. Ein einfacher Prädiktionsschritt kann durch Multiplikation bestimmter Koeffizienten in $O(n)$ umgesetzt werden. Approximiert man die Wurzel der Dichte, muss ein Umweg über ein Grid gegangen werden, was aufgrund der Transformationen zu einer Komplexität von $O(n(\log n)^2)$ führt. Da basierend auf den Koeffizienten keine effizient zu berechnende Operation verfügbar ist, die der Multiplikation der Funktionen entspricht, wurde bei dem Updateschritt in beiden Varianten ein Umweg über ein Grid genommen. Somit ist die Komplexität des Updateschritts ebenfalls $O(n(\log n)^2)$.

In mehreren Evaluationsszenarien wurden die neuen Filter mit einer für periodische Mannigfaltigkeiten angepassten Version des Partikelfilters verglichen. Beispielhafte Ergebnisse für ein Schätzproblem mit nichtlinearem Systemmodell auf dem Torus sind in Abb. 9 dargestellt. Mit nur wenigen Koeffizienten und geringer Laufzeit erreichten die Fourier Filter eine Schätzqualität, die selbst bei Verwendung von 1000 Partikeln nicht erreicht wurde.

Bei den Verfahren für den Kreis und den Hypertorus boten die Versionen mit und ohne Wurzel in mehreren Evaluationsszenarien bei ähnlichen Laufzeiten eine vergleichbare Schätzqualität. Bei der Topologie der Einheitssphäre war die Variante mit Wurzel aufgrund höherer Laufzeiten der Variante ohne Wurzel in dem betrachteten Szenario unterlegen.

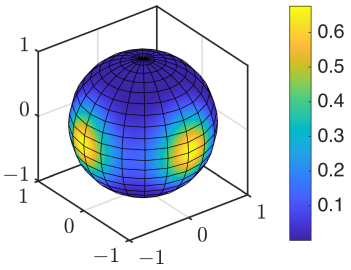



Abb. 8: Approximation einer Dichte bei Verwendung von 16 Koeffizienten.

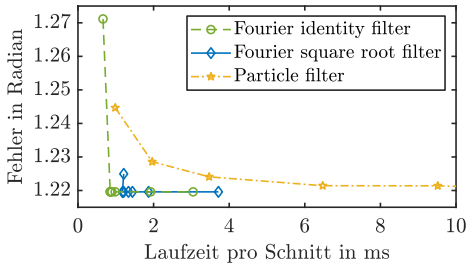



Abb. 9: Evaluationsergebnisse in einem Beispielszenario.

## 4 Integration der Orientierungsschätzung in das Schüttguttracking

Die vorgestellten Fourier Filter erlauben es, die prioren Dichten mit hoher Genauigkeit zu approximieren. Um darauf basierend geeignete Werte für die Assoziationsmatrix zu erhalten, wird die Likelihood betrachtet, dass die Messung $\hat{z}^j$ von dem Track mit Index $i$ stammt. Unter der Annahme, dass die Position eines Teilchens von dessen Orientierung unabhängig ist, zerfällt die Likelihood $\ell(\hat{z}^j|i)$ in ein Produkt der Komponenten $\ell(\hat{z}^{\mathrm{Pos},j}|i)$ für die Position und $\ell(\hat{z}^{\mathrm{Ori},j}|i)$ für die Orientierung. Die Likelihood $\ell(\hat{z}^{\mathrm{Ori},j}|i)$ lässt sich als Marginaldichte der Verbunddichte mit der Orientierung des $i$ten Teilchens $x^{\mathrm{Ori},i}$ schreiben. Somit erhält man die Formel

$$\ell(\hat{z}^{\mathrm{Ori},j}|i) = \int_0^{2\pi} \ell(\hat{z}^{\mathrm{Ori},j}, x^{\mathrm{Ori},i}|i)\, dx^{\mathrm{Ori},i} = \int_0^{2\pi} f^{\mathrm{p,Ori},i}(x^{\mathrm{Ori},i}) f^{\mathrm{L,Ori},j}(\hat{z}^{\mathrm{Ori},j}|x^{\mathrm{Ori},i})\, dx^{\mathrm{Ori},i} \,.$$

Die Multiplikation kann bei dem IFF umgesetzt werden, indem der Koeffizientenvektor der prioren Dichte mit dem der Likelihood gefaltet wird. Um die Marginalisierung umzusetzen, verwirft man alle Koeffizienten außer dem nullten und multipliziert diesen mit $2\pi$. Um selbige Schritte für das SqFF nutzen zu können, müssen lediglich die Koeffizientenvektoren bestimmt werden, welche die Dichte direkt beschreiben. Die wahrscheinlichste Zuordnung basierend auf den Orientierungen lässt sich finden, indem man eine Assoziationsmatrix aufstellt, welche die negativen Logarithmen von $\ell(\hat{z}^{\mathrm{Ori},j}|i)$ enthält. Um auch die Positionen der Teilchen zu berücksichtigen, muss lediglich eine gewichtete Kombination mit den in Abschn. 2 beschriebenen Mahalanobis-Distanzen berechnet werden.

In einem Evaluationsszenario basierend auf Simulationsdaten wurde betrachtet, wie viele Messungen falsch zugeordnet wurden. Hierbei ergab sich, dass bei geringen Messunsicherheiten in der Positionskomponente alle Zuordnungen auch ohne Einbeziehung der Orientierung korrekt waren. Bei höheren Unsicherheiten kam es zu fehlerhaften Assoziationen. Die Anzahl fehlerhafter Assoziationen konnte durch Einbeziehung der geschätzten Orientierungen signifikant reduziert werden.

## 5 Zusammenfassung und Ausblick

Im Rahmen dieser Arbeit wurde gezeigt, dass mithilfe einer Flächenkamera und einem dafür zurechtgeschnittenen Multitarget-Tracking-Verfahren Schüttgutteilchen auf einem Förderband zuverlässig getrackt werden können. Mittels für das Szenario optimierten Bewegungsmodellen konnte eine deutlich verbesserte Vorhersagegenauigkeit erreicht werden, durch die eine Erhöhung der Sortierqualität bei Realanwendungen zu erwarten ist. Die Schätzungen der Orientierungen von Schüttgutteilchen und anderen periodischen Größen konnte durch neuartige flexible Filter basierend auf orthogonalen Basisfunktionen signifikant verbessert werden. Diese Filter können auch bei starken Nichtlinearitäten der Modelle eingesetzt werden und liefern in jedem Zeitschritt eine Beschreibung der Unsicherheit in Form einer kontinuierlichen Wahrscheinlichkeitsdichte. Durch Integration der Orientierungsschätzung konnte die Zuverlässigkeit des Trackings der Schüttgutteilchen weiter erhöht werden. In künftigen Arbeiten ist eine engere Verzahnung des Trackings mit der Bildverarbeitungskomponente denkbar. Die Verfahren zur Orientierungsschätzung basierend auf orthogonalen Basisfunktionen könnten auf andere Mannigfaltigkeiten übertragen werden, um so beispielsweise Schätzer für höherdimensionale Sphären zu herzuleiten.

## Literaturverzeichnis

[BP99] Blackman, Samuel; Popoli, Robert: Design and Analysis of Modern Tracking Systems. 1999.

[KGH16] Kurz, Gerhard; Gilitschenski, Igor; Hanebeck, Uwe D.: Recursive Bayesian Filtering in Circular State Spaces. IEEE Aerospace and Electronic Systems Magazine, 31(3):70–87, März 2016.

[LJ03] Li, X. Rong; Jilkov, Vesselin P.: Survey of Maneuvering Target Tracking. Part I. Dynamic Models. IEEE Transactions on Aerospace and Electronic Systems, 39(4):1333–1364, 2003.

[Ma07] Mahler, Ronald P. S.: Statistical Multisource-Multitarget Information Fusion. Artech House, Inc., 2007.

[Pi16] Pieper, Christoph; Maier, Georg; Pfaff, Florian; Kruggel-Emden, Harald; Wirtz, Siegmar; Gruna, Robin; Noack, Benjamin; Scherer, Viktor; Längle, Thomas; Beyerer, Jürgen; Hanebeck, Uwe D.: Numerical Modeling of an Automated Optical Belt Sorter Using the Discrete Element Method. Powder Technology, Juli 2016.

**Florian Pfaff**, geboren am 13. Juni 1988, ist Postdoc am Lehrstuhl für Intelligente Sensor-Aktor-Systeme am Karlsruher Institut für Technologie. Im Jahr 2013 schloss er sein Studium ab und im Jahre 2018 seine Promotion. Beide Abschlüsse hat er mit Auszeichnung absolviert. Er war Local Arrangements Chair zweier internationaler Konferenzen, die 2016 in Deutschland stattfanden. Seine Forschungsinteressen beinhalten Tracking und Algorithmen für die Schüttgutsortierung, Filterung auf periodischen Mannigfaltigkeiten sowie Methoden zur Modellierung von Unsicherheiten.

Steffen Hölldobler et. al. (Hrsg.): Ausgezeichnete Informatikdissertationen 2018,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2019 181

# Exponentialfamilien auf Ressourcenbeschränkten Systemen[1]

Nico Piatkowski[2]

**Abstract:** Um Maschinelles Lernen (ML) in sicherheitskritischen oder autonomen Systemen einzusetzen sind Gütegarantien und Fehlerschranken erforderlich—eine rein empirische Evaluation von gelernten Modellen reicht für solche Systeme nicht aus. Insbesondere für Methoden die heuristisch motiviert sind, ist eine Herleitung solcher Garantien oft nicht möglich. Dies gilt in der Tat nicht für alle ML Verfahren: Die Exponentialfamilie ist eine Klasse probabilistischer Modelle die es uns erlaubt, das Wahrscheinlichkeitsmaß diskreter Daten zu lernen ohne dabei spezielle Annahmen über die zugrundeliegende Verteilung zu machen. Mitglieder der Exponentialfamilie bilden die Grundlage für eine Vielzahl der heute gängigen Techniken des Maschinellen Lernens, darunter Logistische Regression, Markov Random Fields und Boltzmann Maschinen. Im Gegensatz zu Ansätzen des „Deep-Learning" erlauben generative probabilistische Modelle die konsistente Schätzung der gemeinsamen Verteilung aller beteiligten Zufallsvariablen, sowie der inhärenten bedingten Unabhängigkeitsstruktur. Einmal "gelernt", ermöglichen uns solche Modelle sowohl die Klassifikation von Ereignissen, als auch das Sampling neuer Daten. Diese herausragenden theoretischen Eigenschaften werden jedoch von einer hohen Worst-Case-Komplexität begleitet. Um Exponentialfamilienmodelle in der Praxis—außerhalb von Großrechnern—einzusetzen, ist es von essentieller Bedeutung zu verstehen, unter welchen Umständen der Worst-Case tatsächlich eintritt und wie die Komplexität mit Hilfe von Approximationsalgorithmen verringert werden kann. Daher wird in dieser Arbeit die Klasse der Exponentialfamilien systematisch bezüglich ihres Ressourcenbedarfs untersucht. Aus der formalen Spezifikation der Modellklasse werden Approximationen hergeleitet, die den Speicherverbrauch, die erforderlichen arithmetischen Instruktionen sowie die Berechnungskomplexität der gesamten Modellklasse reduzieren. Dabei wird stets sichergestellt, dass der zusätzliche Fehler, der durch unsere Approximationen entsteht, beschränkt ist. Die theoretischen Erkenntnisse werden auf Datensätzen realer Anwendungen validiert.

[1] Englischer Originaltitel: "Exponential Families on Resource-Constrained Systems"

[2] TU Dortmund, Lehrstuhl für Künstliche Intelligenz, nico.piatkowski@tu-dortmund.de

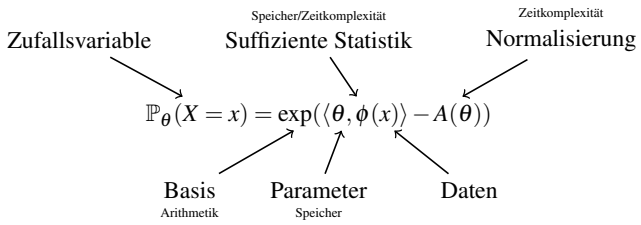



Abb. 1: Wesentliche Bestandteile eines Exponentialfamilienmodells für eine ($n$-dimensionale) Zufallsvariable $X$ sowie die hauptsächlich verbrauchte Ressource.

# 1 Einleitung

Die Künstliche Intelligenz und insbesondere das Maschinelle Lernen erfahren zur Zeit große Aufmerksamkeit in Medien und Politik. Doch auch wenn geringe Klassifikationsfehler mit teilweise übermenschlicher Performanz erreicht werden, reicht eine rein empirische Evaluation für viele Anwendungen nicht aus: Für den Einsatz von ML in sicherheitskritischen oder autonomen Systemen sind Gütegarantien und Fehlerschranken erforderlich. Beispielhafte Szenarien sind autonome Kraftfahrzeuge, Drohnen, oder mobile medizinische Geräte.

Ferner ist der Ressourcenverbrauch von state-of-the-art ML Methoden oft auf Großrechner ausgelegt: Auch wenn es zuerst beeindruckend klingt, dass IBM's Watson „Jeopardy!"-Sieger geworden ist oder das Google DeepMind's AlphaGo einen der weltbesten Spieler des Brettspiels Go besiegen konnte, so sind diese Erfolge weniger beeindruckend wenn man den Ressourcenverbrauch betrachtet. Watson bestand zum Zeitpunkt des Sieges aus neunzig Servern und AlphaGo bestand aus einem System mit 1920 CPUs sowie 280 GPUs. Der genaue Energieverbrauch beider Systeme ist unbekannt, aber die verwendete Hardware legt nahe, dass ihre Leistungsaufnahme bei ca. einhunderttausend Watt liegt—die durchschnittliche Leistungsaufnahme des menschlichen Gehirns liegt bei ca. 20 Watt.

Beide Punkte sind für diese Arbeit von zentraler Bedeutung: Zum einen müssen wir Garantien über das Lernverhalten abgeben können, beispielsweise in Form von asymptotischer Konsistenz oder einer beschränkten Abweichung vom optimalen Lernergebnis. Zum anderen müssen wir den Ressourcenverbrauch von Lernverfahren verstehen und gegebenenfalls an die verfügbare Hardware anpassen können.

Die Menge *aller* Lernverfahren ist selbstverständlich zu groß und zu unstrukturiert für eine systematische Analyse. Wir beschränken uns hier auf die Klasse der Exponentialfamilien, um sowohl die theoretischen Eigenschaften als auch den Ressourcenverbrauch systematisch zu untersuchen. Die funktionale Form sowie die wesentlichen Bestandteile von Verteilungen der Exponentialfamilie sind in Abbildung 1 dargestellt. Die kanonische Verlustfunktion von Exponentialfamilien ist die *negative log-Likelihood* $-\ell(\theta) = -\sum_{x \in \mathscr{D}} \log \mathbb{P}_\theta(x)$. Beim "Lernen" wird diese Funktion bezüglich der Parameter $\theta \in \mathbb{R}^d$ mittels numerischer Optimierungsverfahren minimiert.

Die Klasse der Exponentialfamilienmodelle enthält wichtige Modelle des Maschinellen Lernens, darunter Markov Random Fields, Conditional Random Fields, Lineare- und Logistische Regression und Boltzmann-Maschinen, sowie viele bekannte Wahrscheinlichkeitsverteilungen wie die Normalverteilung, die Poisson-, die Dirichlet- und die Kategorische-Verteilung. Jeder Fortschritt der auf dem Gebiet der Exponentialfamilien erzielt wird, wirkt sich also direkt auf eine Vielzahl von theoretischen Modellen und praktischen Anwendungen aus.

Im Folgenden widmen wir uns drei zentralen Fragestellungen, deren Untersuchung uns letztendlich erlauben wird, Exponentialfamilienmodelle selbst auf sehr schwachen (sog. "ultra-low-power") Berechnungsarchitekturen mit beschränktem Approximationsfehler zu lernen und anzuwenden. Wir beschränken uns hier jeweils auf die Kernresultate in Form von Theoremen sowie einige experimentelle Resultate. Die zugehörigen Beweise sowie vollständige Beschreibung des experimentellen Designs können in [Pi18] gefunden werden. Alle der im Folgenden gezeigten Experimente wurden auf den folgenden drei Datensätzen durchgeführt:

- INSIGHT—SCATS Daten der Stadt Dublin: Der Datensatz enthält Messungen von 2367 SCATS (Sydney Coordinated Adaptive Traffic System) Verkehrs Sensoren, welche in diversen Straßen der Stadt Dublin eingebettet sind. Die Daten wurden zwischen dem 1. Januar 2013 und dem 14. Mai 2013 erhoben. Jeder Sensor gibt pro Minute die Durchschnittsgeschwindigkeit sowie die relative Sensorbelegung (Verkehrsdichte) aus.
- VaVeL—Mobilfunknetzwerknutzung der Stadt Warschau: Der Datensatz enthält die Auslastung von 4988 Zellen des Mobilfunknetzes der Stadt Warschau. Die Daten wurden zwischen dem 15. Mai 2016 und dem 26. Juni 2016 erhoben. Jeder Datenpunkt enthält Ereignisse über An- und Abmeldungen der Nutzer sowie Nutzung von Sprach- und xMS Kanälen. Die Daten jeder Mobilfunkzelle wurden stundenweise aggregiert.
- Intel Lab—Temperatur und Luftfeuchtigkeit: Der Datensatz enthält Messungen der Temperatur und Luftfeuchtigkeit von 56 Sensoren im Intel Berkeley Research Lab. Die Daten wurden zwischen dem 28. Februar und dem 2. April 2004 erhoben. Jeder Sensor erzeugt alle 31 Sekunden eine neue Messung.

Die Daten jedes Datensatzes wurden in gleichlange Sequenzen über 24 Stunden partitioniert. Für jeden Datensatz wurde die bedingte Unabhängigkeitsstruktur, welche die stochastische Interaktion zwischen den Sensoren repräsentiert, mit dem Chow-Liu-Algorithmus [CL68] bestimmt.

## 2 Modelkompression durch Regularisierte Reparametrisierung

Im ersten Beitrag geht es um das Lernen komprimierter Modelle und die damit verbundene Reduktion des Speicherverbrauchs. Ist die zugrundeliegende Zufallsvariable hochdimensional, so müssen oft mehrere Millionen Modellparameter gelernt werden. In Abhängigkeit

von der gelernten Struktur kann die Anzahl der Parameter für die VaVeL Daten $310 \times 10^6$ übersteigen. Probabilistische Modelle erlauben eine Interpretation der gelernten Gewichte als bedingte Transinformation zwischen Zufallsvariablen. Bei genauerer Betrachtung fällt auf, dass viele der Parameter redundant sind. Motiviert durch diese Beobachtung wurde die folgende Reparametrisierung untersucht [PLM13]:

$$\theta(\Delta) = D\Delta \tag{1}$$

Hierbei ist $\theta$ der klassische Parametervektor, $\Delta$ ist der neue Vektor, und $D$ ist eine unitrianguläre untere Dreiecksmatrix mit Einträgen in $[0;1]$. Intuitiv kann die Matrix $D$ so gewählt werden, dass jede Komponente von $\theta$ eine Linearkombination von Komponenten von $\Delta$ sind. Anstatt $\theta$ wird nun $\Delta$ mittels numerischer Minimierung von $-\ell(\Delta) = -\sum_{x\in\mathscr{D}} \log \mathbb{P}_{\theta(\Delta)}(x) + \lambda_1\|\Delta\|_1 + \lambda_2\|\Delta\|_2^2$ gelernt. Durch die normbasierte *Regularisierung* von $\Delta$ ist sichergestellt, dass Redundanzen in $\theta$ mittels einiger weniger Dimensionen von $\Delta$ abgedeckt werden—der Vektor $\Delta$ ist spärlich besetzt wann immer $\theta$ viele Redundanzen enthält. Für (1) haben wir gezeigt, jedes $\theta$ verlustfrei dargestellt werden kann. Die Reparametrisierung (1) ist also *universell* und kann auch eingesetzt werden, falls das optimale Modell gar keine Redundanzen enthält. Zusätzlich haben wir gezeigt, dass bei korrekter Wahl von $\lambda_1$ und $\lambda_2$ der Abstand des gelernten Modells zum optimalen Modell stets beschränkt ist:

**Theorem 2.1 (Beschränkter Fehler)** *Sei $X$ eine Zufallsvariable mit Exponentialfamilienendichte und Parameter $\theta^* \in \mathbb{R}^d$, dessen Reparametrisierung minimale Norm besitzt. Sei $\mathscr{D}$ ein Datensatz mit $N = |\mathscr{D}|$ Realisierungen von $X$. Nimm an, dass $\|\nabla^2 A(\theta^*)^{-1}\|_\infty \leq \kappa$ und $\|\Delta\|_\infty \leq \gamma$, setze $\lambda_1 = 4T\sqrt{\log(d)/N}$ und $\lambda_2 = \gamma^{-1}\lambda_1$. Wenn $N \geq 324\kappa^4 d^{12}\log(d)/(T - d^2)^2$, dann gilt für jede Wahl von $D$:*

*(I) Der Abstand zwischen dem optimalen Modellparameter $\theta^*$ und dem gelernten Schätzer $\eta_D(\hat{\Delta})$ ist beschränkt: $\|\eta_D(\hat{\Delta}) - \theta^*\|_\infty \leq 3\kappa d^2\lambda_1$,*

*(II) jede Spärlichkeit im gelernten Modell impliziert eine Redundanz im optimalen Modellparameter: $\hat{\Delta}_{C=x'}(t) = 0 \Rightarrow$*

$$|\theta^*_{C=x'}(t-1) - \theta^*_{C=x'}(t)| \quad \leq \quad \frac{3d^2\kappa\lambda_1}{T} + (t-1)\left(\max_{i=1}^{t-1} |\hat{\Delta}_{C=x'}(i)| + \frac{3d^2\kappa\lambda_1}{T}\right),$$

*für jede Clique $C$ der bedingten Unabhängigkeitsstruktur und jeden Zeitpunkt $t$. Beide Aussagen gelten mit einer Wahrscheinlichkeit von mindestens $1 - (2/d)$.*

Experimentelle Ergebnisse sind in Abb. 2 dargestellt. Dort bezeichnet M1 das Resultat einer gewöhnlichen Exponentialfamilienschätzung, M2 bezeichnet das Resultat einer $l_1$-regularisierten Exponentialfamilie und alle anderen Resultate entsprechen der hier vorgestellten regularisierten Reparametrisierung mit verschiedenen *Zerfallsmatritzen* $D$. Jeder Punkt ist einen Mittelwert—insgesamt wurden 32805 einzelne Experimente durchgeführt. Die Kompression wird hier durch die Parameterdichte $\rho = (\sum_{i=1}^{d} 1(\Delta_i \neq 0))/d$ gemessen,

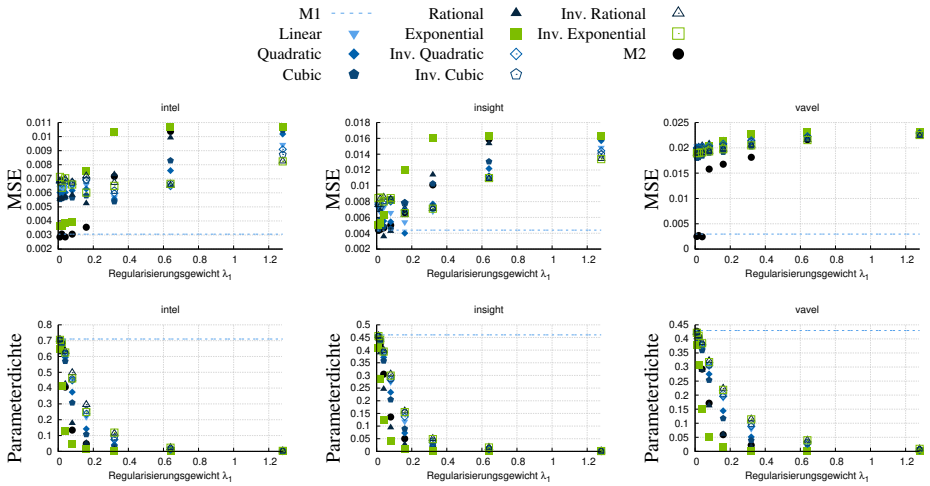



Abb. 2: Experimentelle Ergebnisse der regularisierten Reparametrisierung. Zu sehen ist der mittlere quadratische Fehler (MSE) der geschätzten Randwahrscheinlichkeiten sowie die Dichte der Reparametrisierung $\Delta$ als Funktion von $\lambda_1$. Verschiedene Farben indizieren verschiedene Wahlen für $D$. Schwarze Kreise zeigen Resultate für das Modell ohne Reparametrisierung.

also die relative Anzahl der Parameter die nicht Null sind. Je kleiner dieser Wert, umso größer die Kompression des Modells. Je größer der Wert von $\lambda_1$, desto stärker versucht das Lernen ein möglichst kleines Modell zu finden. Wir sehen, dass für Kompressionsstärken zwischen 0.2 und 0.8 der Fehler unserer reparametrisierten Modelle vergleichbar oder geringer als der Fehler des M2-Modells ist, während eine ähnliche oder höhere Kompression erreicht wird. Die Ergebnisse zeigen auch, dass das Model der VaVel Daten am wenigsten Redundanzen enthält, da die Kompression hier zu einem größeren Fehler führt.

## 3 Lernen und Inferenz ohne Gleitpunktarithmetik

Im zweiten Beitrag geht es um das Lernen ganzzahliger Modelle und die damit verbundene Reduktion der erforderlichen Schaltwerkskomplexität. Kleine Geräte oder Sensoren sind oft nur mit schwachen Mikroprozessoren ausgestattet, um einen möglichst geringen Energieverbrauch zu garantieren. Aus diesem Grund wird oft auf zusätzliche Hardware wie Gleitkommakoprozessoren verzichtet. Aus der Modellspezifikation (Abb. 1) ist jedoch klar, dass die Auswertung von Exponentialfamilienmodellen, und damit auch das Lernen und die Anwendung des Modells, inhärent auf der Auswertung transzendenter Funktionen basiert. Bei genauerer Betrachtung der Herleitung der Exponentialfamilie [Pi36] fällt auf, dass diese äquivalent zu jeder anderen Basis hergeleitet werden kann:

**Theorem 3.1 (Basis-$b$ Familien)** *Jedes Mitglied der Exponentialfamilie kann in Basis-$b$-Form gebracht werden, d.h.,* $\mathbb{P}_{b,\theta}(X = x) = b^{\langle\theta,\phi(x)\rangle - A_b(\theta)}$ *mit* $A_b(\theta) = \log_b Z_b(\theta) =$

$\log_b \int_{\mathscr{X}} b^{\langle\theta,\phi(x)\rangle}\,\mathrm{d}\nu(x)$. *Ein Wechsel der Basis ist mittels der Reparametrisierung* $\eta_{a,b}(\theta) = (\log b/\log a)\theta$ *möglich:* $\forall x \in \mathscr{X} : \mathbb{P}_{a,\theta}(x) = \mathbb{P}_{b,\eta_{a,b}(\theta)}(x)$.

Durch die Wahl von $b = 2$ sowie der Nebenbedingung $\theta \in \mathbb{N}^d$ erhalten wir auf natürliche Weise eine Teilklasse der Exponentialfamilien, für welche der Abstand zur Klasse aller Exponentialfamilienmodelle beschränkt ist:

**Theorem 3.2 (Log-Likelihood Fehler)** *Sei* $\lfloor\theta\rfloor$ *das elementweise Abrunden von* $\theta \in \mathbb{R}^d$ *und sei ferner* $\varepsilon = \theta - \lfloor\theta\rfloor$. *Nimm an, dass* $\theta$ *der Parameter einer Exponentialfamilie mit übervollständiger suffizienter Statistik ist. Dann gilt* $\ell(\lfloor\theta\rfloor;\mathscr{D}) - \ell(\theta;\mathscr{D}) \leq 2\|\varepsilon\|_2|\mathscr{C}(G)|$, *wobei Gleichheit erreicht wird wenn* $\theta$ *bereits ganzzahlig war.*

Hierbei bezeichnet $|\mathscr{C}(G)|$ die Anzahl der Cliquen in der bedingten Unabhängigkeitsstruktur. Durch die Kombination von Basis-2 Familien mit ganzzahligen Parametern kann die transzendente exp-Funktion außerdem durch einen einfachen Bit-shift ersetzt werden:

**Lemma 3.1 (Bit-Shift Potentialfunktion)** *Die Potentialfunktion* $\psi_2(x) = 2^{\langle\theta,\phi(x)\rangle}$ *der Basis-*2 *Familie kann via Bit-shift "von links" (*$a \ll b$ *für* $a,b \in \mathbb{N}$*) ausgewertet werden:* $\psi_2(x) = 1 \ll \langle\theta,\phi(x)\rangle$. *Reellwertige Arithmetik ist nicht erforderlich.*

Die Berechnung von Wahrscheinlichkeiten in Exponentialfamilienmodellen erfolgt mit Hilfe von Inferenzalgorithmen. Für die Basis-2-Familie haben wir einen speziellen Inferenzalgorithmus entwickelt, der die Wahrscheinlichkeiten allein mittels ganzzahliger Arithmetik approximieren kann. Kern des Algorithmus ist eine Datenstruktur für dünnbesetzte ganzzahlige Daten sowie die Beobachtung, dass der Logarithmus zur Basis-2, dessen Auswertung für die Inferenz erforderlich ist, durch die Bitlänge einer ganzen Zahl approximiert werden kann. Für die Methode haben wir gezeigt, dass der Approximationsfehler durch Eigenschaften der zugrundeliegenden bedingten Unabhängigkeitsstruktur beschränkt ist, insbesondere durch die Länge des längsten Pfades im Modell. Beispielhafte experimentelle Ergebnisse sind in Abb. 3 dargestellt. Dort vergleichen wir exakte und approximative-ganzzahlige Inferenz auf zwei synthetischen bedingten Unabhängigkeitsstrukturen. Zum einen „chain", welche einer linearen Liste entspricht, und „star", ein sternförmiger Graph. Die Länge des längsten Pfades in "star" ist konstant 2 (unabhängig von der Anzahl der Variablen), während die längste Pfad in "chain" gleich der Anzahl an Variablen ist. Die Modellparameter wurden zufällig Normalverteilt mit Mittelwert 0 und verschiedenen Varianzen erzeugt und abschließend zum nächst kleineren ganzzahligen Wert gerundet. Die Ergebnisse spiegeln zum einen unsere theoretische Einsicht wieder, dass der Fehler hauptsächlich durch die Länge des längsten Pfades beeinflusst wird. Zum anderen sehen wir, dass die Einschränkung auf ganzzahlige Modellparameter den Raum der Wahrscheinlichkeiten diskretisiert: das Modell kann nur eine kleine endliche Teilmenge an verschiedenen Wahrscheinlichkeiten erzeugen.

Da wir nun in der Lage sind, Inferenz in ganzzahligen Modellen mit beschränktem Fehler zu betreiben, ist der letzte Schritt die Herleitung einer Methode für das Lernen ganzzahliger Modellparameter. Dazu führen wir eine neue Regularisierung ein:

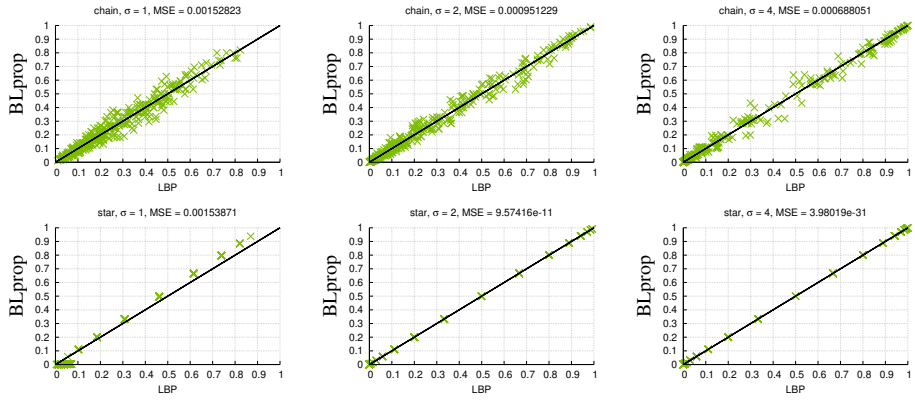



Abb. 3: Vergleich der berechneten Randwahrscheinlichkeiten von exakter Inferenz (LBP) sowie der approximativen ganzzahligen Inferenz (BLprop) auf zwei bedingten Unabhängigkeitsstrukturen („chain" und „star") für zufällige Parameter mit unterschiedlicher Standardabweichung $\sigma \in \{1,2,4\}$.

**Definition 3.1 (Integer-Regularisierung)** *Die Integer-Regularisierung bestraft den Abstand jedes Modellparameters zum nächsten ganzzahligen Wert:*

$$R_{int}(\theta) = \sum_{i=1}^{d} \rho_{int}(\theta_i) \qquad mit \qquad \rho_{int}(\theta_i) = 1 - |1 - 2(\lceil \theta_i \rceil - \theta_i)| \,.$$

Das Modell wird also durch die numerische Minimierung von $-\ell(\theta) = -\sum_{x \in \mathscr{D}} \log \mathbb{P}_\theta(x) + \lambda R_{\mathtt{int}}(\theta)$ gelernt. Da diese Regularisierung sowohl nicht-stetig als auch nicht-konvex ist, musste ein spezieller proximaler-Gradientenabstieg [PB14] inklusive Konvergenzbeweis hergeleitet werden. Entsprechende Ergebnisse sind in Abb. 4 dargestellt. Hier wurde zusätzlich die maximale Bitlänge ($b$) der gelernten ganzzahligen Parameter eingeschränkt, um zu zeigen, dass keine (triviale) Fixpunktdarstellung der reellwertigen Parameter gelernt wird. Tatsächlich zeigen die Ergebnisse, dass bereits drei Bits ausreichen, um Modelle mit geringem Approximationsfehler zu lernen, während die Laufzeit um ein Vielfaches geringer ist. Hierbei ist zu beachten, dass die Laufzeit auf einer gewöhnlichen Workstation gemessen wurde. Im letzten Abschnitt gehen wir kurz auf Ergebnisse ein, die auf einem Mikrocontroller erzielt werden konnten. Zusätzliche Ergebnisse zu Klassifikationsexperimenten können in [PLM16] gefunden werden.

## 4 Lernen und Inferenz durch Stochastische Quadratur

Im dritten Beitrag befassen wir uns mit der Laufzeitkomplexität der probabilistischen Inferenz. Dies entspricht der Auswertung der Normalisierungskonstante $A(\theta) = \log Z(\theta)$ aus Abb. 1. Es kann gezeigt werden, dass dieses Problem #P-vollständig ist [Va79]. Hierbei ist es wichtig zu verstehen, dass diese Eingruppierung der Komplexität den Worst-Case über die Wahl der bedingten Unabhängigkeitsstruktur widerspiegelt. Aus diesem Grund

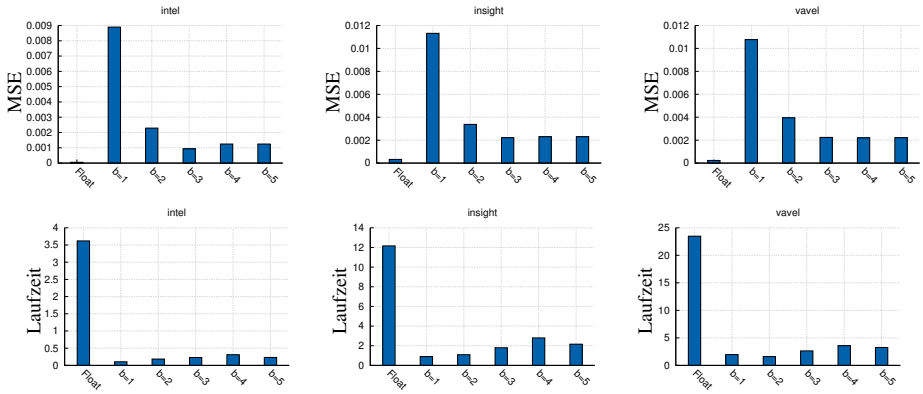



Abb. 4: Experimentelle Ergebnisse der ganzzahligen Exponentialfamilienmodelle. Zu sehen ist der mittlere quadratische Fehler (MSE) der geschätzten Randwahrscheinlichkeiten sowie die Laufzeit als Funktion der maximalen Bitlänge (Wortbreite) pro Parameter. „Float" bezeichnet das gewöhnliche (nicht ganzzahlige) Exponentialfamilienmodell.

basieren die meisten Approximationsalgorithmen für probabilistische Inferenz auf einer Approximation dieser Struktur und damit auf einer Elimination von Abhängigkeiten, die zwischen Zufallsvariablen existieren. Da die Berücksichtigung aller Abhängigkeiten aber gerade die Eigenschaft ist, die es uns erlaubt, Garantien über das Lernergebnis abzugeben, haben wir versucht, eine andere Art der approximativen Inferenz zu finden, bei der die Struktur intakt bleibt. Unser neuer Ansatz [PM16] basiert auf dem Konzept der Quadratur:

$$I[f] = \int_l^u f(x)\mathrm{d}x \approx \int_l^u h_k(x)\mathrm{d}x = \sum_{i=0}^{k} w_i f(x_i) = I_k[f] \tag{2}$$

Hierbei ist $I[f]$ ein Integral über $f$, dessen Auswertung aufgrund zu hoher Komplexität nicht möglich ist. Stattdessen wird eine Approximation von $f$ so gewählt, dass das Integral doch berechnet werden kann. War der Approximationsfehler zwischen $f$ und $h_k$ beschränkt, so ist auch der Fehler der Integralapproximation beschränkt. Dies wenden wir nun auf die Normalisierungskonstante $Z(\theta)$ an. Dazu wählen wir als $h_k$ eine Chebychev-Polynominterpolation von $\exp(\langle\theta,\phi(x)\rangle)$ zum Grad $k$. In diesem Fall spricht man von einer Clenshaw-Curtis Quadratur [CC60]. Da die resultierende Quadraturapproximation von $Z(\theta)$ einem mehrdimensionalen Polynom mit $d^k$ Summanden entspricht, berechnen wir die endgültige Approximation auf einer zufälligen Stichprobe von Summanden—dieses Vorgehen wird auch *Stochastische Quadratur* genannt. Da das Erstellen einer zufälligen Stichprobe möglichst effizient sein muss, haben wir einen spezialisierten Monte-Carlo Stichprobenalgorithmus entwickelt [PM18]. Für die endgültige Prozedur konnten wir folgende Fehlerschranke herleiten:

**Theorem 4.1 (Fehlerschanke)** *Sei $\zeta$ der Koeffizientenvektor einer Grad-$k$ Chebychev-Polynomapproximation von* exp *auf $[l;u] = [-\|\theta\|_1;+\|\theta\|_1]$ mit Worst-Case Fehler $\varepsilon$. Sei $\hat{Z}_\zeta^{N,k}(\theta)$ die Ausgabe des neuen Inferenzalgorithmus'. Sei ferner $\delta \in (0,1]$, $\varepsilon > 0$,*

$N = (\log 2/\delta)\tau^2 2\|\theta\|_\infty^{2k'}\varepsilon^{-2}|\mathscr{X}|^{-2}$, *mit* $(k-1)\,k! \geq 8\exp(2\|\theta\|_1)/(\pi\varepsilon)$, *und* $k' = 1$ *wenn* $\|\theta\|_\infty < 1$ *oder andernfalls* $k' = k$. *Dann gilt*

$$\mathbb{P}[|\hat{Z}_\zeta^{N,k}(\theta) - Z(\theta)| < \varepsilon Z(\theta)] \geq 1 - \delta\ .$$

Eine besondere Einsicht dieses Resultats ist der Zusammenhang zwischen der Norm des Parametervektors $\|\theta\|_1$ und dem Approximationsfehler. Bisherige approximative Inferenzmethoden basieren auf Einschränkungen der Struktur. Unser Theorem zeigt jedoch, dass eine Einschränkung der Parameternorm ausreicht, um den erforderlichen Polynomgrad $k$ zu senken, was wiederum eine Senkung der Worst-Case Komplexität $\mathscr{O}(d^k)$ zur Folge hat. Tatsächlich kann die Norm mittels $l_1$-Regularisierung $\lambda_1\|\theta\|_1$ während des Lernens kontrolliert werden. Exemplarische Lernkurven sind in Abb. 5 dargestellt. Hierbei ist zu beachten, dass sowohl der Gradient als auch der Zielfunktionswert beim Lernen auf der stochastischen Quadratur basieren. Wir sehen, dass in diesem Fall eine kubische Interpolation bereits ausreicht, um eine sehr gute Näherungslösung zu erhalten.

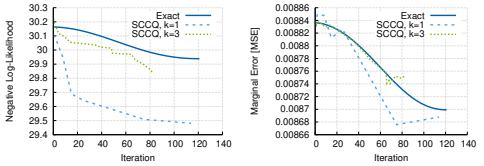



Abb. 5: Exemplarische Lernkurven der Stochastischen Clenshaw-Curtis Quadratur (SCCQ) für eine lineare und eine kubische Approximation.

## 5 Fazit

Unsere systematische Analyse der Ressourcennutzung von Exponentialfamilienmodellen brachte uns neue Erkenntnisse auf den Gebieten der „Model-Compression", dem Lernen unter nicht-konvexen Nebenbedingungen sowie der approximativen Inferenz unter Erhalt der bedingten Unabhängigkeitsstruktur. Alle unsere Ergebnisse verbindet die Regularisierung, die es uns erlaubt, die Menge der erlaubten Modelle während des Lernens einzuschränken. Durch eine theoretisch fundierte Einschränkung der erlaubten Parameter kann die Worst-Case Abweichung zur unrestringierten optimalen Lösung beschränkt werden. Desweiteren können Regularisierungsterme als log-a-priori Wahrscheinlichkeit interpretiert werden. Daher können unsere Ansätze auch als Bayesianische Schätzung interpretiert werden, wobei die Wahrscheinlichkeit, ein bestimmtes Modell zu lernen, proportional zu seinem Ressourcenverbrauch ist. Abschließend sind in Abb. 6 Laufzeitresultate für unsere ganzzahlige Inferenz (BLprop) sowie unsere stochastische Quadratur (SCCQ) auf einem MSP430 Mikrocontroller mit 16 MHz Takt dargestellt. LBP bezeichnet hier die gewöhnliche (Loopy)-Belief-Propagation. Hier können wir eine $>$ 200-fache Beschleunigung beobachten.

Abb. 6: Laufzeit (in ms) einer Iteration bzw. eines Samples der Algorithmen LBP, BLprop und SCCQ.

| **Daten** | $\lvert E\rvert$ | **LBP** | **BLprop** | **SCCQ** |
|---|---|---|---|---|
| Chain | 15 | 4843.4 | 19.0 | 773.8 |
| Star | 15 | 4744.3 | 19.0 | 774.0 |
| Grid | 24 | 7713.9 | 29.5 | 1081.3 |
| Full | 120 | 40422.6 | 141.2 | 4704.2 |

**Nico Piatkowski** studierte Informatik und Wirtschaftswissenschaften an der Technischen Universität Dortmund. Nach dem Abschluss seines Studiums forschte er dort am Lehrstuhl für Künstliche Intelligenz. Nico promovierte über Maschinelles Lernen unter Ressourcenbeschränkung im Rahmen des Sonderforschungsbereichs 876. Für seine Forschung erhielt er den Best-Student-Paper-Award der größten europäischen Konferenz für maschinelles Lernen (ECMLPKDD) sowie den Dissertationspreis der TU Dortmund. Nico ist Autor von über 30 begutachteten Fachartikeln und regelmäßig als Mitglied in Programmkomitees aller großen Machine Learning und Data Mining Konferenzen, sowie als Gutachter für Fachzeitschriften. Nico organisierte 2012 das jährliche Treffen der GI-Fachgruppe "Knowledge Discovery, Data Mining und Maschinelles Lernen" und half als "Proceedings Chair" bei der Organisation der ECMLPKDD 2013.

## Literaturverzeichnis


[CC60] Clenshaw, C. W.; Curtis, A. R.: A method for numerical integration on an automatic computer. Numerische Mathematik, 2(1):197–205, 1960.

[CL68] Chow, C.; Liu, C.: Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory, 14(3):462–467, May 1968.

[PB14] Parikh, Neal; Boyd, Stephen: Proximal Algorithms. Foundations and Trends in Optimization, 1(3):127–239, 2014.

[Pi36] Pitman, Edwin James George: Sufficient statistics and intrinsic accuracy. Mathematical Proceedings of the Cambridge Philosophical Society, 32:567–579, 1936.

[Pi18] Piatkowski, Nico: Exponential families on resource-constrained systems. Dissertation, Technical University of Dortmund, Germany, 2018.

[PLM13] Piatkowski, Nico; Lee, Sangkyun; Morik, Katharina: Spatio-Temporal Random Fields: Compressible Representation and Distributed Estimation. Machine Learning, 93(1):115–139, 2013.

[PLM16] Piatkowski, Nico; Lee, Sangkyun; Morik, Katharina: Integer undirected graphical models for resource-constrained systems. Neurocomputing, 173, Part 1:9–23, 2016.

[PM16] Piatkowski, Nico; Morik, Katharina: Stochastic Discrete Clenshaw-Curtis Quadrature. In: Proceedings of the ICML. Jgg. 48 in JMLR Workshop and Conference Proceedings. JMLR.org, S. 3000–3009, 2016.

[PM18] Piatkowski, Nico; Morik, Katharina: Fast Stochastic Quadrature for Approximate Maximum-Likelihood Estimation. In: Proceedings of the UAI. 2018.

[Va79] Valiant, Leslie Gabriel: The complexity of enumeration and reliability problems. SIAM Journal on Computing, 8(3):410–421, 1979.

# Robuste Sprachverbesserung durch statistische Signalverarbeitung und maschinelles Lernen[1]

Robert Rehr[2]

**Abstract:** In vielen Anwendungen, z. B. für Hörgeräte, zur Interaktion zwischen Mensch und Maschine, oder für Telekommunikation, spielen Sprachsignale eine besondere Rolle. In Umgebungen, in denen neben dem Sprachsignal auch weitere Quellen präsent sind, nehmen die Mikrofone nicht nur das gewünschte Sprachsignal sondern auch weitere Störgeräusche auf. Hiedurch reduziert sich die Qualität und potentiell auch die Verständlichkeit des aufgenommenen Sprachsignals. Zur Reduktion der Hintergrundgeräusche werden Sprachverbesserungsalgorithmen eingesetzt. Diese Arbeit konzentriert sich hierbei auf einkanalige Verfahren, bei denen Synergien zwischen Verfahren, die auf maschinellem Lernen basieren, und Verfahren der statistische Signalverarbeitung genutzt werden.

## 1 Einleitung

Sprache ist eine der natürlichsten Formen der Kommunikation für Menschen und wird zum Austausch von Informationen, Ideen und Gefühlen genutzt. Begünstigt durch die Verfügbarkeit von leistungsfähigen, elektronischen Geräten spielt Sprache eine immer wichtigere Rolle in vielen Anwendungen, z. B. in der mobilen Telekommunikation und in Hörhilfen. Neben der zwischenmenschlichen Kommunikation ist Sprache auch ein wichtiger Bestandteil für die Interaktion zwischen Mensch und Maschine, z. B. mit Robotern oder virtuellen persönlichen Assistenten.

Die für die Umsetzung solcher Anwendungen eingesetzten Geräte werden oft in Umgebungen verwendet, in denen neben dem Zielsignal auch weitere Geräusche auftreten. In solchen Situationen nehmen die Mikrofone nicht nur das gewünschte Sprachsignal sondern zusätzlich auch ungewünschte Signale auf. Dies verschlechtert die wahrgenommene Qualität des Sprachsignals und wirkt sich potentiell auch negativ auf die Verständlichkeit aus. Außerdem erhöht sich auch die Leistung automatischer Spracherkennungsalgorithmen. Um die Qualität und, wenn möglich, auch die Verständlichkeit der gestörten Sprache wiederherzustellen, werden Sprachverbesserungsalgorithmen eingesetzt.

In dieser Arbeit werden einkanalige Sprachverbesserungsalgorithmen betrachtet, die entweder das Signal eines einzelnen Mikrofons oder den Ausgang eines mehrkanaligen räumlichen Filters, d. h. eines Mikrofon-Arrays, verarbeiten. Viele solcher Verfahren transformieren zur Verbesserung das verrauschte Zeitsignal in eine Zeit-Frequenz Repräsentation, wobei häufig eine Kurzzeit Fourier-Transformation (short-time Fourier transform, STFT) verwendet wird. Zeitfrequenzpunkte, die hauptsächlich dem Geräusch zugeordnet werden,

[1] Robust Speech Enhancement Using Statistical Signal Processing and Maching Learning

[2] Universität Hamburg, robert.rehr@uni-hamburg.de

werden anschließend mit einer sogenannten Gewichtungsfunktion unterdrückt und auf Werte nahe Null gesetzt. Abschließend wird das Signal zurück in den Zeitbereich transformiert. Einkanalige Sprachverbesserungsalgorithmen lassen sich grob in zwei Kategorien einteilen:

**1. Verfahren basierend auf statistischer Modellierung** Bei solchen Ansätzen werden die Gewichtungsfunktionen, die auf die STFT-Koeffizienten angewendet werden, in einem statistischen Rahmenwerk hergeleitet. Hierzu werden die Koeffizienten der unverrauschten Sprache und des Rauschens durch parametrische Wahrscheinlichkeitsdichtefunktionen (probability density function, PDFs) modelliert. Die Parameter sind im Allgemeinen durch die Leistungsdichtespektren (power spectral density, PSDs) von Sprache und Rauschen gegeben, die blind aus den verrauschten Beobachtungen geschätzt werden. Hierzu wird häufig angenommen, dass sich das Geräusch über die Zeit weniger stark verändert als Sprache.

**2. Verfahren basierend auf maschinellem Lernen (ML)** Im Gegensatz zu dem konventionellen Ansatz nutzen ML-basierte Algorithmen repräsentative Beispiele, um die statistischen Eigenschaften der Sprache und des Rauschens zu lernen. Häufig sind ML-basierte Ansätze dadurch motiviert, dass konventionelle Ansätze nicht in der Lage sind, hochinstationären, d. h. sich schnell ändernden Geräuschtypen, zu folgen. Im Gegensatz dazu besteht bei ML-basierten Ansätzen allerdings die potentielle Gefahr, dass Eingangsdaten, die stark von den Trainingsdaten abweichen, nicht korrekt verarbeitet werden.

Das Ziel dieser Arbeit ist es, die Robustheit einkanaliger Verfahren zur Sprachverbesserung zu erhöhen. Um dieses Ziel zu erreichen, werden die beiden oben beschriebenen Ansätze betrachtet und Synergien zwischen beiden Ansätzen genutzt. Die sieben Forschungsbeiträge, auf denen die Arbeit [Re19] basiert, sind hierzu in drei Teile gegliedert und präsentieren Verbesserungen für konventionelle, nicht-ML-basierte Ansätze, ML-basierte Ansätze sowie Kombinationen aus beiden Verfahren. Die folgenden Abschnitte dieser Kurzzusammenfassung geben eine Übersicht über die drei Teile der Dissertation.

## 2 Geräusch-PSD Schätzer basierend auf adaptiver Glättung

Das Bestimmen der Geräusch-PSD ist äquivalent zur Bestimmung der Varianz der komplexwertigen spektralen Koeffizienten des Geräuschs $N_{k,\ell}$. Hierbei symbolisiert $N_{k,\ell}$ die STFT des Geräuschs, wobei $k$ der Frequenzindex ist und $\ell$ der Zeitindex. Die Fouriertransformation erlaubt es, die spektralen Koeffizienten durch eine mittelwertfreie Verteilung zu beschreiben. Aufgrund dessen kann die Varianz des Geräuschkoeffizienten $\Lambda_{k,\ell}^{(\mathrm{n})}$ eines spektralen Koeffizienten durch den Erwartungswert $\Lambda_{k,\ell}^{(\mathrm{n})} = \mathbb{E}\{|N_{k,\ell}|^2\}$ definiert werden. In praktischen Anwendungen wird die Berechnung des Erwartungswert häufig durch eine Mittelung der Betragsquadrate $|N_{k,\ell}|^2$ über der Zeit $\ell$ bestimmt. Die Mittelung findet dabei für jedes Frequenzband $k$ separat statt. In der Signalverarbeitung werden hierzu häufig rekursive Glättungsfilter erster Ordnung verwendet. Solche Filter ermöglichen es, Änderungen entlang der Zeit zu verfolgen und zusätzlich lässt sich zeigen, dass diese Filter erwartungstreue Schätzer sind. Im Kontext der Geräusch-PSD-Schätzung ließe sich ein

solches Filter durch

$$\hat{\Lambda}_{k,\ell}^{(\mathrm{n})} = (1-\alpha)|Y_{k,\ell}|^2 + \alpha\hat{\Lambda}_{k,\ell-1}^{(\mathrm{n})} \tag{1}$$

beschreiben, wobei $\hat{\Lambda}_{k,\ell}^{(\mathrm{n})}$ die Schätzung der Geräusch-PSD $\Lambda_{k,\ell}^{(\mathrm{n})}$ ist. Hierbei ist $0 < \alpha < 1$ die Glättungskonstante, welche die Stärke der Glättung kontrolliert. Zusätzlich ist $Y_{k,\ell}$ die STFT des verrauschten Sprachsignals, d. h. $Y_{k,\ell} = S_{k,\ell} + N_{k,\ell}$ und $S_{k,\ell}$ ist die STFT des unverrauschten Sprachsignals.

Aufgrund des zusätzlichen Sprachsignals würde eine einfache Anwendung des Filters in (1) allerdings dazu führen, dass auch Sprachanteile in die Schätzung $\hat{\Lambda}_{k,\ell}^{(\mathrm{n})}$ einfließen. Um dies zu verhindern, tauschen die Geräusch-PSD-Schätzer in [GH11] und [HS04, Kap. 14.1.3] den festen Parameter $\alpha$ gegen einen zeitlich veränderlichen aus. Hierzu wird in Abschnitten, in denen Sprache anwesend ist, der Glättungsparameter auf einen größeren Wert gesetzt, wodurch die Aktualisierung der Schätzung verlangsamt wird. In Abschnitten, in denen nur das Hintergrundgeräusch präsent ist, wird der Glättungsparameter auf einen kleineren Wert zurückgestellt. Die Wahl des Glättungsparameters hängt bei den Verfahren in [GH11, HS04] von dem Verhältnis $|Y_{k,\ell}|^2/\hat{\Lambda}_{k,\ell-1}^{(\mathrm{n})}$ ab. In der vorliegenden Arbeit konnten wir zeigen, dass die Einführung des zeitlich veränderlichen Glättungsparameters allerdings dazu führt, dass das rekursive Filter nicht länger ein erwartungstreuer Schätzer ist. Das bedeutet, dass die Größe entweder unterschätzt wird, was zu einer verringerten Geräuschreduktion führt, oder überschätzt wird, was zu einer Verzerrung des Sprachsignals bei der Verbesserung führt. In [Re19, Kap. 3 und Kap. 4] werden die Geräusch-PSD-Schätzer, welche auf einer solchen adaptiven rekursiven Glättung basieren, analysiert. Darüber hinaus werden in dieser Arbeit Methoden zur Bestimmung des Fehlers und zur Kompensation des Fehlers vorgestellt. Hierzu werden die in [GH11, HS04] vorgestellten Geräusch-PSD-Schätzer als Beispiele verwendet.

In [Re19, S. 45ff] wird gezeigt, dass die betrachteten Geräusch-PSD-Schätzer skalierungsinvariant sind. Wenn das Signal nur Geräusch enthält, lässt sich der systematische Fehler aufgrund dieser Eigenschaft durch die Multiplikation eines einzelnen Faktors $\mathscr{C}$ korrigieren. Der Faktor kann hierzu auf das Eingangs- oder das Ausgangssignal angewendet werden. In [Re19, Kap. 4] wird eine weitere Methode zur Korrektur des Schätzfehlers vorgestellt, bei der ebenfalls nur ein einzelner Faktor $\mathscr{C}^{(\mathrm{a})}$ zur Korrektur des Fehlers notwendig ist, wenn das Signal nur das Geräusch enthält. Im Gegensatz zu der in [Re19, Kap. 3] vorgestellten Methode findet die Korrektur an einer anderen Stelle der rekursiven Struktur statt, sodass im Allgemeinen $\mathscr{C} \neq \mathscr{C}^{(\mathrm{a})}$ gilt.

Aufgrund der Rekursion und der Nichtlinearität der adaptiven Glättungsfilter, ist die analytische Bestimmung von $\mathscr{C}$ als auch $\mathscr{C}^{(\mathrm{a})}$ eine besondere Herausforderung. In [Re19, Kap. 3 und 4] werden analytische Lösungen vorgestellt, mit denen beide Korrekturfaktoren näherungsweise bestimmt werden können. Beide Ansätze verfolgen hierzu das Ziel den Erwartungswert des Geräuschschätzers, d. h. $\mathbb{E}\{\hat{\Lambda}_{k,\ell}^{(\mathrm{n})}\}$, zu bestimmen. Hieraus lässt sich die Abweichung von der wahren Geräusch-PSD bestimmen, woraus sich der benötigte Korrekturfaktor ableiten lässt. Bei der Anwendung der festen Korrekturfaktoren $\mathscr{C}$ und $\mathscr{C}^{(\mathrm{a})}$ wird nicht berücksichtigt, dass das verrauschte Eingangssignal auch Sprache enthält. Daher

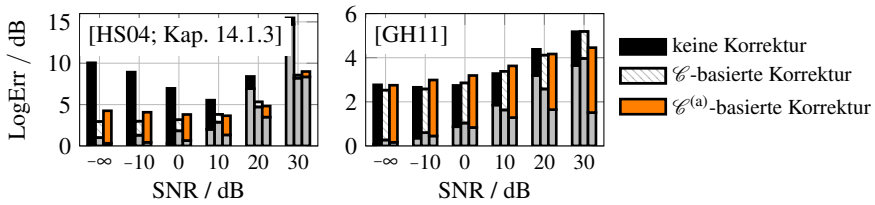



Abb. 1: Überschätzung (unterer, grauer Teil) und Unterschätzung (oberer, farbiger Teil) der Geräusch-PSD für die in [HS04, Kap. 14] und [GH11] beschriebenen Schätzer mit und ohne die in [Re19, Kap. 3 und Kap. 4] vorgestellte Korrektur im Cafeteria-Geräusch.

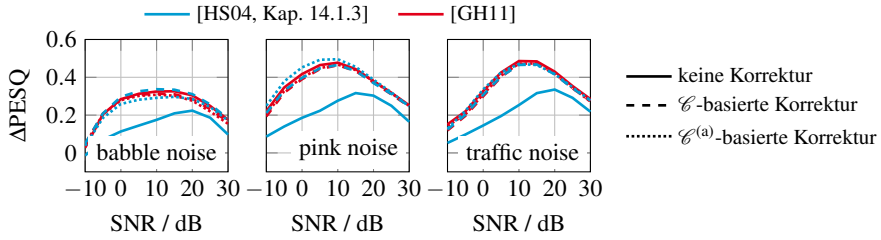



Abb. 2: Vergleich der in [Re19, Kap. 3 und 4] vorgestellten Korrekturmethoden für die Geräusch-PSD-Schätzer in [GH11] und [HS04, Kap. 14.1.3] in Bezug auf PESQ-Verbesserungen.

werden in in [Re19, Kap. 3.5] und [Re19, Kap. 4.2] Erweiterungen vorgeschlagen, um das Sprachsignal bei der Korrektur mit einzubeziehen.

Abbildung 1 zeigt die logarithmische Abweichung für beide in [Re19, Kap. 3 und Kap. 4] untersuchten Geräusch-PSD-Schätzer in einem Cafeteria-Hintergrundgeräusch mit und ohne die vorgeschlagenen Korrekturansätze. Das Fehlermaß erlaubt es die Über- und Unterschätzung der Geräusch-PSD zu quantifizieren. Bei dem in [GH11] vorgestellten Geräusch-PSD-Schätzer ist der Fehler im allgemeinen relativ gering. Aufgrund dessen haben die vorgeschlagenen Korrekturmethoden hier nur eine geringe Auswirkung. Für das in [HS04, Kap. 14.1.3] vorgestellte Verfahren lässt sich hingegen erkennen, dass bei hohen Eingangs-Signal-zu-Rausch Verhältniss (signal-to-noise ratio, SNRs), d. h. wenn das Sprachsignal lauter ist als das Geräusch, die Geräusch-PSD ohne Korrektur überschätzt wird. Im Gegensatz dazu wird die Geräusch-PSD bei niedrigen SNRs, also wenn Sprache deutlich leiser ist als das Geräusch, unterschätzt. Beide Fehlschätzungen lassen sich mit den vorgestellten Korrekturmethoden effektiv verringern.

Zusätzlich wurden die vorgeschlagenen Korrekturverfahren mit Perceptual Evaluation of Speech Quality (PESQ) [P.01] untersucht, einem instrumentellen Maß, dass die Sprachqualität der verbesserten Sprachsignale algorithmisch vorhersagt. Abbildung 2 zeigt die Verbesserung von PESQ gegenüber dem verrauschten Signal, wobei höhere Werte eine bessere Qualität widerspiegeln. Ähnlich zur in Abbildung 1 dargestellten logarithmischen Abweichung hat die Korrektur nur einen geringen Einfluss auf den in [GH11] vorgestellten Geräusch-PSD-Schätzer, während sich für das Verfahren in [HS04, Kap. 14.1.3] ein klarer positiver Effekt beobachten lässt. Die $\mathscr{C}$-basierte und die $\mathscr{C}^{(a)}$-basierte Korrekturmethode liefern ähnliche Verbesserungen.

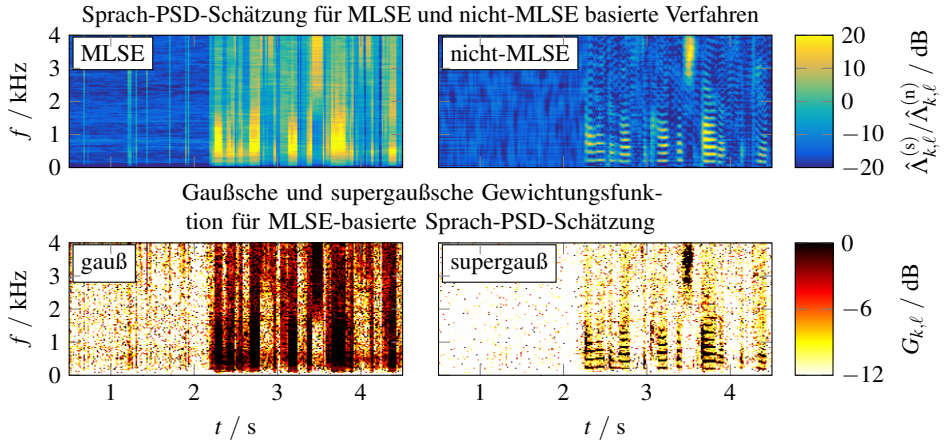



Abb. 3: Oben: Verhältnis der Sprach-PSD und der Rausch-PSD eines nicht-MLSEs und eines MLSEs. Unten: Gewichtungsfunktion eines gaußschen und eines supergaußschen Sprachschätzers für den Fall, dass die MLSE-basierte Sprach-PSD Schätzung verwendet wird. Gleiches Sprachsignal in allen vier Fällen.

## 3 ML-basierte Spracheinhüllendeverfahren

Im zweiten Abschnitt der Arbeit [Re19, Kap. 5 – 7], werden ML-basierte Sprachverbesserungsansätze adressiert, die typische Strukturen der Sprache erlernen. Für die Verarbeitung wird anschließend das Modell selektiert, das die Beobachtung am besten erklärt. Das Besondere an den in diesem Teil betrachteten Methoden ist, dass die betrachteten ML-basierten Ansätze nur eine grobe spektrale Einhüllende der Sprache repräsentieren. Diese Art von Verfahren werden im folgenden als ML-basierte Spracheinhüllendenverfahren (machine-learning spectral envelope, MLSE) bezeichnet und werden unter anderem zur Sprachverbesserung eingesetzt [CGG16]. Die Form der spektralen Einhüllende von Sprache resultiert aus den Resonanzen, die durch den Vokaltrakt des Menschens erzeugt werden, und ermöglicht es verschiedene Phoneme, z. B. die Vokale, akustisch voneinander zu unterscheiden. Allerdings enthält die spektrale Einhüllende nicht die spektrale Feinstruktur, die dem Grundton und deren Harmonischen entspricht, die durch das Schwingen der Stimmlippen in stimmhaften Lauten erzeugt werden. MLSE haben zum einen den Vorteil, dass sie sehr gut generalisieren und zu recheneffizienten Lösungen führen. Allerdings überschätzen MLSE-Ansätze die Sprach-PSD zwischen den spektralen Harmonischen der Sprache. Dadurch wird das Geräusch zwischen diesen Harmonischen typischerweise nicht unterdrückt. Infolgedessen ist die Geräuschreduktion in sprachaktiven Segmenten begrenzt und führt zu hörbaren Aktivierungen des Geräuschs in solchen Abschnitten. Der obere Teil von Abbildung 3 zeigt das Verhältnis der geschätzten Sprach-PSD $\hat{\Lambda}_{k,\ell}^{(\mathrm{s})}$ und der Geräusch-PSD $\hat{\Lambda}_{k,\ell}^{(\mathrm{n})}$ für ein MLSE- und ein nicht-MLSE-Verfahren. Das Hintergrundgeräusch ist ein stationäres rosa Rauschen, das mit einem SNR von 5 dB dem Sprachsignal hinzugefügt wurde. Während sich für das nicht-MLSE-Verfahren klar die harmonischen Strukturen von Sprache erkennen lassen, gehen diese bei dem hier betrachteten MLSE-Verfahren verloren.

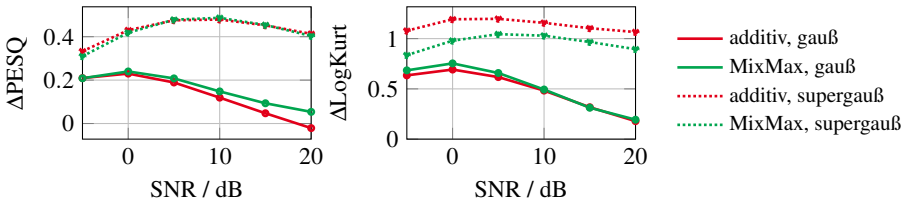



Abb. 4: Vergleich der Sprachqualität und der Artefakte von gaußschen und supergaußschen Sprachschätzern, die jeweils unter dem additiven Modell und dem MixMax-Modell hergeleitet wurden, bei MLSE-Verfahren.

In [Re19] zeigen wir, dass Sprachschätzer, die auf einer supergaußschen Annahme zur Modellierung der Sprachkoeffizienten basieren, das Problem der Unterdrückung des Störgeräuschs ohne heuristische Nachverarbeitung lösen. Die Untersuchungen zeigen, dass Beobachtungen $|Y_{k,\ell}|^2$, die einen ähnlichen Wert haben wie die geschätzte Geräusch-PSD $\hat{\Lambda}_{k,\ell}^{(\mathrm{n})}$, mit einem supergaußschen Schätzer stärker unterdrückt werden. Dies gilt insbesondere auch für den Fall, wenn die Sprach-PSD $\Lambda_{k,\ell}^{(\mathrm{s})}$ deutlich größer als die Geräusch-PSD $\Lambda^{(\mathrm{n})}$ ist, was bei einer Überschätzung der Fall ist. Der untere Teil von Abbildung 3 stellt die Gewichtungsfunktion $G_{k,\ell}$ eines gaußschen und eines supergaußschen Sprachschätzer, wenn eine MLSE-basierte Sprach-PSD-Schätzung verwendet wird. Das Ergebnis zeigt, dass ein Schätzer basierend auf der supergaußschen Annahme die Feinstruktur des Sprachsignals wiederherstellen kann, was mit einem gaußschen Schätzer hingegen nicht möglich ist.

In [Re19] untersuchen wir neben dem additiven Signalmodell auch Sprachschätzer, die unter dem MixMax-Modell hergeleitet wurden. Das MixMax-Modell arbeitet im log-spektralen Bereich, der durch $\log(|Y_{k,\ell}|^2)$ definiert ist, und modelliert das verrauschte Log-Spektrum durch $\log(|Y_{k,\ell}|^2) = \max(\log(|S_{k,\ell}|^2), \log(|N_{k,\ell}|^2))$. Log-spektrale Repräsentationen sind gut geeignet, um generalisierende Sprachmodelle zu trainieren und werden daher häufig auch zur Spracherkennung eingesetzt. Durch die nichtlineare Transformation ist allerdings der sonst üblicherweise genutzte additive Zusammenhang zwischen Sprache und Rauschen mathematisch nicht länger handhabbar. Hierbei wird das MixMax-Modell zur mathematischen Vereinfachung genutzt [CGG16]. In [GM09] wird gezeigt, dass Spektralkoeffizienten, die einer supergaußschen Verteilung folgen, eine höhere Varianz im Log-Spektralbereich aufweisen als gaußverteile Spektralkoeffizienten. Mit diesem Zusammenhang wird gezeigt [Re19, Kap. 6], dass sich der Vorteil von supergaußschen Schätzverfahren auch für MixMax-basierte Sprachschätzer übertragen lässt. Die Verwendung des MixMax-Modells hat dabei den Vorteil, dass weniger wahrnehmbare Prozessierungsartefakte auftreten. Dieses Ergebnis wurde instrumentell mit dem Log-Kurtosis Verhältnis verifiziert, das in Abbildung 4 neben dem Sprachqualitätsmaß PESQ dargestellt ist. Höhere Werte des Log-Kurtosis-Verhältnis entsprechen dabei einem häufigeren Auftreten von Prozessierungsartefakten. Das supergaußsche Modell unter dem additiven und dem MixMax-Modell liefern ähnliche Ergebnisse im Bezug auf die Sprachqualität, die durch PESQ vorhergesagt wird. Allerdings ist Log-Kurtosis-Verhältnis für den MixMax-basierten sußergaußschen Schätzer deutlich kleiner, was auf weniger Artefakte hindeutet. Zusätzlich wurde die Effektivität von supergaußschen Schätzern in einem Hörversuch verifiziert [Re19, Kap. 5.6].

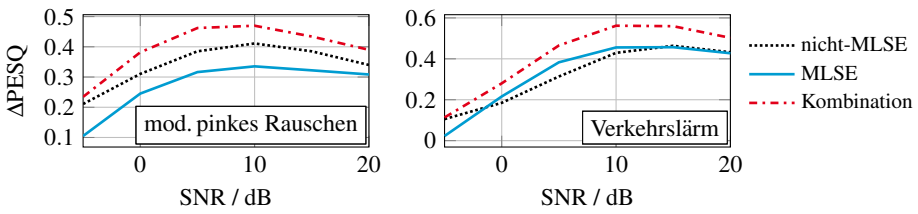



Abb. 5: Verbesserungen in PESQ für die statistische Kombination von konventionellen Sprachverbesserungsalgorithmen und MLSE-Verfahren.

In [Re19, Kap. 7] wird ein Ansatz verfolgt, bei dem mit Hilfe einer Kombination von MLSE-Verfahren und konventionellen Verfahren die Feinstruktur wiederhergestellt wird. Die Kombination der beiden Verfahren wird mit Hilfe eines statistischen Rahmenwerks realisiert. Das Resultat der Kombination ist, dass die konventionellen Algorithmen das Geräusch zwischen spektralen Harmonischen unterdrücken, während das MLSE-Verfahren hauptsächlich die Unterdrückung von Koeffizienten verhindert, die von Sprache dominiert sind. Dementsprechend wird bei einer Geräuschreduktion, die vergleichbar mit konventionellen Ansätzen ist, weniger Sprache verzerrt. Abbildung 5 zeigt die Verbesserungen der vorgeschlagenen Kombination von konventionellen und MLSE-Ansätzen gegenüber reinen nicht-MLSE- und MLSE-Ansätzen in Form von PESQ-Verbesserungen.

## 4 Generaliserung bei DNN-basierter Sprachverbesserung

Im dritten Teil der Arbeit [Re19, Kap. 8] wird die Generalisierung von ungesehenen akustische Umgebungen im Zusammenhang mit ML-basierten Verbesserungsverfahren, die auf tiefen neuronalen Netzwerken (deep neural networks, DNNs) basieren, untersucht. Bei DNNs handelt es sich um eine Methode des MLs, die es ermöglicht beliebige Funktionen auf einem beschränkten Raum zu approximieren. Solche Verfahren wurden kürzlich mit vielversprechenden Ergebnissen für die Sprachverbesserung eingesetzt [Xu15]. Ein übliches Problem, dass für viele Verfahren des MLs gilt, trifft auch auf DNNs zu und zwar, wie gut das gelernte Modell ungesehene Eingangsdaten generalisieren kann. Diese Fragestellung ist auch für DNN-basierte Sprachverbesserungssysteme untersucht worden [Xu14, KTJ17], wobei in [Xu14] das geräuschbasierte Training (noise-aware training, NAT) vorgeschlagen wurde, um eine erhöhte Robustheit von DNN-basierter Sprachverbesserung zu erzielen. Hierbei wird eine Schätzung der Geräusch-PSD an die Merkmale, die aus dem verrauschten Eingangssignal extrahiert werden, angehängt. Die Geräusch-PSD kann mit konventionellen Verfahren, z. B. [GH11], erfolgen, die im Allgemeinen den Vorteil haben, unabhängig von der akustischen Umgebung zu sein.

In [Re19, Kap. 8] stellen wir eine neuartige Methode zur Erhöhung der Robustheit von DNN-basierten Sprachverbesserungsalgorithmen vor. Hierzu werden Schätzungen der Sprach- und der Geräusch-PSD als Eingangsmerkmale für ein DNN-basiertes Sprachverbesserungsverfahren verwendet, die aufgrund der Robustheit mit einem konventionellen bestimmt werden. Im Gegensatz zum NAT werden aber SNRs als Eingangsmerkmale verwendet. Zum einen wird das *a priori* SNR, d. h. das Verhältnis zwischen Sprach- und

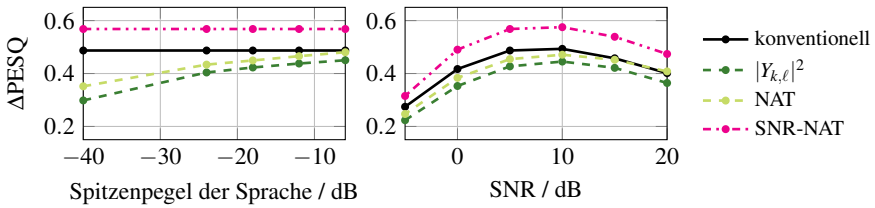



Abb. 6: Vergleich der verschiedener Eingangsmerkmale für DNN-basierte Sprachverbesserungsverfahren mittels PESQ. Zum Vergleich sind auch die Ergebnisse eines konventionellen Ansatzes dargestellt.

Geräusch-PSD $\Lambda_{k,\ell}^{(\mathrm{s})}/\Lambda_{k,\ell}^{(\mathrm{n})}$, vorgeschlagen. Zum anderen wird auch das *a posteriori* SNR, d. h. das Verhältnis der verrauschten Eingangsperiodogramm $|Y_{k,\ell}|^2$ und der Geräusch-PSD $\Lambda_{k,\ell}^{(\mathrm{n})}$, genutzt. Die beiden SNRs können separat verwendet werden oder durch Aneinanderhängen kombiniert werden. Letzteres wird als SNR-basiertes NAT (SNR-based NAT, SNR-NAT) bezeichnet wird.

Die NAT- und SNR-NAT-Merkmale werden in [Re19] experimentell mit der Methode der Kreuzvalidierung verglichen. Hierzu wird eine Menge von neun verschiedene Geräuschtypen, mit denen Sprachsignale künstlich verrauscht werden, betrachtet. Jedes Sprachsignal wird mit allen verfügbaren Störsignalen verrauscht und für jedes Sprachsignal wird das SNR als auch der Gesamtpegel zufällig anders gewählt, um das DNN lernen zu lassen, auf solche Variationen richtig zu reagieren. Für jedes der untersuchten Eingangsmerkmale, werden neun verschiedene Modelle trainiert, wobei alle verfügbaren Geräusche bis auf eins für das Training verwendet werden. Bei jedem dieser neun Modelle gibt es daher ein Geräusch, das nicht während des Trainings gesehen wurde. Außerdem wurden zusätzliche Geräuschtypen in dem Trainingsset aller Modelle eingefügt, um die Robustheit des DNN-basierten Verfahrens allgemein zu stärken. Die Größe des Trainingssatzes für jedes Modell entspricht dabei ungefähr 20 Stunden Audiomaterial, die genutzt werden, um ein Feedforward-Netzwerk mit drei versteckten Ebenen zu trainieren.

Die in Abbildung 6 dargestellten Ergebnissen zeigen den Mittelwert über alle ausgewerteten Geräuschtypen, d. h. ohne die für das Training zusätzlich eingefügten Geräuschtypen. Die Ergebnisse zeigen, dass das vorgeschlagene SNR-NAT gegenüber NAT zwei wesentliche Vorteile hat: Im linken Teil von Abbildung 6 sind die PESQ-Verbesserung für verschiedene Spitzenpegel der Sprache, die im direkten Zusammenhang mit dem Gesamtpegel des Signals stehen, dargestellt. Das Resultat zeigt, dass das DNN-basierte Verfahren mit der Verwendung des vorgeschlagenen SNR-NAT über den gesamten Pegelbereich nahezu identische Ergebnisse liefert, während die Qualität von NAT abhängig vom Gesamtpegel ist. Das bedeutet, dass das vorgeschlagenen SNR-NAT nicht durch den Gesamtpegel des Eingangssignals beeinflusst wird. Zweitens führt das vorgeschlagene SNR-NAT zu robusteren Modellen als NAT, wie der rechte Teil in Abbildung 6 zeigt. Dies lässt sich an den PESQ-Werten für das SNR-NAT-Verfahren erkennen, die über einen weiten SNR-Bereich über den verglichenen Verfahren liegen. Dies gilt insbesondere für den Fall, wenn wie im durchgeführten Experiment wenige Trainingsdaten zur Verfügung stehen.

Die Ergebnisse sind mit Hilfe eines Hörversuchs mit elf Teilnehmern verifiziert worden, bei dem die Qualität der verbesserten Signale bewertet wurde. Hierfür wurden zwei Hintergrundgeräusche untersucht und zwar, Fabriklärm und Verkehrslärm, die beide für die Verbesserungsverfahren nicht während des Trainings zur Verfügung standen. Die Resultate des Hörversuchs zeigen, dass die vorgeschlagenen SNR-NAT Merkmale signifikant besser bewertet werden als NAT, wenn das Geräusch nicht aus dem Training bekannt ist.

# 5 Zusammenfassung

Die Dissertation betrachtet verschiedene Aspekte der einkanaligen Sprachverbesserung und nutzt Synergien zwischen konventionellen Sprachverbesserungsalgorithmen und ML-basierten Verfahren aus. In diesem Rahmen sind verschiedene Beiträge zur Verbesserung konventioneller und ML-basierter Verfahren entstanden.

Zum einen wurden Verbesserungen für konventionelle Verfahren zur Schätzung der Geräusch-PSD vorgeschlagen, um systematische Schätzfehler zu vermeiden. Hierzu wurden verschiedene analytische Lösungen vorgestellt, die es ermöglichen den Schätzfehler approximativ zu bestimmen. Zusätzlich sind Methoden vorgestellt worden, mit denen der systematische Schätzfehler kompensiert werden kann. Die Evaluation zeigt, dass insbesondere Geräusch-PSD-Schätzer, die einen großen Schätzfehler aufweisen, von der Korrektur des Fehlers profitieren.

Zusätzlich wurden supergaußsche Sprachschätzer zur Verbesserung von MLSE-basierten Sprachverbesserungsalgorithmen vorgeschlagen. MLSE-basierte Verfahren verwenden trainierte Sprachmodelle, bei denen aber nur die spektrale Einhüllende des Sprachsignals abgebildet wird. Im Gegensatz zu Methoden, bei denen eine heuristische Nachverarbeitung eingesetzt wird, um die spektrale Feinstruktur wiederherzustellen, ermöglichen es supergaußsche Schätzer die Feinstruktur ohne diesen Umweg zu rekonstruieren. Die Effektivität dieser Methode ist durch die Evaluationen mit instrumentellen Maßen als auch durch Hörversuche verifiziert worden. Neben dem Einsatz von supergaußschen Schätzern ist außerdem eine Kombination von konventionellen Methoden und MLSE-Methoden vorgeschlagen worden, die in einem statistischen Rahmenwerk eingebettet wurde. Die Effektivität der Methode konnte auch hier mittels instrumenteller Maße verifiziert werden.

Zuletzt wurde die Generalisierung DNN-basierter Sprachverbesserungsverfahren betrachtet. Hierzu wurden SNR-basierte Merkmale vorgestellt, wobei für die Schätzung der Sprach- und Geräusch-PSD konventionelle Verfahren eingesetzt werden. In den Auswertungen konnte gezeigt werden, dass das vorgeschlagene SNR-NAT zwei Vorteile gegenüber dem zuvor vorgestellten NAT hat. Zum einen sind die Merkmale skalierungsinvariant, sodass die Leistung des Sprachverbesserungsalgorithmus nicht länger vom Gesamtpegel des Eingangssignals abhängen. Des Weiteren zeigen die Evaluationen mit instrumentellen Maßen und Hörversuche, dass diese Merkmale außerdem zu robusteren Modellen im Bezug auf den Einsatz in ungesehenen akustischen Umgebung führen.

## Literaturverzeichnis

[CGG16] Chazan, S. E.; Goldberger, J.; Gannot, S.: A Hybrid Approach for Speech Enhancement Using MoG Model and Neural Network Phoneme Classifier. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(12):2516–2530, Dezember 2016.

[GH11] Gerkmann, Timo; Hendriks, Richard. C.: Noise Power Estimation Based on the Probability of Speech Presence. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). New Paltz, NY, USA, S. 145–148, 2011.

[GM09] Gerkmann, Timo; Martin, Richard: On the Statistics of Spectral Amplitudes after Variance Reduction by Temporal Cepstrum Smoothing and Cepstral Nulling. IEEE Transactions on Signal Processing, 57(11):4165–4174, 2009.

[HS04] Hänsler, Eberhard; Schmidt, Gerhard: Acoustic Echo and Noise Control: A Practical Approach. Adaptive and Learning Systems for Signal Processing, Communication and Control. Wiley & Sons, 2004.

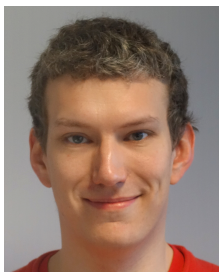
[KTJ17] Kolbæk, M.; Tan, Z. H.; Jensen, J.: Speech Intelligibility Potential of General and Specialized Deep Neural Network Based Speech Enhancement Systems. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(1):149–163, Januar 2017.

[P.01] : P.862: Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs. ITU-T recommendation, International Telecommunication Union, Januar 2001.

[Re19] Rehr, Robert: Robust Speech Enhancement Using Statistical Signal Processing and Machine-Learning. Dissertation, Universität Hamburg, Hamburg, Januar 2019.

[Xu14] Xu, Yong; Du, Jun; Dai, Li-Rong; Lee, Chin-Hui: Dynamic Noise Aware Training for Speech Enhancement Based on Deep Neural Networks. In: Interspeech. Singapore, Singapore, S. 2670–2674, September 2014.

[Xu15] Xu, Y.; Du, J.; Dai, L. R.; Lee, C. H.: A Regression Approach to Speech Enhancement Based on Deep Neural Networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(1):7–19, Januar 2015.



**Robert Rehr** hat Hörtechnik und Audiologie studiert und seinen B.Eng. an der Jade Hochschule in Oldenburg 2011 abgeschlossen und seinen M.Sc. an der Carl-von-Ossietzky Universität in Oldenburg 2013 erhalten. Von 2013 bis 2016 hat er in der Speech Signal Processing Group der Carl-von-Ossietzky Universität, Oldenburg seine Doktorarbeit begonnen, wo er von Sep. 2014 bis Feb. 2015 an einem Projekt mit „Sivantos – the hearing company" zusammengearbeitet hat. Er hat seine Doktorarbeit von 2017 bis. 2018 in der Signal Processing Group der Universität Hamburg abgeschlossen und war dort bis 2019 als wissenschaftlicher Mitarbeiter tätig. Er arbeitet seit Mai 2019 bei Oticon an Geräuschreduktionsalgorithmen für Hörgeräte.

Steffen Hölldobler et. al. (Hrsg.): Ausgezeichnete Informatikdissertationen 2018,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2019 201

# Sichere Ausführung von LLVM-basierten Sprachen auf der Java Virtual Machine[1]

Manuel Rigger[2]

**Abstract:** Sprachen wie C/C++ sind unsicher, da gewisse ihrer Operationen zu *undefiniertem Verhalten* führen können. Undefiniertes Verhalten ist problematisch, da es oft von AngreiferInnen ausgenutzt wird und zu schwer auffindbaren Fehlern führen kann. Um dieses Problem zu lösen haben wir Safe Sulong entwickelt, ein Werkzeug, das es erlaubt, unsichere Sprachen in einem sicheren Modus auszuführen. Die Grundidee dabei ist, die unsichere Sprache in einem Interpreter auszuführen, der in einer sicheren Sprache geschrieben ist, und damit automatisch undefiniertes Verhalten zu eliminieren. Die Evaluierung von Safe Sulong zeigt, dass es Fehler findet, die von anderen Werkzeugen übersehen werden, während Sulongs Ausführungsgeschwindigkeit vergleichbar mit der anderer Werkzeuge ist. Um die Implementierung von Inline-Assembly und Compiler-Builtins in Safe Sulong zu unterstützen, haben wir ferner empirische Studien durchgeführt, in denen wir diese Elemente untersuchten. Außerdem haben wir Introspektionsfunktionen entwickelt, mit denen man Metadaten von Fehlerfindungstools abfragen kann.

## 1 Problemkontext

In unsicheren Sprachen wie C ist die Semantik von Operationen nur für gültige Eingabewerte spezifiziert. Während z.B. das Dereferenzieren eines gültigen Zeigers spezifiziert ist, ist das Dereferenzieren eines Zeigers, der außerhalb eines Objektes zeigt (was als *Pufferüberlauf* bezeichnet wird), nicht spezifiziert. Solch ein Fehler verursacht *undefiniertes Verhalten*. Compiler sind nicht verpflichtet, Maschinencode zu produzieren der undefiniertes Verhalten erkennt, daher fügen sie z.B. keine Überprüfungen ein, die Pufferüberläufe verhindern. In der Tat optimieren weit verbreitete Compiler wie Clang or GCC Programme unter der Annahme, dass undefiniertes Verhalten nie auftritt [Wa12].

Das Ausführen von Programmen mit undefiniertem Verhalten, die mit Clang or GCC kompiliert wurden, kann zu verschiedenen unbeabsichtigten oder sogar katastrophalen Ergebnissen führen. Zum Beispiel kann das Lesen außerhalb von Objekten sensitive Daten preisgeben, die zu anderen Objekten gehören, auch wenn diese nicht explizit referenziert wurden. Aktuelle Beispiele für solche Fehler sind *Heartbleed* in der OpenSSL Bibliothek und *Cloudbleed* im Cloudfare Online-Service, die es beide ermöglichten sensitive Daten auf Webservern zu entwenden, was potenziell Millionen von Nutzern betraf. Pufferüberläufe können aber nicht nur dazu ausgenutzt werden, um private Daten zu lesen; Schreibzugriffe außerhalb von Objekten können benutzt werden um Kontrollflussdaten (z.B. Funktionsadressen) zu überschreiben, was AngreiferInnen ausnutzen können, um die Kontrolle über den Prozess zu erlangen.

Undefiniertes Verhalten ist auch dann problematisch, wenn es nicht von AngreifernInnen ausgenutzt wird. Programmfragmente, die undefiniertes Verhalten auslösen, können zu Maschinencode kompiliert werden, der sich bei der Ausführung anders verhält als erwartet. Solche Fragmente können auch komplett vom Compiler entfernt (d.h. wegoptimiert) werden, was zu „fehlerhaft kompiliertem" Code führt, der schwer zu debuggen ist. Manchmal ist undefiniertes Verhalten nicht unmittelbar ein Problem, z.B. wenn eine Operation so kompiliert wird, dass sie zufällig das erwartete Verhalten aufweist. Zum Beispiel kann eine Ganzzahladdition zu einer x86-`add`-Instruktion kompiliert werden,

[1] Englischer Titel der Dissertation: „Safe Execution of LLVM-based Languages on the Java Virtual Machine"

[2] Johannes Kepler Universität Linz, `manuel.rigger@jku.at`

die bei einem arithmetischen Überlauf das Vorzeichen wechselt). Solche Fehler sind jedoch „tickende Zeitbomben“, da zukünftige Compiler undefiniertes Verhalten ausnutzen könnten, um stärker zu optimieren.

## 2 Stand der Technik

Sowohl die Industrie als auch die Wissenschaft haben undefiniertes Verhalten, insbesondere Pufferüberläufe, seit Jahrzehnten behandelt. Daher gibt es eine Fülle von Ansätzen, die es auf unterschiedliche Weise angehen. Ein wichtiger Ansatz sind dynamische Fehlersuchwerkzeuge (sogenannte *Sanitizer*), die bestimmte Klassen von undefiniertem Verhalten erkennen, indem sie das Programm instrumentieren und die Ausführung bei einem Fehler abbrechen. Diese Werkzeuge erkennen Fehler während der Ausführung des Programms und müssen daher mit konkreten Programmeingaben (z. B. Programmargumenten oder Kommandozeileneingaben) versorgt werden. Sie erkennen Fehler in den Programmen ohne Fehlalarme, d.h. Fehler, die von solchen Werkzeugen angezeigt werden, sind immer echte Fehler.

Sanitizer können entweder mittels dynamischer Binärcode-Instrumentierung oder mittels Instrumentierung zur Compilezeit entwickelt werden. Dynamische Binärcode-Instrumentierung fügt Prüfungen zum Maschinencode hinzu, nachdem das Programm gestartet wurde. Ein Beispiel dafür ist Valgrind [NS07], das z.B. Speicherfehler im Programm erkennt. Ein großer Vorteil dieses Ansatzes ist, dass Werkzeuge keinen Zugriff auf den Quellcode benötigen, da sie direkt auf dem Binärcode arbeiten. Ein Nachteil ist jedoch, dass beim Übersetzungsvorgang Informationen verloren gehen, so dass nur bestimmte Fälle von undefiniertem Verhalten erkannt werden. Da es unser Ziel ist, möglichst viele Fehler zu finden, konzentrieren wir uns auf die Instrumentierung zur Compilezeit, die Prüfungen in den Quellcode oder in die Zwischensprache des Compilers einfügen und deshalb typischerweise alle Fehler unterschiedlicher Fehlerkategorien während der Ausführung erkennen können. Ein Beispiel für diesen Ansatz ist LLVM's AddressSanitizer (ASan), der Pufferüberläufe und andere Fehler erkennt [Se12].

## 3 Herausforderungen

Unser Hauptziel war es, vorhandene Sanitizer zu verbessern und Programme sicher auszuführen, auch wenn sie undefiniertes Verhalten verursachen. *Sicher* bedeutet in diesem Kontext, dass Operationen mit undefiniertem Verhalten eine definierte Semantik zugeordnet wird; z.B. soll eine undefinierte Operationen zum kontrollierten Beenden des Programms mit einer Fehlermeldung führen. Wir haben drei Bereiche identifiziert, in denen aktuelle Ansätze Schwächen aufweisen, die wir in Angriff nehmen wollten.

**Unsichere Optimierungen.** Wenn Compileroptimierungen aktiviert sind, können Sanitizer nicht alle Fehler erkennen, da die Compiler undefiniertes Verhalten für Optimierungen nutzen; wenn Optimierungen jedoch deaktiviert sind, ist das kompilierte Programm langsam. Außerdem ist das Hinzufügen von Laufzeitprüfungen sowohl im Compiler als auch im Binärcode fehleranfällig, da die Instrumentierung für Sonderfälle vergessen werden kann, was die Fehlererkennung beeinträchtigt. Wir glauben daher, dass ein neuartiger Ansatz erforderlich ist, der umfassenden Schutz vor undefiniertem Verhalten bietet, indem undefiniertes Verhalten nicht wegoptimiert werden kann, das entstehende Programm aber dennoch schnell genug ist, um in der Praxis eingesetzt zu werden.

**Fehlendes Wissen über Inline-Assembly und Compiler-Builtins.** Werkzeuge für unsichere Sprachen wie C erfordern einen hohen Implementierungsaufwand. C-Programme enthalten beispielsweise unstandardisierte Elemente wie Inline-Assembly und Compiler-Builtins. Nach bestem Wissen und Gewissen erkennen aktuelle Sanitizer undefiniertes Verhalten in ihnen nicht. Eine vollständige Implementierung solcher Elemente würde die Unterstützung verschiedener

Maschinenarchitekturen mit Hunderten oder Tausenden verschiedener Instruktionen oder Builtins erfordern. Es wäre hilfreich, ihre Verwendung in der Praxis zu untersuchen, da Werkzeugentwickler dadurch die Implementierung von Fehlerprüfungen für solche Elemente priorisieren können.

**Fehlende Programmierschnittstellen.** Sanitizer zeichnen Metadaten auf (z.B. Objektgrößen) und verwalten diese, um Fehlerprüfungen zu implementieren. Sie stellen diese Informationen jedoch nicht ProgrammiererInnen zur Verfügung, die diese Information beispielsweise verwenden könnten, um Invarianten im Programm zu überprüfen. Wir glauben, dass es hilfreich wäre, solche Metadaten über eine Programmierschnittstelle zur Verfügung zu stellen, die von verschiedenen Werkzeugen implementiert werden könnte. Das würde ProgrammierInnen helfen, undefiniertes Verhalten manuell zu verhindern.

## 4 Wissenschaftliche und Technische Beiträge

Die vorliegende Arbeit weist drei Beiträge zur Lösung der oben beschriebenen Probleme auf: *Safe Sulong*, *Introspektion* und *Empirische Studien*.

**Safe Sulong.** Wir stellen einen neuartigen Ansatz vor, um potenziell unsichere LLVM-basierte Sprachen sicher auszuführen, indem sie auf einer virtuellen Maschine für Java ausgeführt werden. Wir haben diese Idee in einem Werkzeug namens *Safe Sulong* umgesetzt. LLVM ist ein Compiler-Framework, das die Übersetzung vieler unsicherer Eingabesprachen in eine Zwischendarstellung ermöglicht, die wir ausführen. Indem wir uns auf den Just-in-Time-Compiler der Java Virtual Machine (JVM) verlassen, der für eine sichere Sprache entwickelt wurde, können wir unsichere Sprachen sicher optimieren (d.h. ohne undefiniertes Verhalten in unseren Optimierungen ausnutzen), während die spezifizierte Semantik der unsicheren Sprachen eingehalten wird. Wir haben zwei Arten entwickelt, wie undefiniertes Verhalten in diesem Ansatz behandelt wird. Wir stellen einen Sanitizer-Modus vor, der ungültige Speicherzugriffe und andere Fehler erkennt. Wir haben ihn in Hinblick auf seine Wirksamkeit beim Finden von Fehlern in Open-Source Projekten und in Bezug auf seine Ausführungsgeschwindigkeit evaluiert. Außerdem haben wir einen *Lenient C* Modus entwickelt, der undefiniertes Verhalten in C so spezifiziert, wie es häufig von ProgrammierInnen erwartet wird.

**Empirische Studien.** Wir präsentieren zwei empirische Studien zur Verwendung unstandardisierter Elemente in C-Projekten, um WerkzeugentwicklerInnen zu helfen, diese Elemente bei der Implementierung von Tools für C-Code zu priorisieren. In diesen Studien haben wir die Verwendung von x86-64 Inline-Assembly und die Verwendung von GCC-Builtins in einer großen Anzahl von Open-Source-Projekten analysiert. Nach unserem besten Wissen sind dies die ersten Studien zur Verwendung von Inline-Assembly und Compiler-Builtins.

**Introspektion.** Wir stellen einen neuen Ansatz vor, mit dem ProgrammiererInnen die Robustheit ihrer C-Bibliotheken erhöhen können. Dieser Ansatz basiert auf einer Introspektions-Schnittstelle, die ProgrammiererInnen das Abfragen von Metadaten wie z.B. Objektgrößen und -typen ermöglicht. Wir haben diesen Ansatz in einer Fallstudie einer *libc*-Implementierung für Safe Sulong evaluiert. Außerdem haben wir einen Ausführungsgmodus entwickelt, der Fehler abwehren kann ohne die Ausführung zu stoppen (namens *Context-aware Failure-oblivious Computing*). Wir haben gezeigt, dass Introspektion nicht nur in Safe Sulong, sondern auch in anderen Sanitizern wie ASan [Se12], MPX-basiertes Prüfen von Objektgrenzen und SoftBound+CETS [RPM18] implementiert werden kann. Wir haben sowohl die Geschwindigkeit des Ansatzes, als auch die Wirksamkeit in einer Studie von Fehlern in weit verbreiteten Programmen evaluiert.

**Publikationen, Auszeichnungen und Vorträge.** Im Zuge der Dissertation haben wir einen Journal-, fünf Konferenz-, zwei Workshopartikel und fünf weitere begutachtete Beiträge publiziert. Desweiteren befinden sich zur Zeit zwei Manuskripte unter Begutachtung. Den wichtigsten Beitrag der Dissertation, Safe Sulong, haben wir in *ASPLOS '18* publiziert, einer Konferenz, die eine Akzeptanzrate von 18% hatte. Der Autor gewann die *ACM Student Research Competition* auf der *Programming '18* Konferenz, wurde zu Vorträgen an der *University of Cambridge*, am *Imperial College London* und an der *Universität Salzburg* eingeladen und hielt Vorträge an diversen Entwicklerkonferenzen.

**Auswirkungen auf die Industrie.** Sulong wurde in die *GraalVM* integriert,[1], eine mehrsprachige virtuelle Maschine, die von Oracle entwickelt wurde. Hierbei wird Sulong zum Ausführen von LLVM-basierten Sprachen verwendet und wird aktuell von mehreren Oracle-Ingenieuren weiterentwickelt. Durch den Mechanisms für Sprachinteroperabilität von GraalVM können andere Sprachimplementierungen Sulong als sichere native Funktionsschnittstelle [Gr15] verwenden. Der Kern von Sulong ist auf *GitHub* verfügbar und ist mit über 600 Sternen sehr beliebt.[2] In der Evaluierung von Safe Sulong haben wir außerdem insgesamt 68 Fehler in kleinen Open-Source Projekten gefunden, für die wir eine Fehlerbeschreibung erstellt und eine Behebung des Problems vorgeschlagen haben.[3] Desweiteren haben wir unimplementierte Funkionalität und Probleme in anderen Werkzeugen wie z.B. ASan entdeckt, die wir den Entwicklern gemeldet haben.[4] Die empirischen Studien haben uns dabei geholfen, Fehler in Tools zu finden, die für sicherheitskritische Anwendungen verwendet werden, z. B. fehlerhafte GCC-Builtin Implementierungen in Frama-C[5] [Cu12] und CompCert[6] [Le09].

## 5 Safe Sulong

Um undefiniertes Verhalten in C Programmen in Angriff zu nehmen haben wir einen Ansatz für die Ausführung unsicherer Sprachen auf der JVM entwickelt. Die Kernidee dieses Ansatzes ist, dass die Semantik eines unsicheren Programms auf die sichere Semantik eines Java-Programms abgebildet werden kann. Durch das Übersetzen einer unsicheren Operation auf eine Folge äquivalenter Java-Operationen verhalten sich beide Ausführungen für legale Eingaben gleich. Da Java jedoch eine sichere Sprache ist und die Semantik der Operationen vollständig spezifiziert ist, verhält sich der Java-Code auch für Operationen, die im C-Code unzulässig sind, in einer definierten Weise. Ein Pufferüberlauf in Java resultiert beispielsweise immer in einer *Exception*, die die Ausführung in einer definierten Weise abbricht. Die Java Virtual Machine optimiert das Programm basierend auf der klar definierten Java-Semantik. Pufferüberläufe und andere Fehler werden somit nicht wegoptimiert. Wir haben diesen Ansatz als ein System namens *Safe Sulong* implementiert.

**Sichere Ausführung.** Safe Sulong verfügt über zwei Ausführungsmodi, die auf undefiniertes Verhalten unterschiedlich reagieren. Der erste Ausführungsmodus erkennt undefiniertes Verhalten und bricht die Ausführung in einem solchen Fall ab. Der Modus stützt sich auf die automatischen Laufzeitprüfungen der zugrunde liegenden JVM, die z.B. durchgeführt werden, um Pufferüberläufe zu erkennen. In diesem Modus kann Safe Sulong als Sanitizer verwendet werden. Dies ist besonders für ProgrammiererInnen hilfreich, die Fehler in ihren Programmen während der Entwicklung und beim Testen korrigieren können. Der Modus verhindert jedoch, dass Software mit undefiniertem Verhalten ausgeführt werden kann, auch wenn beispielsweise SystemadministratorInnen dieses in bestimmten Fällen als harmlos betrachten. Um die Ausführung solcher Programme zu

[1] `http://www.graalvm.org/`
[2] `https://github.com/graalvm/sulong`
[3] `http://ssw.jku.at/General/Staff/ManuelRigger/ASPLOS18-SafeSulong-Bugs.csv`
[4] `https://github.com/google/sanitizers/issues?utf8=%E2%9C%93&q=author%3Amrigger`
[5] `https://lists.gforge.inria.fr/pipermail/frama-c-discuss/2018-July/005483.html`
[6] `https://github.com/AbsInt/CompCert/issues/243`

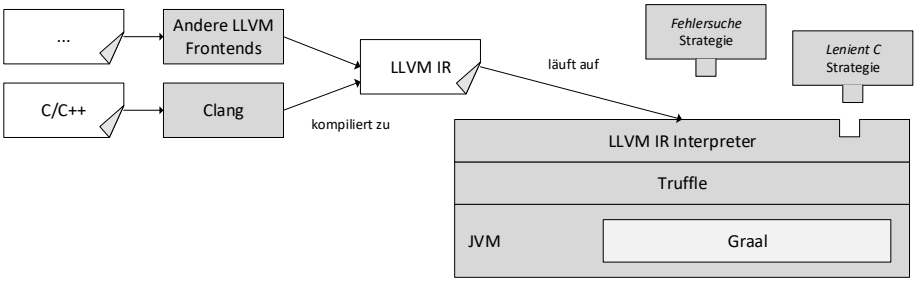



Figure 1: Überblick über die Architektur von Safe Sulong

ermöglichen, haben wir einen zweiten *Lenient C* Ausführungsmodus entwickelt, der eine Semantik für undefinierte Operationen definiert, basierend auf den Erwartungen von ProgrammiererInnen. Bei einem arithmetischen Überlauf definiert Lenient C beispielsweise ein Verhalten, das äquivalent zu dem von x86 Prozessoren ist.

**Architektur.** Abbildung 1 stellt die Architektur von Safe Sulong überblicksmäßig dar. Safe Sulong basiert auf existierenden Sprach-Frontends des LLVM [LA04]-Compiler-Frameworks, z.B. Clang, um unsichere Eingabesprachen in LLVM-Zwischencode (LLVM IR) zu übersetzen. Beim Kompilieren einer unsicheren Sprache nach LLVM IR deaktivieren wir alle Optimierungen, sodass Operationen die undefiniertes Verhalten verursachen nicht wegoptimiert werden. Der RISC-ähnliche Zwischencode wird auf dem LLVM-IR-Interpreter, den wir in Java implementiert haben (er umfasst ca. 60.000 Codezeilen), ausgeführt. Wie bereits beschrieben, umfasst der Interpreter verschiedene Modi, wie den Lenient-C-Modus und den Fehlersuche-Modus, die als Laufzeitstrategien konfiguriert werden können. Die Implementierung des Interpreters basiert auf dem Sprachimplementierungs-Framework Truffle, das den Just-in-Time-Compiler Graal verwendet, um häufig ausgeführte Funktionen zu Maschinencode zu kompilieren [Wu13]. Da Graal so entwickelt wurde, dass er effizient Prüfungen optimiert (z. B. durch das Entfernen redundanter Begrenzungsprüfungen gegen Pufferüberläufe), erreicht unser Prototyp eine Ausführungsgeschwindigkeit, die mit bestehenden Compilern, die unsicheren Code erzeugen, konkurrenzfähig ist. Wir haben Safe Sulong implementiert und im Hinblick auf die Effektivität des Fehlersuchmodus und der Leistung evaluiert.

**Wirksamkeit.** Um zu testen, ob Safe Sulong ein wirksames Werkzeug zur Fehlersuche ist, haben wir C-Projekte von Github ausgewählt und mit Safe Sulong ausgeführt, um potenzielle Fehler darin zu finden. Wir wollten auch zeigen, dass in aktuellen Ansätzen Fehler übersehen werden, die jedoch Safe Sulong erkennt. Zu diesem Zweck haben wir jedes der fehlerhaften Programme unter den gleichen Bedingungen mit ASan und Valgrind ausgeführt, d.h. mit den gängigsten Werkzeugen der Compilezeit- und Laufzeitinstrumentierung, um zu überprüfen, ob auch sie die von Safe Sulong erkannten Fehler finden. Insgesamt fanden wir mit Safe Sulong 68 Fehler, wobei es sich beim Großteil um Pufferüberläufe handelte. Danach haben wir die Programme mit Clang und ASan ohne Optimierungen (-O0) kompiliert, da wir auch mit den anderen Werkzeugen möglichst viele Fehler finden wollten. Um zu zeigen, dass das Anwenden von Optimierungen dazu führt, dass bestimmte Fehler nicht erkannt werden, haben wir zusätzlich die Programme auf der Optimierungsstufe -O3 für ASan und Valgrind kompiliert. Wie erwartet, fand Valgrind in beiden Konfigurationen nur etwa die Hälfte der Fehler, da es Pufferüberläufe am Stack und auf globalen Variablen nicht zuverlässig erkennt. ASan ohne Optimierungen hat 60 der 68 Fehler gefunden. Mit Optimierungen wurden nur 56 Fehler gefunden, da die anderen von Clang wegoptimiert wurden. Insgesamt konnten von den 68 Fehlern, die Safe Sulong entdeckte, 8 weder von Valgrind noch von ASan in allen Konfigurationen gefunden werden. Einerseits konnten diese Werkzeuge Fehler aufgrund grundlegender Ein-

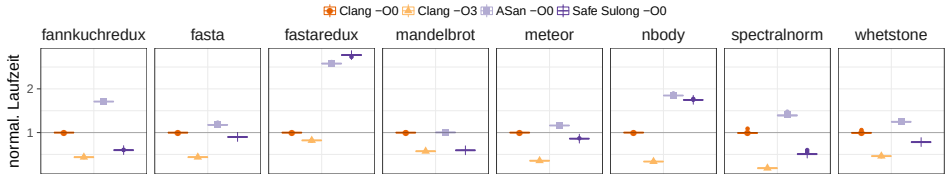



Figure 2: Geschwindigkeit auf den *Computer Language Game Shootouts*.

schränkungen ihrer Ansätze nicht erkennen und weil sie auf Compilern basieren, die undefiniertes Verhalten unter Umständen wegoptimieren. Diese Probleme können nicht ohne weiteres gelöst werden. Andererseits übersahen sie Fehler aufgrund von Problemen in ihrer Implementierung, was durch Verbesserungen oder Korrekturen behoben werden könnte. Allerdings ist unser Ansatz von Natur aus zuverlässiger, da er auf der gut getesteten Implementierung von Java und seiner klar definierten Semantik beruht, was das Wegoptimieren von Fehlern vermeidet.

**Geschwindigkeit.** In unserer Geschwindigkeitsevaluierung wollten wir zeigen, dass Safe Sulong vergleichbar schnell wie andere Werkzeuge läuft. Ein direkter Vergleich der Laufzeitleistung zwischen verschiedenen Tools ist jedoch nicht fair, da diese unterschiedliche Funktionen bieten. Unsere Messungen sollen daher nur zeigen, dass unser Ansatz praxistauglich ist. Neben der Geschwindigkeit von Safe Sulong haben wir daher auch die von ASan, basierend auf LLVM Version 3.9, und die von Valgrind, Version 3.12, gemessen. Als Baseline haben wir auch die Geschwindigkeit von Programmen gemessen, die mit Clang Version 3.9 kompiliert wurden (jeweils mit und ohne Optimierungen), deren Ausführung daher unsicher ist. Für alle anderen Werkzeuge (auch für Safe Sulong) haben wir die Optimierungen von Clang ausgeschaltet, da die Annahme weiterhin galt, dass es unser Ziel ist, so viele Fehler wie möglich zu finden. Wir evaluierten die Werkzeuge auf den Benchmarks des *Computer Language Benchmark Games*, die entworfen wurden, um die Geschwindigkeitsunterschiede zwischen verschiedenen Programmiersprachen zu zeigen. Wir mussten die adaptiven Kompilierungstechniken von Truffle und Graal berücksichtigen, indem wir einen Benchmark-Harness implementierten, der Aufwärmiterationen durchführte. Durch die Ausführung von 50 In-Prozess-Warmup-Iterationen haben wir sichergestellt, dass jeder Benchmark einen stabilen Zustand erreicht hat. Abbildung 2 zeigt Box-Plots für die Ausführungsgeschwindigkeit im Verhältnis zu `Clang -O0`. Wir haben Valgrind vom Graphen ausgeschlossen, da es `10`× bis `58`× langsamer war als `Clang -O0` bei 5 Benchmarks. Die geringste Verlangsamung stellten wir auf `spectralnorm`, `fasta`, und `fannkuchredux` fest (`2.3`, `3.6` und `5.1`). Die Ergebnisse für den Benchmark `binarytrees` haben wir auch ausgeschlossen, weil ASan `14`× langsamer und Valgrind `58`× langsamer als `Clang -O0` war. Diese Verlangsamung war darauf zurückzuführen, dass `binarytrees` zuweisungsintensiv war, was darauf hindeutet, dass die derzeitigen Fehlersuchansätze mit allokationsintensiven Benchmarks nicht gut umgehen können. Auf diesem Benchmark war Safe Sulong nur `1.7`× langsamer als `Clang -O0`. Auf fast allen Benchmarks war Safe Sulong schneller als `ASan`; nur auf `fastaredux` waren beide Werkzeuge etwa gleich schnell. Safe Sulong war außerdem schneller als Programme kompiliert mit `Clang -O0`, außer bei den `fastaredux` und `nbody` Benchmarks. Auf `fannkuchredux` und `mandelbrot` war Safe Sulong sogar gleich schnell wie Programme die mit `Clang -O3` kompiliert wurden. Die langsamste Geschwindigkeit hatte Safe Sulong bei `fastaredux`, wo es `2.5`× langsamer war als `Clang -O0`. Insgesamt zeigen die Ergebnisse, dass die Leistung von Safe Sulong mit anderen Werkzeugen konkurrenzfähig ist und in einigen Fällen sogar die Leistung von Programmen erreicht, die mit unsicheren Compilern kompiliert wurden.

## 6 Empirische Studien

Als wir Safe Sulong verwendeten um Programme „aus der echten Welt“ auszuführen stellen wir fest, dass viele Programme *unstandardisierte Elementen* wie Inline-Assembly und GCC-Compiler-

Builtins verwendeten. Sie sind Compiler-Erweiterungen und nicht Bestandteil eines offiziellen C-Standards. Inline-Assembly ermöglicht das Einbetten von Maschinencode-Anweisungen direkt in den C-Code. Compiler-Builtins ähneln Standard-Bibliotheksfunktionen, werden jedoch direkt im Compiler implementiert. Unter Linux sind die GCC-Builtins weit verbreitet. Der Aufwand für die Implementierung unstandardisierter Elemente in einem Werkzeug (z. B. einem Compiler oder einem Sicherheitsanalyse-Werkzeug) ist umständlich, da selbst eine Architektur mit einem einzigen komplexen Befehlssatz, wie z. B. x86, etwa 1.000 Anweisungen enthält, und da über 12.000 GCC-Builtins existieren. Um die Implementierung unstandardisierter Elemente in Safe Sulong und anderen Tools zu priorisieren, haben wir deren Verwendung in C/C++-Open-Source-Projekten von GitHub mit Repository-Mining-Methoden untersucht. Wir haben mehr als 1.200 Projekte für die Inline-Assembly Studie analysiert, und fast 5.000 Projekte für die Studie zu Compiler-Builtins.

**Häufigkeit.** Wir haben festgestellt, dass sowohl Inline-Assembly Fragmente als auch GCC-Builtins weit verbreitet sind; 38% der Projekte verwendeten GCC-Builtins, während 28% der populärsten Projekte Inline-Assembly verwendeten. Normalerweise werden sie jedoch nur an wenigen Stellen im Quellcode verwendet. GCC-Builtins werden normalerweise alle 6.000 Codezeilen (LOC) verwendet, während Inline-Assembly Fragmente sogar nur alle 40.000 LOCs verwendet werden. Unsere Ergebnisse legen nahe, dass ProgrammiererInnen nur einen Teil der verfügbaren unstandardisierten Elemente verwenden. Insgesamt wurden nur ca. 3.000 verschiedene GCC-Builtins und ca. 200 verschiedene Inline-Assembly Fragmente verwendet. Darüber hinaus analysierten wir die historische Entwicklung der Projekte und stellten fest, dass sie meistens GCC-Builtins hinzufügten und selten entfernten. Dies legt nahe, dass unstandardisierte Elemente wahrscheinlich auch in zukünftigen Tools unterstützt werden müssen.

**Werkzeug-Unterstützung.** Um Werkzeug-Entwickler zu informieren, haben wir verschiedene Implementierungsstrategien evaluiert. Beispielsweise zeigt Abbildung 3 die Anzahl der GCC-Builtins, die implementiert werden müssen, um einen bestimmten Prozentsatz von Projekten zu unterstützen, wenn eine *gierige* Implementierungsreihenfolge angenommen wird. Die Anzahl der zu implementierenden Builtins steigt exponentiell an; Um die Hälfte der Projekte zu unterstützen, müssen 22 Builtins implementiert werden. Um 90% der Projekte zu unterstützen, müssen 106 Builtins implementiert werden und um 99% zu unterstützen, müssen 1.600 Builtins implementiert werden. Wir haben auch untersucht, inwieweit beliebte Werkzeuge Builtins unterstützen, indem wir eine Testsuite schrieben, mit der wir die Implementierung der am häufigsten verwendeten Builtins testeten. Während ausgereifte Compiler alle unsere Testfälle bestanden haben, konnten viele andere Tools aufgrund von falsch implementierten und fehlenden Builtin-Implementierungen nicht alle Testfälle erfolgreich ausführen. Interessanterweise fanden wir zwei fehlerhaft implementierte Builtins im formal verifizierten CompCert-Compiler, der hauptsächlich für sicherheitskritische Anwendungen verwendet wird, da seine Builtin-Implementierungen nicht verifiziert wurden.

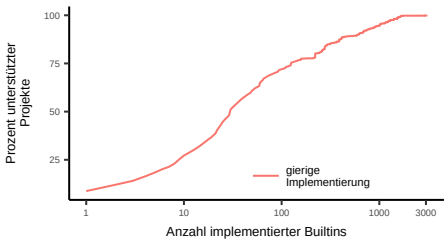



Figure 3: Implementierungsaufwand der GCC-Builtins; Beachten Sie die Exponentialachse.

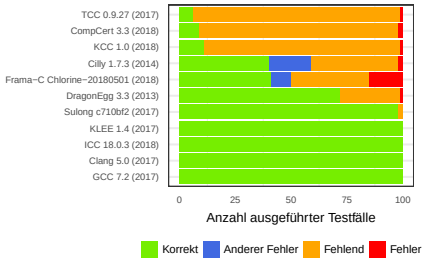



Figure 4: Evaluierung der Unterstützung von GCC-Builtins in verschiedenen Werkzeugen

List. 1: Robuste Implementierung von `strlen()`, die auch für unterminierte Strings funktioniert.

```
size_t strlen(const char *str) {
  size_t len = 0;
  while (size_right(str) > 0 && *str != '\0') {
    len++; str++;
  }
  return len;
}
```

## 7 Introspektion

In einigen Fällen können ProgrammiererInnen besser als automatische Ansätze wie Safe Sulong auf undefiniertes Verhalten reagieren, zum Beispiel indem sie diese Invarianten überprüfen und etwaige Fehler beheben. ProgrammiererInnen können jedoch keine Objekt-Metadaten wie z.B. Objektgrößen abfragen um illegale Zustände oder Parameter zu überprüfen, da der C-Sprache solche Mechanismen fehlen. Da immer mehr Tools zur Fehlererkennung und Fehlersuche Metadaten aufzeichen, um ihre Prüfungen zu implementieren, haben wir eine Introspektionsschnittstelle entwickelt, mit der ProgrammiererInnen diese Metadaten abfragen können, um undefiniertes Verhalten manuell anzugehen. Diese Introspektionsschnittstelle bietet beispielsweise die Funktion `size_right()` um die Grenze eines Objekts abzufragen. Wir haben die Introspektionsfunktionen in Safe Sulong implementiert und gezeigt, dass das Verfahren auch auf andere Fehlersuchwerkzeuge anwendbar ist, indem wir `size_right()` in ASan, SoftBound, und in MPX-basierter Grenzüberprüfung implementiert haben. Außerdem haben wir basierend auf Introspektion mit verschiedenen Anwendungsfällen experimentiert.

**Context-aware Failure-oblivious Computing.** Ein Anwendungsfall von Introspektion ist das Abschwächen von Fehlern, ohne die Ausführung des Programms zu beenden, was für Systeme mit hohen Verfügbarkeitsanforderungen (z.B. Servern) nützlich ist. Unsere Idee basiert auf dem *Failure-oblivious Computing* Ansatz, bei dem ungültige Schreib- und Lesevorgänge ignoriert werden [Ri04]. In unserem *Context-aware Failure-oblivious Computing* Ansatz implementieren ProgrammiererInnen manuell Code, der illegale Eingaben behandelt, wobei sie die Semantik der Funktion berücksichtigen, in der die Logik implementiert ist. Dies wird durch List. 1 demonstriert, in dem `strlen()`, eine Funktion die die Länge eines Strings berechnet, auch die Länge berechnen kann wenn das '\0' Zeichen, das für gültige C Strings notwendig ist, fehlt. In einem solchen Fall nimmt die Funktion mittels `size_right()` die Länge des zugrunde liegenden Puffers als Stringlänge an. In vielen Fällen hilft dies dabei, die Ausführung ohne Folgefehler fortzusetzen. Wir haben auch verschiedene andere libc-Funktionen verbessert, um Fehler, die durch ungültige Argumente verursacht werden, abzuschwächen. Ein anderer Anwendungsfall, mit dem wir experimentiert haben, ist die Härtung der C-Standardbibliothek, indem Invarianten der übergebenen Argumente überprüft werden.

**Wirksamkeit.** Um die Anwendbarkeit von Context-aware Failure-oblivious Computing in realen Projekten zu demonstrieren, haben wir jüngste (d.h. weniger als ein Jahr alte) Pufferüberläufe in weit verbreiteter Software wie Dnsmasq, Libxml2 und GraphicsMagick betrachtet. Wir haben fünf geeignete Fehler in der *Common Vulnerabilities and Exposures*-Datenbank gefunden, bei denen in einer libc-Funktion ein Pufferüberlauf aufgetreten ist. Anschließend haben wir die Anwendungen mit den durch Introspektion erweiterten libc-Funktionen ausgeführt, mit Eingabeparametern die den Fehler auslösen. In vier Fällen konnte die Ausführung erfolgreich fortgesetzt werden. In einem Fall führte ein nachfolgender Pufferüberlauf in der Anwendung dazu, dass die Werkzeuge die Ausführung kontrolliert beendeten. Dies war zu erwarten, da unser Ansatz nur für Funktionen funktioniert, die mit Introspektionsfunktionen erweitert wurden.

**Geschwindigkeit.** Um die Geschwindigkeit des auf Introspektion basierenden Failure-Oblivious-Computing Ansatzes zu ermitteln, haben wir unterschiedliche Arbeitslasten betrachtet. Erstens haben wir Server in Betracht gezogen, weil sie eine hohe Angriffsoberfläche bieten und sie hochverfügbar sein müssen. Wir haben festgestellt, dass der Introspektions-Mehraufwand keinen Einfluss auf ihre Laufzeitleistung hat. Zweitens haben wir CPU-intensive Lasten betrachtet, nämlich die SPEC CPU 2016-Benchmarks. Bei MPX und SoftBound war der Mehraufwand meist gering und reichte bis zu 13%. Bei ASan war der Mehraufwand jedoch signifikant und reichte bis zu 130%. Dies war zu erwarten, da ASan die Grenzen eines Objektes nicht explizit speichert. Die Berechnung in `size_right()` hat daher eine Komplexität von *O(n)*.

# 8 Zusammenfassung

Programme, die in unsicheren Sprachen wie C/C++ geschrieben sind, können undefiniertes Verhalten verursachen, was problematisch für die korrekte Ausführung der Programme und die Sicherheit von Computersystemen darstellt. Um undefiniertes Verhalten in Angriff zu nehmen, hat die vorliegende Arbeit in drei Bereichen beigetragen.

**Safe Sulong.** Erstens haben wir einen Ansatz zur sicheren Ausführung unsicherer Sprachen entwickelt, bei dem ein Interpreter in einer sicheren Programmiersprache geschrieben ist. Wir haben diesen Ansatz als Werkzeug namens Safe Sulong implementiert. Unsere Evaluierung hat gezeigt, dass Safe Sulong Fehler in Sonderfällen erkennt, die andere Tools übersehen (z. B. weil Instrumentierung für Sonderfälle weggelassen wurde). Der Vorteil von Safe Sulong ist es, dass es auf automatischen Laufzeitprüfungen der zugrunde liegenden virtuellen Maschine basiert und einen optimierenden Compiler verwendet, der undefiniertes Verhalten nicht ausnutzt. Außerdem haben wir demonstriert, dass Programme, die von Safe Sulong ausgeführt laufen, in einigen Fällen in etwa gleich schnell ausgeführt werden wie Programme, die von unsicheren Compilern kompiliert wurden. Safe Sulong ist jedoch immer noch ein Forschungsprototyp und muss noch hinsichtlich seiner Vollständigkeit erweitert werden, um sein Verhalten bei großen Programmen zu evaluieren.

**Empirische Studien.** Zweitens haben wir die Verwendung von C-Spracherweiterungen, nämlich Inline-Assembly und GCC-Builtins, in C-Projekten analysiert. Wir haben festgestellt, dass diese unstandardisierte Elemente häufig verwendet werden, was Werkzeugautoren Anreize bietet, sie in vorhandenen Werkzeugen zur Fehlersuche zu unterstützen. Aufgrund des hohen Implementierungsaufwands ist eine vollständige Implementierung jedoch manchmal nicht realisierbar, da beispielsweise über 10.000 GCC-Builtins existieren. Unsere Ergebnisse deuten darauf hin, dass bereits durch die Implementierung einer kleinen Teilmenge von GCC-Builtins und Inline-Assembly eine große Anzahl von Projekten unterstützt werden kann, da die meisten Projekte eine gemeinsame Teilmenge dieser Elemente verwenden. Unsere Analyse der historischen Entwicklung von GCC-Builtins in Projekten legt nahe, dass es sich nicht um eine stagnierende Funktionalität handelt, so dass Werkzeuge sie wahrscheinlich auch in Zukunft unterstützen müssen.

**Introspektion.** Drittens haben wir einen Ansatz vorgeschlagen, um ProgrammierInnen Metadaten zur Verfügung zu stellen, die von vorhandenen dynamischen Fehlersuchwerkzeugen aufgezeichnet werden, damit diese die Metadaten verwenden können, um die Robustheit von Bibliotheken zu verbessern. Wir haben diesen Ansatz in verschiedenen Werkzeugen implementiert um zu zeigen, dass der Ansatz nicht nur mit Safe Sulong funktioniert. Wir haben Context-aware Failure-oblivious Computing als eine auf Introspektion basierende Technik vorgeschlagen, die es erlaubt, Programme auszuführen auch wenn Pufferüberläufe auftreten. Wir haben diesen Ansatz hinsichtlich der Wirksamkeit in einer Fallstudie mit realen Fehlern in gängigen Anwendungen und hinsichtlich der Geschwindigkeit evaluiert. Die Ergebnisse deuten an, dass Context-aware Failure-oblivious Computing dazu verwendet werden kann, um Fehler zu mindern und Programme weiter auszuführen. Bei Tools, bei denen der Introspektionszugriff effizient implementiert werden konnte,

waren die Geschwindigkeitseinbußen von Introspektion gering. Daher glauben wir, dass Context-aware Failure-oblivious Computing in Industrieszenarien eingesetzt werden könnte.

## References

[Cu12] Cuoq, Pascal; Kirchner, Florent; Kosmatov, Nikolai; Prevosto, Virgile; Signoles, Julien; Yakobowski, Boris: Frama-C: A Software Analysis Perspective. In: Proceedings of SEFM'12. Springer-Verlag, Berlin, Heidelberg, pp. 233–247, 2012.

[Gr15] Grimmer, Matthias; Seaton, Chris; Würthinger, Thomas; Mössenböck, Hanspeter: Dynamically Composing Languages in a Modular Way: Supporting C Extensions for Dynamic Languages. In: Proceedings of MODULARITY 2015. ACM, New York, NY, USA, pp. 1–13, 2015.

[LA04] Lattner, Chris; Adve, Vikram: LLVM: A Compilation Framework for Lifelong Program Analysis & Transformation. In: Proceedings of CGO '04. IEEE Computer Society, Washington, DC, USA, pp. 75–, 2004.

[Le09] Leroy, Xavier: Formal Verification of a Realistic Compiler. Commun. ACM, 52(7):107–115, July 2009.

[NS07] Nethercote, Nicholas; Seward, Julian: Valgrind: A Framework for Heavyweight Dynamic Binary Instrumentation. In: Proceedings of PLDI '07. ACM, New York, NY, USA, pp. 89–100, 2007.

[Ri04] Rinard, Martin; Cadar, Cristian; Dumitran, Daniel; Roy, Daniel M.; Leu, Tudor; Beebee, Jr., William S.: Enhancing Server Availability and Security Through Failure-oblivious Computing. In: Proceedings of OSDI'04. USENIX Association, Berkeley, CA, USA, pp. 21–21, 2004.

[RPM18] Rigger, Manuel; Pekarek, Daniel; Mössenböck, Hanspeter: Preventing Buffer Overflows by Context-aware Failure-oblivious Computing. In: Proceedings of NSS '18. 2018.

[Se12] Serebryany, Konstantin; Bruening, Derek; Potapenko, Alexander; Vyukov, Dmitry: AddressSanitizer: A Fast Address Sanity Checker. In: Proceedings of USENIX ATC'12. USENIX Association, Berkeley, CA, USA, pp. 28–28, 2012.

[Wa12] Wang, Xi; Chen, Haogang; Cheung, Alvin; Jia, Zhihao; Zeldovich, Nickolai; Kaashoek, M. Frans: Undefined Behavior: What Happened to My Code? In: Proceedings of APSYS '12. ACM, New York, NY, USA, pp. 9:1–9:7, 2012.

[Wu13] Wuerthinger, Thomas; Wimmer, Christian; Wöß, Andreas; Stadler, Lukas; Duboscq, Gilles; Humer, Christian; Richards, Gregor; Simon, Doug; Wolczko, Mario: One VM to Rule Them All. In: Proceedings of Onward! 2013. ACM, New York, NY, USA, pp. 187–204, 2013.

**Manuel Rigger** wurde am 13.09.1990 in Schwaz, Österreich geboren. Er begann seine Informatiker- und Forscher-Karriere mit einem Bachelorstudium an der *Johannes Kepler Universität (JKU) Linz*. Während seines *Software Engineering* Masterstudiums an der JKU absolvierte er ein Auslandssemester an der *National Taiwan University*. Bevor er sein Masterstudium abschloss verweilte er zwei Jahre lang in China, wo er an der *Xiamen University* ein Masterstudium in Chinesischer Philosophie absolvierte. Seine neu-erlangten Chinesischkenntnisse inspirierten den Namen seines Informatik-Dissertationsprojektes *Sulong*. In 2018 promovierte er an der JKU Linz unter der Betreuung von Prof. Dr. Dr. h.c. Hanspeter Mössenböck. Ab Februar 2019 wird er an der ETH Zürich als Postdoktorand unter der Betreuung von Prof. Dr. Zhendong Su arbeiten.

# Wissensbildung in der visuellen Datenanalyse: Integration menschlicher und maschineller Intelligenz für die Exploration großer Datenmengen [1]

Dominik Sacha[2]

**Abstract:** Das Volumen heutiger Datensätze macht eine manuelle Inspektion der Daten unmöglich und automatisierte Analyseverfahren aus dem Data Mining oder maschinellen Lernen können meist nicht vollautomatisiert auf reale Probleme angewandt werden. Um dieses Problem zu lösen, bindet Visual Analytics den Menschen in den Analyseprozess ein. Diese Dissertation wählt eine konzeptionelle, modell-getriebene Herangehensweise an Visual Analytics, welche die Zusammenarbeit von Mensch und Maschine während des Wissensbildungsprozesses fokussiert. Zusätzlich trägt diese Dissertation neuartige Verfahren zur Untersuchung und Unterstützung von menschlichen Wissensbildungsprozessen bei. Diese Verfahren werden in enger Zusammenarbeit mit echten Experten aus verschiedenen Anwendungsgebieten entwickelt. Die Ergebnisse zeigen, dass die konzeptionelle Herangehensweise zu neuartigen, eng integrierten Verfahren führt, welche die Analysten bei der Wissensbildung unterstützen. Diese haben das gemeinsame Ziel, die Exploration großer Datenmengen effektiver, effizienter und transparenter zu machen.

## 1 Motivation

Die Menschheit ist einem enormen Informationsüberfluss ausgesetzt [Ke10] und es bleibt eine Herausforderung, diese oft verrauschten Daten zu analysieren und die darin versteckten Informationen zu extrahieren. Rein automatische Verfahren oder konventionelle Statistik reichen Domänen-Experten oft nicht aus, um die echten Analyseprobleme der Realität anzugehen. Dafür wird eine explorative Datenanalyse benötigt, welche es ermöglicht, Hypothesen zu formen und schrittweise zu verfeinern, um diese letztendlich zu bestätigen bzw. zu widerlegen. Computer sind in der Lage in kürzester Zeit eine enorme Anzahl an Berechnungen durchzuführen und können dabei auf einen Speicher zurückgreifen, welcher nur durch Hardware begrenzt ist. Menschen können dagegen Wissen, das mehrere Facetten hat (z.B. Erfahrungen, taktisches und Domänen-Wissen, Fähigkeiten), nutzen – Informationen, die der Maschine oft nicht zugänglich sind –, was ihnen erlaubt, über die Probleme nachzudenken und Schlussfolgerungen zu ziehen. Visual Analytics (visuelle Datenanalyse, VA) bietet eine visuelle Schnittstelle zwischen automatisierten Techniken und dem menschlichen Analysten mit dem Ziel, die menschlichen und maschinellen Stärken effektiv zu vereinen [Ke10]. Diese Dualität resultiert in einer Konversation zwischen Mensch und Maschine, wobei berechnete (Zwischen-) Ergebnisse durch Visualisierungen kommuniziert und menschliches Feedback durch Benutzer-Interaktion ausgedrückt wird. Vielen

[1] Englischer Titel der Dissertation: “Knowledge Generation in Visual Analytics: Integrating Human and Machine Intelligence for Exploration of Big Data”

[2] AG für Datenanalyse und Visualisierung (DBVIS), Universität Konstanz, sacha@dbvis.inf.uni-konstanz.de

Analysesystemen mangelt es an der Fähigkeit, den Menschen effektiv während des Analyseprozesses dabei zu unterstützen, Wissen aus den Daten zu erzeugen [Sa14]. Menschen können sich nur eine begrenzte Anzahl an Informations-Stücken (eng. information chunks) merken [Mi56] und unsere Aufmerksamkeit innerhalb des Analyseprozesses wird durch bestimmte Antwortzeiten begrenzt [Mi68]. Zudem sind die Analysten oft von den Datenmengen und den Konfigurationsmöglichkeiten überfordert. Visuelle Benutzerschnittstellen müssen daher effektiv auf die menschlichen Stärken zugeschnitten werden. Um die VA-Forschung weiter voran zu bringen, werden "bessere und benutzbarere Lösungen" benötigt, damit Wissen aus den Daten extrahiert werden kann, mit der Herausforderung, den Bedürfnissen der Benutzer gerecht zu werden [Ke10].

## 2 Forschungsmethodik & Beiträge zur Forschung

Diese Doktorarbeit befasst sich mit der zentralen Fragestellung: **"Wie eine enge Integration zwischen automatisierter Analyse und visueller Interaktion erreicht werden kann, um die Wissensbildung in VA besser zu unterstützen?"**. Die Forschungsbeiträge dieser Doktorarbeit sind zweifältig (konzeptionell und methodisch):

**C1 Modell-getriebener Ansatz für VA:** Der Hauptbeitrag ist ein modell-getriebener Ansatz für VA, welcher neue Perspektiven auf die aktuell unter-erforschten Aspekte innerhalb des Wissensbildungsprozesses liefert. Diese resultierenden Artefakte (Ontologien, Richtlinien, konzeptionelle Prozessmodelle) bauen auf den existierenden Vorarbeiten auf und sind von echten Beispielsystemen inspiriert. Außerdem werden sie instantiiert, indem sie auf eine große Anzahl an Beispielsystemen angewandt werden, um deren Nützlichkeit zu testen und um sie mit den existierenden konzeptionellen Modellen in Einklang zu bringen (evaluierende Nutzung). Dieser Prozess resultiert in beidem, 1.) der Erweiterung der Prozessmodelle um fehlende Features und 2.) der Zusammenfassung (Vereinfachung) von gemeinsamen Aspekten über mehrere Beispielsysteme und existierende theoretische Arbeiten. Außerdem ermöglicht der Prozess die Identifikation offener Forschungsgebiete für die Verbesserung des aktuellen Forschungsstandes und um neue Methoden zu generieren, die den Wissensbildungsprozess unterstützen (generative Nutzung). Dieser Ansatz ist durch die *Grounded Theory* [Ch14] inspiriert, in welcher Daten (hier Publikationen und Beispiele) systematisch analysiert werden, um Kategorien zu identifizieren und zu verfeinern, welche dann verwendet werden, um schrittweise ein theoretisches Modell zu bilden. Dieser Ansatz wurde bereits in der Visualisierungsforschung ( z.B. [Is13] und verwandten Forschungsgebieten, wie beispielsweise der Mensch-Maschine Interaktion [Ho06] angewandt und dessen Wichtigkeit, solch eine benötigte theoretische Grundlage in der Visualisierungsforschung aufzubauen, wurde bereits erkannt [Ch17]. C1 adressiert die Erfordernis einer theoretischen Grundlage und deren evaluierende und generative Nutzung mit einem konzeptionellen Beitrag zur VA Forschung.

**C2 Neuartige Methoden für die Analyse & Unterstützung von Wissensbildung:** Diese Dissertation beschreibt vier VA Systeme, welche die spezifischen, identifizierten Aspekte der Analyse und Unterstützung der Wissensbildung adressieren. Diese Systeme streben

eine **"enge Integration von visueller Interaktion und automatisierter Analyse"** an, um menschliche analytische Aktivitäten zu unterstützen. Die entsprechenden theoretischen Artefakte (C1) werden dabei dazu eingesetzt, die relevanten Forschungsmöglichkeiten zu identifizieren und die Erforschung von neuartigen VA Methoden zu lenken. Diese Methoden werden in enger Zusammenarbeit mit Domänen-Experten mit echten Daten und Analyseproblemen in einem *Benutzer-zentrierten Design Prozess* entwickelt. Dafür adoptiert diese Arbeit den *Design Study Methodology* Ansatz [SMM12], um die Prototypen zu erstellen und zu evaluieren. Die ersten Prototypen wurden in *Pair Analytics*-Settings (in denen ein Experte eines Fachgebiets mit einem VA Experte zusammenarbeitet) erprobt, um Daten zu analysieren, die Analysten zu trainieren, den Analyseprozess zu beobachten und um qualitatives Feedback einzuholen. Mit den fortgeschrittenen Versionen der Systeme wurden ebenfalls quantitative Benutzerstudien durchgeführt. Die neuartigen Methoden liefern Lösungsansätze zur Überwindung von Problemen während der Wissensbildung.

## 3 Aufbau der Arbeit und Kurzfassung der Kapitel

Die Dissertation ist in sechs Kapitel aufgeteilt, wobei das erste Kapitel den Forschungskontext und deren Struktur beschreibt. Es folgen die vier Hauptkapitel sowie ein abschließendes Kapitel mit einer finalen Zusammenfassung, einer Ontologie und Schlussfolgerungen. Jedes der Hauptkapitel (Kapitel 2–5) enthält zwei Unterkapitel: Der erste Teil stellt jeweils einen konzeptionellen Beitrag dar (konzeptionelles Prozessmodell), welcher im zweiten Unterkapitel zu einem methodischen Beitrag führt. Im Folgenden werden die Hauptkapitel im einzelnen etwas genauer beschrieben.

**Kapitel 2 - Konstruktion und Anwendung eines Wissensbildungsmodells für VA:** Der erste Teil des Kapitels beschreibt den konzeptionellen Kernbeitrag dieser Arbeit. Das **Wissensbildungsmodell für Visual Analytics** [Sa14] bietet eine Charakterisierung menschlicher und maschineller Konzepte innerhalb des Datenanalyseprozesses.

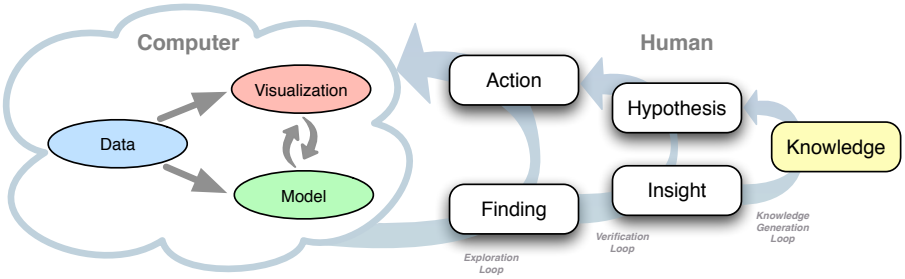



Abb. 1: Wissensbildungsmodell für VA: Es besteht aus Computer und menschlichen Teilen. Die linke Seite stellt ein VA System dar, die rechte Seite den menschlichen Wissensbildungsprozess.

Die linke Seite in Abbildung 1 illustriert ein VA System, welches **Daten** vorbereitet, auf automatische Methoden und **Modelle** anwendet und dann durch die Abbildung von Daten-Attributen auf visuelle Strukturen eine **Visualisierung** erzeugt. Die rechte Seite

zeigt den Wissensbildungsprozess des Menschen, ein Prozess schlussfolgernden Denkens, der sich aus **Explorations**-, **Verifikations**- und **Wissensbildungs**-Schleifen zusammensetzt. VA verfolgt eine enge Integration von Mensch und Maschine, die durch Benutzerinteraktion mit dem System erreicht wird. Es werden verschiedene Interaktionen beschrieben, welche beispielsweise die Datenvorverarbeitung, das Model-Building oder das visuelle Mapping, durch menschliche **Aktionen** beeinflussen. Jede Interaktion bewirkt eine observierbare Reaktion des Systems, in denen Analysten Muster erkennen können, die als Untersuchungsergebnisse genutzt werden (**Finding**). Aktionen und Findings sind Teil der **Explorationsschleife**. Die Findings werden durch deren Interpretation und Gegenüberstellung zu Erkenntnissen (**Insights**), welche immer wieder mit den **Hypothesen** abgeglichen werden. Die Hypothesen werden vom Vorwissen und von den erzeugten Erkenntnissen gebildet und geformt. Findings und Insights sind Teil der **Verifikationsschleife**. Wenn genügend Erkenntnisse für bestimmte Hypothesen gesammelt wurden und das Vertrauen in diese stark genug ist, um diese zu widerlegen oder zu bestätigen, dann bildet sich menschliches Wissen. Dieser Prozess wird **Wissensbildungsschleife** genannt, welche die Verifikationsschleife steuert. Gleichermaßen steuert die Verifikationsschleife die Explorationsschleife. Das Prozessmodell wird außerdem in Relation zu anderen existierenden Arbeiten wie die maschinellen Prozesse (z.B. [Ke10]), der Mensch-Maschine Interaktion (z.B. [No02]) oder den menschlichen Denkprozessen (z.B. [PC05]) gebracht. Es wird außerdem auf verschiedene Datenanalysesysteme angewandt, um diese zu evaluieren, um deren Stärken und Schwächen herauszuarbeiten. Im Anschluss werden die wichtigsten Aspekte diskutiert und zukünftige Forschungsmöglichkeiten für die Unterstützung des Wissensbildungsprozesses abgeleitet.

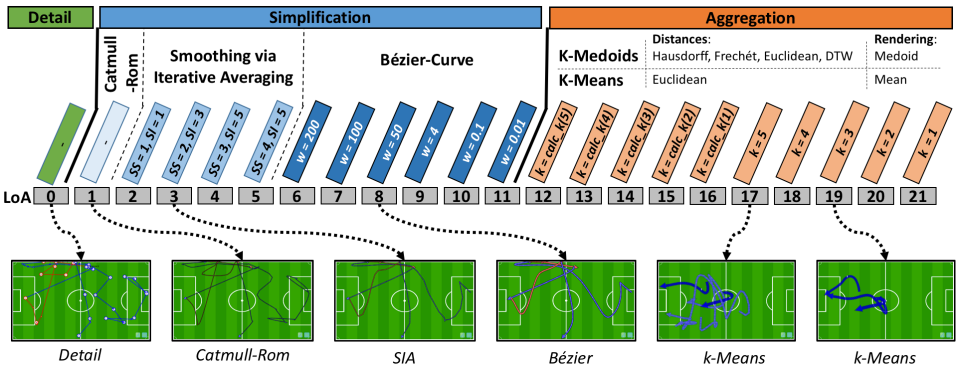



Abb. 2: Die Abstraktionstechniken und deren Parameter werden im oberen Teil gezeigt und in der Mitte auf einen globalen Abstraktionsparameter (LoA) abgebildet. Unten werden verschiedene visuelle Ergebnisse für eine Auswahl der Abstraktionsstufen gezeigt.

Der zweite Teil beschreibt einen **dynamischen und interaktiven Ansatz für die visuelle Abstraktion von Bewegungen im Fußball**, welcher in einer Designstudie [Sa17a] zusammen mit echten Fußballexperten entwickelt und evaluiert wurde. Der Ansatz ist durch das Wissensbildungsmodell inspiriert und zielt darauf ab, die Exploration verschiedener Abstraktionstechniken zu unterstützen. Dies wird durch die interaktive Navigation durch den Raum aller möglichen Abstraktionsstufen und durch flüssige Übergänge zwischen den Teilergebnissen erreicht. Um die Probleme der visuellen Überzeichnungen

und Überladungen von vielen Trajektorien zu lösen, werden die drei Abstraktions-Ebenen **Detail**, **Simplifizierung** und **Aggregation** angewandt (siehe Abbildung 2). In der ersten Detail-Ebene werden nur die wichtigsten Daten gezeichnet. In der Simplifizierungs-Ebene werden die Trajektorien durch verschiedene Techniken vereinfacht und in der Aggregations-Ebene werden mehrere Trajektorien zusammengefasst. Die Einführung eines globalen Abstraktions-Parameters **LoA** (Level of Abstraction) ermöglicht es, die verschieden parametrierten Techniken sinnvoll zu verketten. Das ermöglicht dem Benutzer, den Abstraktionsraum durch eine einzige Interaktion (Mausrad) dynamisch und interaktiv zu explorieren. Der Analyst kann dadurch die zugrundeliegenden Berechnungen explorieren und verstehen und somit die erzielten Ergebnisse besser bewerten. Es wurde außerdem ein Empfehlungssystem hinzugefügt, welches den LoA automatisch durch explizites Benutzerfeedback personalisiert lernen und dann automatisch vorkonfigurieren kann. Das Kapitel enthält viele weitere technische Details, verschiedene qualitative und quantitative Evaluierungen mit Fußballexperten, sowie abschließender Diskussion.

**Kapitel 3 - Analytisches Verhalten und Vertrauensbildung während der Wissensbildung:** Der erste Teil des Kapitels taucht tiefer in die menschlichen Vertrauensbildungs- und Verifikations-Prozesse ein und beschreibt Relationen zu den Systemkomponenten und Analyseproblemen, welche durch Unschärfen in den Berechnungen und durch hohe Komplexität verursacht werden. Ein schärferer Fokus auf die Definition von "validem Wissen" und dessen Relation zu dem ganzen Wissensbildungsprozess führt zu einem neuen konzeptionellen Prozessmodell, welches die Rolle von **Unschärfen, Bewusstsein und Vertrauen** in VA definiert [Sa16b]. Es beschreibt schrittweise alle Systemkomponenten über mögliche Unschärfen und bezieht diese auf die menschliche Wahrnehmung und die Vertrauensbildung in den verschiedenen Stufen des Wissensbildungsprozesses. Das Bewusstsein über die Existenz solcher Unschärfen beeinflusst die menschliche Vertrauensbildung auf verschiedenen Ebenen. Dazu wird eine **Unschärfen – Bewusstsein – Vertrauen**-Klassifikation eingeführt, welche problematische Situationen aufzeigt und den Bezug zu kognitiven Täuschungen herstellt. Danach werden Leitlinien, Beispiele und Herausforderungen für den Umgang mit Unschärfen beschrieben. Auf der Systemseite werden Methoden zur Erfassung, Messung und Darstellung für die interaktive Exploration der Unschärfen erläutert. Auf der menschlichen Seite hingegen wird das Augenmerk auf die Bedienbarkeit, die Unterstützung der Denkprozesse, die automatische Erkennung und Erfassung von Hinweisen auf Probleme und Täuschungen, wie auf das Nachverfolgen und Verifizieren des Analyseprozesses gelegt.

Der zweite Teil des Kapitels stellt eine generalisierte **Notizumgebung mit analytischer Provenienz Komponente** vor, welche an jedes beliebige, bereits bestehende VA System angeschlossen werden kann. Das System implementiert eine neuartige Methodik, um analytische Verhaltensweisen und die Vertrauensbildung während des Analyseprozesses zu erforschen. Es unterstützt das Wissensmanagement, wie die Beweisführung und Hypothesen-Verfeinerung und integriert analytische Provenienz (Herkunft der erzielten Ergebnisse) Funktionalitäten innerhalb des eigentlichen VA Systems und der Notizumgebung. Das System wird mit dem visuellen Abstraktionssystem für Fußballbewegungsdaten integriert und evaluiert, um menschliche Explorations- und Verifikations-Prozesse in einer quantitativen Benutzerstudie zu untersuchen [Sa16a]. Die Ergebnisse enthüllen, dass

Analyse-Strategien und die Vertrauensbildung sehr individuell sind. Es stellte sich heraus, dass vor allem die Selbstbewertung von Vertrauen bei allen Teilnehmern sehr unterschiedlich ist. Bei einigen Teilnehmern war es jedoch möglich, signifikante Korrelationen in den Metriken zu Vertrauen und Interaktion zu finden. Künftig sollte an solchen Methoden weiter geforscht werden, um es zu ermöglichen, verschiedene Benutzercharakteristiken automatisch zu erkennen und die Benutzer entsprechend ihrer Bedürfnisse zu unterstützen.

**Kapitel 4 - Visuell-interaktives maschinelles Lernen:** Dieses Kapitel legt den Fokus auf die Interpretation und die Steuerung komplexer rechnergestützter Methoden und liefert im ersten Teil ein generelles Menschen-zentriertes **visuell-interaktives Modell des maschinellen Lernens** (ML) [Sa17d], welches verschiedene Arten an Benutzer-Feedback beleuchtet und Interaktionen entlang der visuell-interaktiven ML Pipeline aufzählt. Nach einer Beschreibung der verwandten Arbeiten beispielsweise aus der VA, dem interaktiven ML oder dem Menschen-basierten Design wird das konzeptionelle Prozessmodell schrittweise mit Hilfe von Beispielsystemen und Methoden zur Unterstützung des Analyseprozesses beschrieben (siehe Abbildung 3). **Daten** werden für die jeweiligen ML Methoden

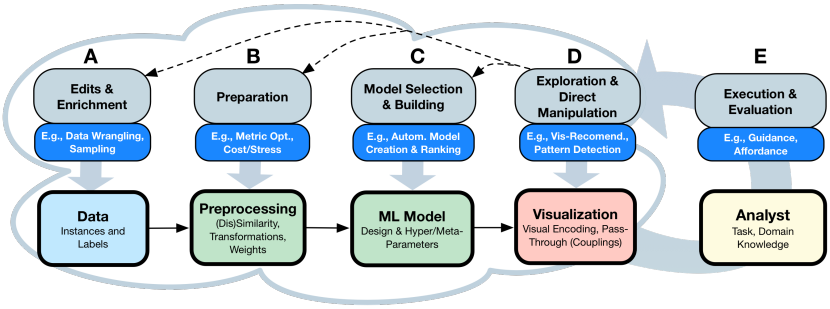



Abb. 3: Eine interaktive VA/ML Pipeline wird links unten gezeigt (*A–D*), angereichert mit mehreren Interaktions-Möglichkeiten (graue Boxen darüber) und exemplarischen automatisierten Methoden diese Interaktionen zu unterstützen (blaue Boxen). Interaktionen lösen Systemreaktionen aus, die vom Analysten beobachtet, interpretiert, validiert und verfeinert werden (*E*).

**vor-verarbeitet**. Im Anschluss wird das **ML Modell** trainiert bzw. erzeugt und dessen Ergebnisse (z.B. Klassifizierungen) können durch eine **Visualisierung** von einem **Analysten** untersucht werden. Die wesentlichen Interaktionen umfassen entsprechend die **Editierung & Anreicherung** von Daten, deren **Vorbereitung** für die ML Komponenten, die **Modell Selektion & Erstellung**, die **Exploration & Direkte Manipulation** der Ergebnisse, sowie der menschlichen **Ausführung & Evaluierung** von Aktionen. Dann wird das Modell auf verschiedene Beispielsysteme angewandt, um deren Funktionalitäten zu sammeln und zu vergleichen. Zusätzlich werden die menschlichen Phasen der Interaktion [No02], also die **Ausführung** und **Evaluierung** von Aktionen, genauer beschrieben. Dieser Teil charakterisiert außerdem weitere spezifische Analyseszenarien, in denen Domänen-Experten Teil der visuell-interaktiven ML Schleife sind. Dazu gehören die **Bestätigende Analyse**, die **Hypothesen-Formung**, die **Konfrontation von ML Resultaten oder Strukturen**, die **Anpassung von ML Pipelines**, **Was Wäre Wenn-Szenarien** oder die **Experten-Verifikation**. Am Ende werden weitere Forschungsmöglichkeiten beschrieben.

Der zweite Teil des Kapitels stellt eine neuartige visuell-interaktive Methode für das explorative Clustering von Zeitserien mit **selbstorganisierenden Karten, automatischer Nutzerführung und analytischer Provenienz (SOMFlow)** [Sa17c] vor. Der Ansatz wird auf die Domäne der linguistischen Intonationsforschung angewandt. Das System wurde über einen längeren Zeitraum in enger Zusammenarbeit mit Sprachwissenschaftlern entwickelt und wurde dann schrittweise verbessert und für die Nutzung in anderen Domänen generalisiert (beispielsweise für die Analyse von Aktienkursen oder Temperaturveränderungen auf der Welt). Das System integriert Visualisierung eng mit interaktivem ML und beinhaltet zusätzlich Notiz-Funktionalitäten mit analytischer Provenienz, um die Wissensbildung zu unterstützen. Es macht sich Qualitäts- und Interessantheits-Maße zu nutze, um den Analysten während der Wissensbildung zu leiten. Dabei unterstützt das SOMFlow System vier Explorations-Aufgaben: Der Analyst kann die Ergebnisse visuell **analysieren** und bei Bedarf **anpassen**. Neue Ergebnisse werden durch das iterative **Aufteilen** der Daten erzeugt. Alle Ergebnisse werden in einen Graphen eingebettet, der es dem Analysten erlaubt, den Analyseprozess zu **reflektieren**.

**Kapitel 6 - Visuell-interaktive Dimensionsreduktion:** Im ersten Teil des Kapitels wird das vorherige Kapitel für das konkrete ML Problem der Dimensionsreduktion (DR) durch eine strukturierte Literaturanalyse spezialisiert [Sa17e]. Unter der Fragestellung "wie genau Analysten mit den aktuellen DR Techniken interagieren" werden sieben allgemeine Szenarien für die interaktive DR enthüllt. Die DR Pipeline, welche Daten über den Feature-Raum auf DR Methoden abbildet, wird durch eine visuell-interaktive Schnittstelle erweitert, über welche die sieben Interaktions-Szenarien realisiert werden. Diese Szenarien umfassen die **Daten Selektion & Erweiterung**, die **Datenmanipulation**, das **Annotieren & Labeln** von Daten, die **Feature Selektion & Vorbereitung**, das **Parameter Tuning**, die **Spezifikation von Einschränkungen**, sowie die **Selektion der DR Art**. Das Prozessmodell wird dann genutzt, um existierende Systeme zu evaluieren, zu vergleichen und weitere Forschungsgebiete abzuleiten.

Abb. 4: Das System wird genutzt um Feature-Gewichtungen, DR- und Clustering-Algorithmen zu verstehen und anzupassen. Die Cluster können in der tabellarischen Darstellung (rechts) interpretiert werden. Diese unterscheiden sich hauptsächlich in den Features "force" und "window".

Im zweiten Teil wird ein System von dem Prozessmodell abgeleitet, das verschiedene **visuell-interaktive Dimensionsreduktions-Techniken** auf die Domäne der polizeilichen Analyse von Kriminalfällen anwendet [Sa17b]. Es ermöglicht den Ermittlern, Kriminalfälle in verschiedene Cluster einzuteilen und diese zu evaluieren. Zudem kann jederzeit die Robustheit der Kriminalitäts-Cluster über verschiedene Algorithmen und Parametrisierungen

getestet werden. Das System ermöglicht Ermittlern, wie beispielsweise Polizeibeamten ähnliche Fälle und Muster zu identifizieren, welche auf hoch-dimensionalen Features in den Daten basieren (siehe Abbildung 4). Außerdem implementiert das System die Erfassung und Visualisierung von Benutzer-Interaktionen. Die Analysten können dabei durch verschiedene Visualisierungstechniken, wie in Scatterplots, Matrizen oder einer tabellarischen Darstellung der Cluster und derer Merkmale, verschiedene algorithmische Varianten explorieren und interpretieren. Dabei können die Analysten zwischen verschiedenen DR Algorithmen (linear, Distanz-basiert, Nachbarschafts-basiert) wechseln und direkt in allen Visualisierungen die Wichtigkeit bestimmter Merkmale anpassen. Zusätzlich wird auf die Ergebnisse ein Clustering im visuellen Raum angewandt. Das Kapitel diskutiert außerdem, wie komplexe Algorithmen und verschiedene Ergebnisse den Domänen-Experten durch Visualisierungen transparent, verständlich und zugänglich gemacht werden können.

## 4 Schlussteil

Das letzte Kapitel fasst die konzeptionellen und methodischen Beiträge mit einer Diskussion um Implikationen und Forschungsmöglichkeiten in einem breiteren Forschungskontext zusammen. Der Umfang der Dissertation und die diskutierten Erweiterungspunk-

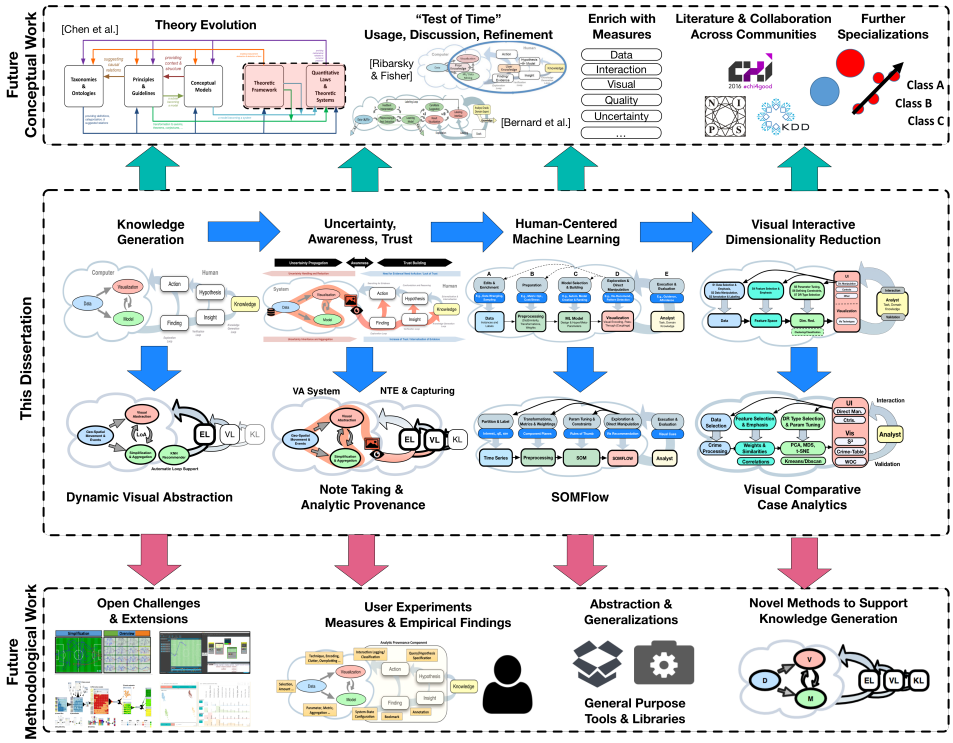



Abb. 5: Zusammenfassung der konzeptionellen und methodischen Beiträge (Mitte) mit Erweiterungspunkten (oben und unten).

te werden in Abbildung 5 dargestellt. Diese Dissertation verdeutlicht, dass es sich bei VA nicht nur um Computer-Anwendungen mit visuellen Daten-Repräsentationen handelt. Es handelt sich vielmehr um einen kollaborativen Prozess, der sich aus menschlichen und maschinellen analytischen Aktivitäten zusammensetzt, mit dem Ziel, Wissen aus den Daten zu erzeugen. Die zentrale Fragestellung: **"Wie eine enge Integration zwischen automatisierter Analyse und visueller Interaktion erreicht werden kann, um die Wissensbildung in VA besser zu unterstützen?"** wurde mit einer Serie aufeinander aufbauender konzeptioneller Prozessmodelle (Abbildung 5 Mitte) adressiert, welche verschiedene Wege oder "Hebel" zur interaktiven Kontrolle identifizieren, um den Menschen (und dessen Expertenwissen) effektiv in die komplexen Analyseprozesse einzubinden. Die konzeptionellen Prozessmodelle wurden genutzt, um spezifische Forschungsgebiete mit dem Ziel den menschlichen Wissensbildungsprozess zu unterstützen zu beleuchten. Dazu gehören Halb-Automatisierung und Empfehlungen, analytisches Verhalten und Vertrauen und die visuelle Interaktion mit maschinellem Lernen. Diese Dissertation liefert exemplarische Lösungen für diese Forschungsgebiete und schlägt weitere Forschungsrichtungen vor (Abbildung 5 Mitte und unten). Der Autor hofft, dass seine Perspektive und der modell-getriebene Ansatz für VA ebenso andere Forscher aller involvierten Forschungsfelder inspirieren und leiten wird, mit dem ultimativen Ziel gemeinsam eine interdisziplinäre theoretische Grundlage für die VA Forschung zu schaffen (Abbildung 5 oben). Diese wird zu neuartigen, eng gekoppelten VA Methoden führen, welche die Analysten adaptiv nach deren Bedürfnissen unterstützen.

## Literaturverzeichnis

[Ch14] Charmaz, Kathy: Constructing grounded theory. Sage, 2014.

[Ch17] Chen, Min; Grinstein, Georges; Johnson, Chris R.; Kennedy, Jessie; Tory, Melanie: Pathways for Theoretical Advances in Visualization. IEEE Computer Graphics and Applications, 2017.

[Ho06] Hornbæk, Kasper: Current practice in measuring usability: Challenges to usability studies and research. International Journal of Man-Machine Studies, 64(2):79–102, 2006.

[Is13] Isenberg, Tobias; Isenberg, Petra; Chen, Jian; Sedlmair, Michael; Möller, Torsten: A systematic review on the practice of evaluating visualization. IEEE Trans. on Visualization and Computer Graphics, 19(12):2818–2827, 2013.

[Ke10] Keim, Daniel A.; Kohlhammer, Jörn; Ellis, Geoffrey P.; Mansmann, Florian: Mastering the Information Age - Solving Problems with Visual Analytics. Eurographics Association, 2010.

[Mi56] Miller, George A: The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological review, 63(2):81, 1956.

[Mi68] Miller, Robert B.: Response time in man-computer conversational transactions. In: American Federation of Information Processing Societies: Proc. of the AFIPS Joint Computer Conf. - Part I. S. 267–277, 1968.

[No02] Norman, D.A.: The Design of Everyday Things. Basic Books, 2002.

[PC05] Pirolli, Peter; Card, Stuart: The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In: Proc. of the Intern. Conf. on Intelligence Analysis. Jgg. 5, S. 2–4, 2005.

[Sa14] Sacha, Dominik; Stoffel, Andreas; Stoffel, Florian; Kwon, Bum Chul; Ellis, Geoffrey P.; Keim, Daniel A.: Knowledge Generation Model for Visual Analytics. IEEE Trans. on Visualization and Computer Graphics, 20(12):1604–1613, 2014.

[Sa16a] Sacha, Dominik; Boesecke, Ina; Fuchs, Johannes; Keim, Daniel A.: Analytic Behavior and Trust Building in Visual Analytics. In (Bertini, Enrico; Elmqvist, Niklas; Wischgoll, Thomas, Hrsg.): EuroVis 2016 - Short Papers. The Eurographics Association, 2016.

[Sa16b] Sacha, Dominik; Senaratne, Hansi; Kwon, Bum Chul; Ellis, Geoffrey P.; Keim, Daniel A.: The Role of Uncertainty, Awareness, and Trust in Visual Analytics. IEEE Trans. on Visualization and Computer Graphics, 22(1):240–249, 2016.

[Sa17a] Sacha, Dominik; Al-Masoudi, Feeras; Stein, Manuel; Schreck, Tobias; Keim, Daniel A.; Andrienko, Gennady; Janetzko, Halldr: Dynamic Visual Abstraction of Soccer Movement. Computer Graphics Forum, 2017.

[Sa17b] Sacha, Dominik; Jentner, Wolfgang; Zhang, Leishi; Stoffel, Florian; Ellis, Geoffrey: Visual Comparative Case Analytics. In (Sedlmair, Michael; Tominski, Christian, Hrsg.): EuroVis Workshop on Visual Analytics (EuroVA). The Eurographics Association, 2017.

[Sa17c] Sacha, Dominik; Kraus, Matthias; Bernard, Jürgen; Behrisch, Michael; Schreck, Tobias; Asano, Yuki; Keim, Daniel A.: SOMFlow: Guided Exploratory Cluster Analysis with Self-Organizing Maps and Analytic Provenance. IEEE Trans. on Visualization and Computer Graphics, 2017.

[Sa17d] Sacha, Dominik; Sedlmair, Michael; Zhang, Leishi; Lee, John A.; Peltonen, Jaakko; Weiskopf, Daniel; North, Stephen C.; Keim, Daniel A.: What you see is what you can change: Human-centered machine learning by interactive visualization. Neurocomputing, 2017.

[Sa17e] Sacha, Dominik; Zhang, Leishi; Sedlmair, Michael; Lee, John Aldo; Peltonen, Jaakko; Weiskopf, Daniel; North, Stephen C.; Keim, Daniel A.: Visual Interaction with Dimensionality Reduction: A Structured Literature Analysis. IEEE Trans. on Visualization and Computer Graphics, 23(1):241–250, 2017.

[SMM12] Sedlmair, Michael; Meyer, Miriah D.; Munzner, Tamara: Design Study Methodology: Reflections from the Trenches and the Stacks. IEEE Trans. on Visualization and Computer Graphics, 18(12):2431–2440, 2012.

**Dominik Sacha** wurde am 31. Januar 1988 in Tübingen geboren. Nach seinem Abitur im Jahre 2007 studierte er Medien- und Kommunikationsinformatik an der Fachhochschule Reutlingen (Bachelor) und führte sein Studium in Informatik (Master) an der HTWG Konstanz fort. Danach promovierte er am Lehrstuhl für Datenanalyse und Visualisierung der Universität Konstanz und erhielt im Jahre 2018 einen Doktortitel. Parallel zum Studium konnte er Erfahrungen in mehreren Unternehmen und Forschungsprojekten sammeln. Seit 2018 arbeitet er bei Siemens Postal, Parcel & Airport Logistics, wobei er sich mit großen Datenmengen und Fragestellungen aus der Logistik beschäftigt.

Steffen Hölldobler et. al. (Hrsg.): Ausgezeichnete Informatikdissertationen 2018,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2019 221

# Petrinetz-Synthese und modale Spezifikationen[1]

Uli Schlachter[2]

**Abstract:** Bei der Petrinetz-Synthese soll ein gegebenes Verhalten durch ein Petrinetz erzeugt werden, was bedeutet, dass der Erreichbarkeitsgraph des Petrinetzes genau das geforderte Verhalten hat. In dieser Arbeit wird die Synthese von Petrinetz-Teilklassen untersucht, beispielsweise schlichten und schlingen-freien Petrinetzen. Unlösbare LTS werden durch Überapproximation behandelt. Außerdem wird die Synthese von modalen Transitionssystemen (MTS), dem modalem $\mu$-Kalkül und einer Teilmenge des $\mu$-Kalküls, die konjunktiver $\nu$-Kalkül heißt, behandelt. Das Synthese-Problem für MTS und den $\nu$-Kalkül ist unentscheidbar, aber durch eine kleine Einschränkung der betrachteten Petrinetze wird dieses Problem sogar für den ausdrucksmächtigeren $\mu$-Kalküls entscheidbar.

## 1 Einleitung

In der Informatik wird viel mit Modellen gearbeitet. Ein Modell ist dabei eine abstrakte Beschreibung eines Systems über das Aussagen gemacht werden sollen. Für verteilte Systeme muss hierbei die Nebenläufigkeit abgebildet werden, also das unabhängige, evtl. zeitgleiche, Agieren von verschiedenen, räumlich getrennten Teilen des Systems. Mit *Petrinetzen* können solche Systeme gut beschrieben werden, da es hier einen intuitiven Begriff von Nebenläufigkeit gibt [BCD02; BD11; GGS13]. Daher werden sie beispielsweise in der Geschäftsprozessmodellierung [Aa16] und in der Biologie [PWM03] eingesetzt.

Das Verhalten eines Petrinetzes wird durch seinen *Erreichbarkeitsgraphen* beschrieben. Dieser Graph enthält alle erreichbaren Zustände des Petrinetzes und die möglichen Übergänge zwischen diesen. Hiermit kann z.B. untersucht werden, ob es im Modell Zustände gibt, in denen das System nicht weiter arbeiten kann, sogenannte Verklemmungen. Jedoch ist es aufwendig zu prüfen, ob Verklemmungen existieren.

Eine Möglichkeit um gewünschte Eigenschaften für ein System, z.B. Verklemmungs-Freiheit, sicherzustellen, ist die *Synthese*. Hierbei ist eine Beschreibung des gewünschten Verhaltens gegeben und ein System soll automatisch gefunden werden. In der Petrinetz-Synthese wird beispielsweise ein prototypischer Erreichbarkeitsgraph, ein sogenanntes beschriftetes Transitionssystem (labelled transition system; LTS) gegeben und ein Petrinetz soll gefunden werden, das dieses Transitionssystem erzeugt. Hierbei wird üblicherweise gefordert, dass jede Beschriftung, die in der Eingabe vorkommt, von genau einer einzigen Transition im Petrinetz erzeugt wird.

[1] Englischer Titel der Dissertation: “Petri Net Synthesis and Modal Specifications” [Sc18]

[2] Carl von Ossietzky Universität Oldenburg, uli.schlachter@uol.de

Durch ein LTS ist das Verhalten eines Systems exakt beschrieben. Allerdings ist es wünschenswert flexiblere Spezifikationen zu haben, in denen noch Variationen möglich sind. Beispielsweise sollte es möglich sein, Verklemmungen zu verbieten, ohne konkrete Lösungen vorzugeben. Daher werden in dieser Dissertation modale Spezifikationen eingesetzt. Hierbei gibt es zwei Modi von Verhalten: Es gibt *gefordertes* Verhalten, dass auf jeden Fall im System möglich sein muss, und es gibt *erlaubtes* Verhalten, dass das System haben kann, aber das auch weggelassen werden darf. Eine modale Spezifikation beschreibt dabei eine Familie von LTS. Bei der Petrinetz-Synthese soll ein Petrinetz gefunden werden, dessen Verhalten einem dieser LTS entspricht bzw. dessen Verhalten die modale Spezifikation *implementiert*. Beispiele für modale Spezifikationen sind modale Transitionssysteme [Kř17; La89] und der modale $\mu$-Kalkül [AN01; Ko83]. Die Petrinetz-Synthese von modalen Spezifikationen wurde in [BBD15] als interessantes offenes Problem genannt und wird in dieser Dissertation untersucht.

Der nächste Abschnitt führt Petrinetz-Synthese formal ein. In Abschnitt 3 und 4 werden Inhalte der Dissertation vorgestellt. Nachdem in Abschnitt 5 modale Spezifikationen eingeführt wurden, geht Abschnitt 6 auf den zweiten Teil der Dissertation ein: Petrinetz-Synthese aus modalen Spezifikationen.

## 2 Petrinetze und Petrinetz-Synthese

Ein Petrinetz ist ein Tupel $N = (P, T, F, M_0)$. $P$ und $T$ sind endliche und disjunkte Mengen an *Stellen* und *Transitionen*. $F\colon ((P \times T) \cup (T \times P)) \to \mathbb{N}$ ist eine *Kantenrelation*, die *Kantengewichte* angibt. Eine *Markierung* ist eine Funktion $P \to \mathbb{N}$, die jeder Stelle eine Anzahl an *Token* zuweist und $M_0$ ist eine ausgezeichnete Markierung, die *Initialmarkierung*.

Ein Beispiel für ein Petrinetz $N = (P, T, F, M_0)$ zeigt der linke Teil von Abbildung 1. Dieses Netz hat die Stellen $P = \{p_0, p_1, p_2, p_3\}$ und Transitionen $T = \{a, b, c\}$. Es gibt eine Kante mit Gewicht 1 von $p_0$ nach $a$, also gilt $F(p_0, a) = 1$. Außerdem gilt $F(a, p_0) = 0$, es gibt also keine Kante von $a$ nach $p_0$. Die Initialmarkierung wird beschrieben durch $M_0(p_0) = M_0(p_2) = 1$ und $M_0(p_1) = M_0(p_3) = 0$.

Eine Transition $t \in T$ ist in einer Markierung $M$ *aktiviert*, wenn jede Stelle mindestens so viele Token enthält wie durch die ausgehenden Kanten gefordert: $\forall p \in P\colon F(p,t) \leq M(p)$. Dies wird als $M[t\rangle$ geschrieben. Eine aktivierte Transition kann *feuern*, wodurch eine neue Markierung $M'$ erreicht wird. Hierfür wird anhand der Kantengewichte die aktuelle Markierung verändert: $\forall p \in P\colon M'(p) = M(p) - F(p,t) + F(t,p)$. Dies wird als $M[t\rangle M'$ geschrieben und diese Schreibweise wird induktiv auf Sequenzen $\sigma \in T^*$ erweitert.

Beispielsweise ist im Petrinetz aus Abbildung 1 in der aktuellen Markierung Transition $a$ aktiviert, da diese nur ein Token von der Stelle $p_0$ benötigt, welche aktuell genau ein Token besitzt. Dieses Token wird beim Feuern von $a$ konsumiert und ein Token auf der Stelle $p_1$ produziert. Die neue Markierung $M'$ erfüllt $M'(p_0) = 0$, $M'(p_1) = 1 = M'(p_2)$ und $M'(p_3) = 0$.

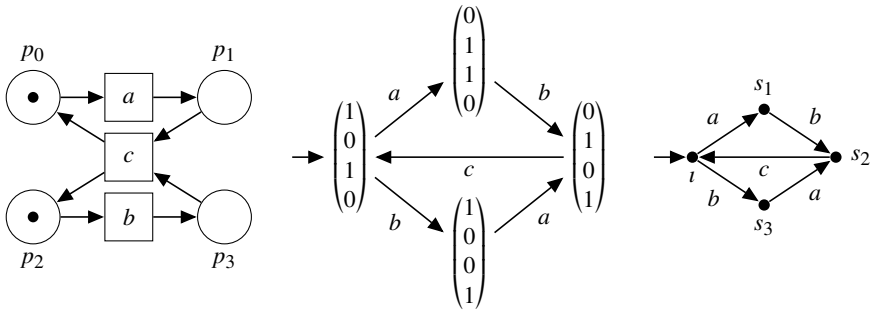



Abb. 1: Beispiel für ein Petrinetz (links), sein Verhalten (Mitte) und ein LTS (rechts)

Das Verhalten eines Prozesses, wie z.B. eines Petrinetzes, kann als LTS dargestellt werden. Ein LTS ist eine Struktur $A = (S, \Sigma, \rightarrow, \iota)$. Hier ist $S$ eine Menge an *Zuständen*, $\Sigma$ ist ein Alphabet mit *Beschriftungen* oder *Ereignissen*, $\rightarrow \subseteq S \times \Sigma \times S$ ist eine *Kantenrelation* und $\iota$ ist der *Initialzustand*.

Ein Beispiel für ein LTS ist im rechten Teil von Abbildung 1 zu sehen. Es hat $\iota$ als Initialzustand. Das Alphabet ist nicht explizit gegeben, aber enthält mindestens $\{a, b, c\}$, da dies die Beschriftungen der Kanten sind. Dieses LTS hat unter anderem die Kanten $(\iota, a, s_1)$ und $(s_1, b, s_2)$. Somit ist z.B. auch die Sequenz $ab$ möglich, die vom Zustand $\iota$ aus zu $s_2$ führt.

Das Verhalten eines Petrinetzes $N$ kann als LTS dargestellt werden, das dessen *Erreichbarkeitsgraph* $\mathrm{RG}(N)$ genannt wird. Hierfür wird die Menge aller erreichbaren Markierungen $\mathfrak{E}(N) = \{M \mid \exists \sigma \in T^* \colon M_0[\sigma\rangle M\}$ als Zustandsmenge verwendet. Eine Kante mit Beschriftung $t \in T$ existiert zwischen zwei Markierungen $M, M' \in \mathfrak{E}(N)$ genau wenn $M[t\rangle M'$ gilt. Der Erreichbarkeitsgraph des Petrinetzes aus dem linken Teil von Abbildung 1 ist in der Mitte der gleichen Abbildung illustriert. Hierfür wird eine Markierung $M$ als Spaltenvektor $M = (M(p_0)\ M(p_1)\ M(p_2)\ M(p_3))$ geschrieben.

Für ein gegebenes Petrinetz ist es recht einfach seinen Erreichbarkeitsgraphen zu berechnen, solange das Petrinetz nur endlich viele verschiedene erreichbare Markierungen hat. Beispielsweise kann durch Tiefen- oder Breitensuche der gesamte Erreichbarkeitsgraph bestimmt werden. *Petrinetz-Synthese* ist die umgekehrte Operation: Ein endliches LTS ist gegeben und es soll ein Petrinetz bestimmt werden, dass dieses LTS löst. Dies bedeutet, dass der Erreichbarkeitsgraph des Petrinetzes isomorph zum gegebenen LTS sein soll, also die gleiche Form haben muss. Von den konkreten Markierungen wird dabei abstrahiert.

Die Grundlage für die Petrinetz-Synthese sind sogenannte *Regionen*. Eine Region eines LTS entspricht einer möglichen Stelle eines Petrinetzes, dass das LTS lösen soll. Die Regionen-Theorie beschreibt, wann das berechnete Petrinetz das gegebene LTS löst und geht auf Arbeiten von Ehrenfeucht und Rozenberg [ER90] zurück. Hierfür gibt es sogenannte

*Separierungsprobleme*, die von Regionen gelöst werden müssen. Davon gibt es zwei Arten: Bei einem *Beschriftungs-Zustands-Separierungsproblem (event-state separation problem; ESSP)* muss in einem Zustand $s \in S$ des LTS die Beschriftung $e \in \Sigma$ verhindert werden, da $s$ keine ausgehende Kante mit Beschriftung $e$ hat. Dies bedeutet, das eine Stelle gesucht wird, die Transition $e$ deaktiviert, wenn der Zustand $s$ eingenommen wird. Ein *Zustands-Separierungsproblem (state separation problem; SSP)* besteht aus zwei verschiedenen Zuständen $s, s' \in S$. Diese beiden Zustände sollen in einer Lösung durch verschiedene Markierungen repräsentiert werden.

## 3 Gezielte Synthese von Petrinetz-Teilklassen

Es existieren schon Algorithmen zur Petrinetz-Synthese. Beispielsweise kann das Problem als lineares und ganzzahliges Ungleichungssystem ausgedrückt werden [BBD15]. In der Dissertation wird dieser Ansatz erweitert, um zusätzliche Eigenschaften für das synthetisierte Petrinetz zu garantieren. Anstatt ein beliebiges Petrinetz zu berechnen, wird auf diese Art auf eine gegebene Teilklasse von Petrinetzen abgezielt. Beispiele für solche Teilklassen sind schlichte Petrinetze, bei denen keine Kantengewichte größer als eins erlaubt sind, schlingenfreie Petrinetze, bei denen zwischen einem Paar aus Transition und Stelle eine Kante in höchstens einer Richtung existieren darf, und $k$-beschränkte Petrinetze, bei denen jede Stelle höchstens $k$ Token in jeder erreichbaren Markierung enthalten darf. Diese Teilklassen werden durch Ungleichungen charakterisiert, beziehungsweise durch prädikatenlogische Formeln, die über ganzen Zahlen interpretiert werden.

Indem diese Formeln zu den Gleichungssystemen hinzugefügt werden, die mittels des bereits bekannten Ansatz aufgestellt werden, wird garantiert, dass die berechneten Petrinetze zur gewünschten Teilklasse gehören. Außerdem lassen sich Teilklassen kombinieren. Beispielsweise werden schlichte und $k$-beschränkte Petrinetze synthetisiert, indem die Formeln für beide Teilklassen mit in die Gleichungssysteme aufgenommen werden.

In der Arbeit werden eine Reihe von bekannten Teilklassen von Petrinetzen identifiziert, die auf diese Weise ausgedrückt werden können. Es werden aber auch Beispiele für Teilklassen gegeben, die nicht mit diesem Ansatz umsetzbar sind.

## 4 Kleinste Lösbare Überapproximation

Nicht jedes LTS kann durch ein Petrinetz gelöst werden. Außerdem sind die meisten Teilklassen ausdrucksschwächer als Petrinetze im Allgemeinen, wodurch noch weniger LTS synthetisiert werden können. Der Algorithmus erkennt diese Situation daran, dass ein Separierungsproblem bzw. das hierfür aufgestellte Gleichungssystem unlösbar ist. In der Dissertation wird ein Algorithmus vorgestellt, der diesen Fall behandelt. Anstatt als Ergebnis „unlösbar“ auszugeben, approximiert dieser Algorithmus das gegebene LTS.

Das Ergebnis des Approximierungs-Algorithmus ist eine kleinste Überapproximation der Eingabe. Hierfür wird eine Partialordnung für LTS namens LTS-Homomorphismus definiert. Diese Partialordnung hat Ähnlichkeit zur bekannten Simulations-Partialordnung. Anhand dieser Ordnung kann das Ergebnis des Algorithmus formal definiert werden: Alle LTS, die größer als das Eingabe-LTS $A$ sind, sind Überapproximationen desselben. Nur manche dieser Überapproximationen sind Petrinetz-lösbar. Der Algorithmus berechnet die kleinste dieser Petrinetz-lösbaren Überapproximationen, welche Approx($A$) genannt wird. Die Eindeutigkeit von Approx($A$) wird in der Arbeit gezeigt.

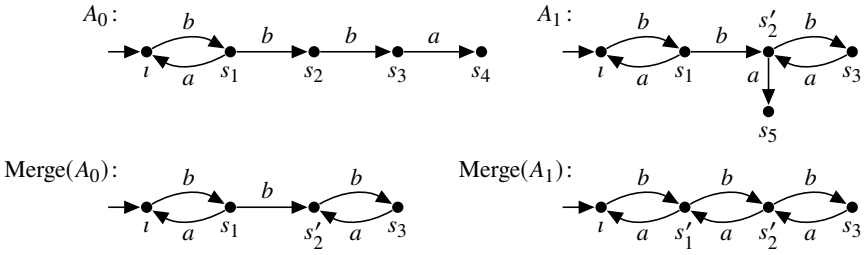



Abb. 2: Ein Beispiel des Approximierungs-Algorithmus

Ein Beispiel für den Algorithmus wird in Abbildung 2 gezeigt. Das LTS $A_0$ soll approximiert werden. Das LTS $A_1$ ist ein Zwischenergebnis. Der Algorithmus ist ein Fixpunkt-Algorithmus, bei dem zwei Schritte immer wiederholt werden, bis keine weiteren Änderungen mehr notwendig sind.

Bei Merge($A$) werden unlösbare SSP-Instanzen behandelt, also Zustände, denen zwingend dieselbe Markierung zugeordnet werden muss. Hierzu werden die Zustände zu einem einzelnen Zustand zusammengefasst. Im Beispiel wurden im LTS Merge($A_0$) die Zustände $s_2$ und $s_4$ zu $s'_2$ zusammengefasst, da die zugehörige SSP-Instanz unlösbar ist.

Im zweiten Schritt wird Merge($A_0$) zu einem LTS $A_1$ expandiert, indem unlösbare ESSP-Instanzen eliminiert werden. Im Beispiel kann im Zustand $s'_2$ die Beschriftung $a$ nicht verhindert werden und wird daher zum LTS hinzugefügt. Dies geschieht, indem ein neuer Zustand $s_5$ eingeführt wird und als Ziel der aus $s'_2$ ausgehenden Kante mit Beschriftung $a$ verwendet wird.

Nun wird der erste Schritt mit $A_1$ wiederholt und Merge($A_1$) bestimmt. Im abgebildeten Beispiel werden hierbei die Zustände $s_1$ und $s_5$ zu $s'_1$ vereinigt. Anschließend gibt es keine weiteren unlösbaren Separierungsprobleme mehr. Das resultierende LTS ist Petrinetz-lösbar und erlaubt u.a. jegliches Verhalten, das schon in $A_0$ möglich war.

Dieser Algorithmus ist auf einige Petrinetz-Teilklassen anwendbar, während für andere Teilklassen die kleinste Überapproximation ein unendlich großes LTS sein kann und dann daher nicht berechnet werden kann.

Die vorgestellte Überapproximation ist minimal nach einer in der Dissertation definierten

Partialordnung. Anhand eines Beispieles wird gezeigt, dass der Algorithmus nicht die kleinste Überapproximation laut Sprachinklusion liefert. Hiermit ist gemeint, dass zu einem gegebenen LTS $A$ die Sprache $L(A)$ approximiert werden soll. Minimalität bedeutet hierbei, dass das berechnete Petrinetz $N$ eine kleinere Sprache hat als jede andere Überapproximation von $A$, also $\forall N' \colon L(A) \subseteq L(N') \Rightarrow L(N) \subseteq L(N')$ erfüllt. Es wird auch gezeigt, wie die kleinste Überapproximation gemäß Sprachinklusion bestimmt werden kann.

## 5 Modale Spezifikationen

Ein LTS spezifiziert das exakte Verhalten eines Systems, ohne dass noch Variationsmöglichkeiten bestehen. Modale Spezifikationen erlauben eine weniger exakte Spezifikation. Eine modale Spezifikation beschreibt eine Menge von LTS. Jedes dieser LTS implementiert die Spezifikation.

In der Dissertation werden drei Arten von modalen Spezifikationen betrachtet: Modale Transitionssysteme (MTS) [La89], der modale $\mu$-Kalkül [AN01; Ko83] und eine syntaktische Teilmenge des $\mu$-Kalküls, die konjunktiver $\nu$-Kalkül heißt. Im folgenden wird der $\mu$-Kalkül kurz eingeführt.

Der $\mu$-Kalkül besteht aus Formeln, die wie folgt aufgebaut sind:

$$\beta ::= \texttt{true} \mid \texttt{false} \mid \beta_1 \wedge \beta_2 \mid \beta_1 \vee \beta_2 \mid \neg\beta_1 \mid \langle a \rangle \beta_1 \mid [a]\beta_1 \mid \nu X.\beta_1 \mid \mu X.\beta_1 \mid X$$

Der $\mu$-Kalkül umfasst die Aussagenlogik, also die Konstanten `true` und `false` und die Konnektoren Konjunktion $\wedge$, Disjunktion $\vee$ und Negation $\neg$. Formeln werden in Zuständen eines LTS interpretiert. Es gibt eine existentielle Modalität $\langle a \rangle \beta$ und eine universelle Modalität $[a]\beta$. Die existentielle Modalität $\langle a \rangle \beta$ fordert, dass der aktuelle Zustand eine ausgehende Kante hat, die mit $a$ beschriftet ist und zu einem Zustand führt, der die Formel $\beta$ erfüllt. Die universelle Modalität macht die gleiche Aussage für alle ausgehenden Kanten mit Beschriftung $a$. Da Petrinetze deterministisch sind, gibt es in einem Erreichbarkeitsgraphen höchstens eine ausgehende Kante für jede Beschriftung. Somit bedeutet $\langle a \rangle \beta$ in diesem Kontext, dass es eine $a$-Kante geben muss. Im Gegensatz dazu fordert $[a]\beta$ diese Kante nicht, sondern macht nur darüber Aussagen, was gelten muss, falls diese $a$-Kante vorhanden ist. Die Formel $[a]\texttt{false}$ sagt z.B. aus, dass es keine ausgehende Kante mit Beschriftung $a$ geben darf, da kein Zustand die Formel `false` erfüllen kann.

Weiterhin gibt es Variablen $X$, den größten Fixpunkt $\nu$ und den kleinsten Fixpunkt $\mu$. Mit diesen kann unendliches Verhalten beschrieben werden. Beispielsweise bedeutet $\langle a \rangle X$, dass nach einer Kante mit Beschriftung $a$ die Variable $X$ erfüllt werden muss. Diese freie Variable kann durch einen Fixpunkt-Operator gebunden werden. So bedeutet $\nu X.\langle a \rangle X$, dass $a$ unendlich oft hintereinander möglich ist. In einem endlichen System fordert dies also eine Schleife.

Der Unterschied zwischen den beiden Fixpunkt-Operatoren ist, dass der größte Fixpunkt $\nu$ unendlich oft durchlaufen werden darf, während der kleinste Fixpunkt $\mu$ nur endliche

Rekursion erlaubt. Beispielsweise bedeutet $\mu X.\langle a\rangle X \vee \langle b\rangle \texttt{true}$, dass nach endlich vielen $a$-Kanten, eine ausgehende Kante mit Beschriftung $b$ existieren muss. Im Gegensatz hierzu würde $\nu X.\langle a\rangle X \vee \langle b\rangle \texttt{true}$ auch durch unendlich viele Wiederholungen der Beschriftung $a$ erfüllt werden.

Ein LTS $A$ *erfüllt* bzw. *implementiert* eine Formel $\beta$, geschrieben $A \models \beta$, falls der Initialzustand des LTS die Formel $\beta$ erfüllt, wie gerade skizziert wurde.

Der $\mu$-Kalkül ist ein sehr ausdrucksmächtiger Formalismus. Er umfasst beispielsweise die bekannten Logiken LTL, CTL, and CTL* [CGR11], als auch modale Transitionssysteme. Letzteres wird in der Dissertation gezeigt, indem ein syntaktisches Fragment des $\mu$-Kalküls definiert wird, die konjunktiver $\nu$-Kalkül heißt, und gezeigt wird, dass dieses Fragment im Wesentlichen äquivalent zu modalen Transitionssystemen ist.

## 6 Petrinetz-Realisierungen Modaler Spezifikationen

Eine modale Spezifikation definiert eine Menge von LTS, die es implementieren. Ein Petrinetz erzeugt einen Erreichbarkeitsgraphen, welches ein LTS ist. Diese beiden bereits existierenden Begriffe werden in der Arbeit zu einem neuen Begriff verbunden: Ein Petrinetz *realisiert* eine modale Spezifikation, falls sein Erreichbarkeitsgraph diese Spezifikation implementiert.

Nun kann das *Realisierungsproblem* formuliert werden: Gegeben eine modale Spezifikation, gibt es eine Petrinetz-Realisierung? Verschiedene Varianten dieses Problems können betrachtet werden. In der Arbeit werden dafür drei Arten von modalen Spezifikationen vorgestellt. Es kann außerdem gefordert werden, dass das Petrinetz zu einer gegebenen Teilklasse gehört, wie beispielsweise schlingen-freie und $k$-beschränkte Petrinetze.

In der Dissertation wird für das Realisierungsproblem noch gefordert, dass das Petrinetz nur endlich viele erreichbare Markierungen hat. Für den unendlichen Fall mit schlingen-freien Petrinetze wurde bereits die Unentscheidbarkeit gezeigt [Fe05].

Es wird in der Dissertation gezeigt, dass das Realisierungsproblem auch für endliche Erreichbarkeitsgraphen unentscheidbar ist. Dieses Ergebnis kann auch für schlingen-freie Petrinetze und einige andere Teilklassen gezeigt werden. Es gilt auch für sämtliche betrachteten Arten von modalen Spezifikationen, also nicht nur für den sehr ausdrücksmächtigen $\mu$-Kalkül, sondern auch für modale Transitionssysteme und den konjunktiven $\nu$-Kalkül, die nicht einmal Disjunktionen ausdrücken können und somit sehr eingeschränkt sind. Der Beweis basiert auf einer Simulation von Zwei-Zähler-Maschinen.

Um Entscheidbarkeit zu erreichen, werden die betrachteten Petrinetze eingeschränkt, und zwar auf die Klasse der $k$-beschränkten Petrinetze, wobei $k$ eine weitere Eingabe an den Algorithmus ist, also a priori gegeben ist. Durch diese Einschränkung gibt es nur noch endlich viele mögliche Petrinetze, die in Frage kommen können: Im Petrinetz können nur

Zahlen bis $k$ auftauchen und das Alphabet $\Sigma$, also die Transitionsmenge des Petrinetzes, ist durch die Spezifikation bereits festgelegt. Diese Petrinetze könnten theoretisch alle darauf geprüft werden, ob sie die Spezifikation implementieren, jedoch wäre der Aufwand dieses Ansatzes riesig: Eine Abschätzung der Anzahl in Frage kommender Petrinetze ergibt $2^{(k+1)^{1+2|\Sigma|}}$, wobei für jedes dieser Petrinetze noch der Erreichbarkeitsgraph berechnet und geprüft werden muss, ob dieser die Spezifikation implementiert.

Daher wird in der Arbeit ein direkter Ansatz verfolgt. Es soll ein LTS konstruiert werden, das Petrinetz-lösbar ist und die Spezifikation implementiert. Falls ein Zustand eine Formel $\langle a \rangle \beta$ erfüllen muss, aber noch keine ausgehende Kante mit der Beschriftung $a$ hat, dann wird eine solche Kante zu einem neuen Zustand hinzugefügt. Die anderen Operatoren des $\mu$-Kalküls können auf ähnliche Weise behandelt werden. Hierfür wird auf einen Algorithmus von Stirling und Walker [SW89; SW91] zur lokalen Modellprüfung zurückgegriffen. Normalerweise schlägt dieser Algorithmus fehl, wenn ein Zustand $\langle a \rangle \beta$ erfüllen soll, aber keine passende ausgehende Kante hat. Dies ist jedoch genau die Information, die zum Erweitern des LTS benötigt wird.

Ein auf diese Weise konstruiertes LTS ist nicht notwendigerweise Petrinetz-lösbar. Um dies sicherzustellen, wird wieder die kleinste Überapproximation eingesetzt. Der Algorithmus besteht darin, die beiden Schritte, Erweiterung des LTS und Überapproximation, zu wiederholen bis eine Petrinetz-lösbare Implementierung gefunden wurde oder die Erweiterung fehlschlägt. Ein Fehlschlag passiert beispielsweise, wenn ein Zustand eine Kante mit der Beschriftung $a$ hat, aber $[a]$`false` erfüllen soll.

Da nur die Überapproximation spezifisch für Petrinetze ist, unterstützt dieser Algorithmus alle Teilklassen von Petrinetzen, für die eine Überapproximation möglich ist.

Der vorgestellte Algorithmus und seine Implementierung werden am Ende der Dissertation für eine Fallstudie eingesetzt. In dieser Fallstudie wird Dijkstras bekanntes Philosophenproblem als LTS und als modale Spezifikation umgesetzt.

## 7 Fazit

Die hier beschriebene Dissertation besteht aus zwei Teilen. Im ersten Teil wird Petrinetz-Synthese aus LTS untersucht. Nach der Einführung der Grundlagen und existierender Algorithmen wird einer dieser Algorithmen erweitert, um auf Teilklassen von Petrinetzen zielen zu können. Dies bedeutet, dass weitere Anforderungen an das Petrinetz gestellt werden, beispielsweise können Kantengewichte verboten werden. Als nächstes wird ein Ansatz zur Approximation vorgestellt. Nicht jedes LTS kann in ein Petrinetz überführt werden. In einem solchen Fall kann das LTS strukturell verändert werden, um es Petrinetz-lösbar zu machen. Hierzu wird ein Algorithmus zur kleinsten Überapproximation eines gegeben LTS vorgestellt. Dieser Algorithmus funktioniert auch für manche der zuvor untersuchten

Teilklassen, während für andere Teilklassen die kleinste Überapproximation unendlich groß werden kann.

Im zweiten Teil der Dissertation geht es um Petrinetz-Synthese aus modalen Spezifikationen. Hierzu werden MTS, der $\mu$-Kalkül und der $\nu$-Kalkül zunächst eingeführt und eine Beziehung zwischen ihnen hergestellt. Dann wird gezeigt, dass die Petrinetz-Synthese aus einer sehr ausdrucksschwachen Spezifikationssprache, den MTS, unentscheidbar ist. Als nächstes werden $k$-beschränkte Petrinetze betrachtet, für die das Synthese-Problem wieder entscheidbar ist, und ein Synthese-Algorithmus wird vorgestellt.

Alle Algorithmen, die in der Dissertation entwickelt wurden, wurden in einem Open-Source-Werkzeug implementiert. Diese Implementierung wurde für eine Fallstudie herangezogen, in der das bekannte Philosophenproblem mit modalen Spezifikationen modelliert und anschließend in ein Petrinetz synthetisiert wurde. Hierbei zeigte sich, dass die unabhängigen Teile eines verteilten Systems mit modalen Spezifikationen gut abgebildet werden können.

Eine der offenen Fragen ist die Komplexität der untersuchten Probleme und der vorgestellten Algorithmen. Hierzu gibt es bereits erste Ansätze, aber noch keine abschließenden Antworten. Eine andere Richtung für weitere Forschung wäre es, den Ansatz zur Realisation von $\mu$-Kalkül-Formeln auf andere Spezifikationssprachen zu erweitern. Beispielsweise ist die monadische Logik zweiter Ordnung (monadic second order logic) ausdrucksmächtiger als der $\mu$-Kalkül und kann beispielsweise ausdrücken, dass zwei Pfade zum gleichen Zustand führen müssen. Eine solche Anforderung könnte durch Zusammenfassen von Zuständen behandelt werden. Somit sollte auch die Synthese entsprechender Formeln möglich sein. Allerdings fehlt noch ein Ansatz, um ein gegebenes LTS um benötigtes Verhalten zu erweitern. Die Grundidee für einen Algorithmus existiert jedoch schon.

## Literatur

[Aa16] van der Aalst, W. M. P.: Process Mining - Data Science in Action, Second Edition. Springer, 2016.

[AN01] Arnold, A.; Niwiński, D.: Rudiments of $\mu$-calculus. North Holland, 2001.

[BBD15] Badouel, E.; Bernardinello, L.; Darondeau, P.: Petri Net Synthesis. Springer, 2015.

[BCD02] Badouel, E.; Caillaud, B.; Darondeau, P.: Distributing Finite Automata Through Petri Net Synthesis. Formal Aspects of Computing 13/6, S. 447–470, 2002.

[BD11] Best, E.; Darondeau, P.: Petri Net Distributability. In (Clarke, E. M.; Virbitskaite, I.; Voronkov, A., Hrsg.): PSI 2011. Bd. 7162. LNCS, Springer, S. 1–18, 2011.

[CGR11] Cranen, S.; Groote, J. F.; Reniers, M. A.: A linear translation from CTL* to the first-order modal $\mu$-calculus. Theoretical Computer Science 412/28, S. 3129–3139, 2011.

[ER90] Ehrenfeucht, A.; Rozenberg, G.: Partial (Set) 2-Structures. Part I: Basic Notions and the Representation Problem and Part II: State Spaces of Concurrent Systems. Acta Inf. 27/4, S. 315–368, 1990.

[Fe05] Feuillade, G.: Spécification logique de réseaux de Petri, Diss., Université de Rennes I, 2005, URL: http://www.irisa.fr/s4/download/papers/Feuillade-these2005.pdf.

[GGS13] van Glabbeek, R. J.; Goltz, U.; Schicke-Uffmann, J.: On Characterising Distributability. Logical Methods in Computer Science 9/3, 2013.

[Ko83] Kozen, D.: Results on the Propositional mu-Calculus. Theoretical Computer Science 27/3, S. 333–354, 1983.

[Kř17] Křetínský, J.: 30 Years of Modal Transition Systems: Survey of Extensions and Analysis. In (Aceto, L.; Bacci, G.; Bacci, G.; Ingólfsdóttir, A.; Legay, A.; Mardare, R., Hrsg.): Models, Algorithms, Logics and Tools. Bd. 10460. LNCS, Springer, S. 36–74, 2017.

[La89] Larsen, K.: Modal Specifications. In (Sifakis, J., Hrsg.): AVMFSS. Bd. 407. LNCS, Springer, S. 232–246, 1989.

[PWM03] Pinney, J.; Westhead, D.; McConkey, G.: Petri Net representations in systems biology. Biochemical Society Transactions 31/6, S. 1513–1515, 2003, ISSN: 0300-5127.

[Sc18] Schlachter, U.: Petri Net Synthesis and Modal Specifications, Diss., Carl von Ossietzky Universität Oldenburg, 2018, URL: http://nbn-resolving.de/urn:nbn:de:gbv:715-oops-38362.

[SW89] Stirling, C.; Walker, D.: Local Model Checking in the Modal Mu-Calculus. In (Diaz, J.; Orejas, F., Hrsg.): TAPSOFT'89 (CAAP'89). Bd. 351. LNCS, Springer, S. 369–383, 1989.

[SW91] Stirling, C.; Walker, D.: Local Model Checking in the Modal mu-Calculus. Theoretical Computer Science 89/1, S. 161–177, 1991.

**Uli Schlachter** wurde am 12. Dezember 1989 in Oldenburg geboren. Nach dem Abitur an der Kooperativen Gesamtschule Rastede im Jahr 2009 begann er ein Bachelorstudium der Informatik an der Carl von Ossietzky Universität Oldenburg. An dieses schloss er ein Masterstudium an, welches er 2014 mit Auszeichnung abschloss. Anschließend arbeitete er dort in der theoretischen Informatik als wissenschaftlicher Mitarbeiter in der Abteilung für parallele Systeme bei Prof. Dr. Eike Best, wo er im November 2018 promovierte.

Steffen Hölldobler et. al. (Hrsg.): Ausgezeichnete Informatikdissertationen 2018,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2019 231

# Lernen von Repräsentationen für Atomistische Systeme mit Tiefen Neuronalen Netzen[1]

Kristof T. Schütt[2]

**Abstract:** Tiefes Lernen hat in den vergangenen Jahren zu Durchbrüchen beim Lernen von Repräsentationen strukturierter Daten in zahlreichen Anwendungen wie der Bilderkennung oder der maschinellen Übersetzung geführt. Allerdings ist die Funktionsweise von neuronalen Netzen schwer nachzuvollziehen, während ihr Training riesige Datensätze erfordert. In der zusammengefassten Dissertation wird die Nutzung anwendungsspezifischen Vorwissens beim Lernen von Repräsentationen sowohl zur Verbesserung der Dateneffizienz als auch der Interpretierbarkeit am Beispiel von atomistischen Systemen demonstriert. So wie tiefes Lernen Repräsentationen für strukturierte Daten wie Bilder, Texte oder Audio lernen kann, sind die hier entwickelten neuronalen Netze in der Lage atomistische Repräsentationen direkt aus den Positionen und Elementen der Atome zu lernen. Damit ermöglichen sie genaue Vorhersagen von chemischen Eigenschaften sowie atomistische Simulationen, welche mit konventionellen *ab initio* Methoden aufgrund der benötigten Rechenzeit nicht durchführbar wären. Eine Analyse der trainierten Modelle zeigt, dass lokale Repräsentationen gelernt wurden, die mit chemischem Grundwissen übereinstimmen, was den Gewinn von wissenschaftlichen Erkenntnissen aus dem Lernmodell ermöglicht.

## 1 Einführung

Die Fortschritte im tiefen Lernen, u.a. in der Bilderkennung oder der Verarbeitung natürlicher Sprache [KSH12, SVL14, Vi15, Mn15], wurde durch Ende-zu-Ende-Lernen von Repräsentationen strukturierter Daten ermöglicht [LBH15, Sc15]. Dazu werden große Datenmengen benötigt, beispielsweise der ImageNet-Datensatz [De09] in der Bilderkennung, welcher mehr als 14 Millionen annotierte Bilder enthält. In vielen Anwendungen stehen jedoch nur wesentlich kleinere Datenmengen zur Verfügung. Um dort ebenfalls Repräsentationen lernen zu können, ist es zwingend notwendig anwendungsspezifisches *a priori* Wissen in die Entwicklung von neuronalen Netzen einfließen zu lassen. Weiterhin stellen tiefe Netze hochkomplexe, statistische Modelle dar, deren Funktionsweise für den Nutzer schwer nachvollziehbar ist. Während eine Möglichkeit zur Erklärung neuronaler Netze Methoden sind, die den Eingabedimensionen jeweils eine Relevanz für die getroffene Vorhersage zuweisen [Mo17, Ki18], ist es oft besser während des Entwurfs der Architektur eines neuronalen Netzwerkes bereits auf Interpretierbarkeit zu achten. In der zusammengefassten Dissertation wird die Nutzung anwendungsspezifischen Vorwissens beim Lernen von Repräsentationen sowohl zur Verbesserung der Dateneffizienz als auch der Gewinn wissenschaftlicher Erkenntnisse am Beispiel von atomistischen Systemen demonstriert.

[1] Englischer Titel der Dissertation: „Learning Representations of Atomistic Systems with Deep Neural Networks“

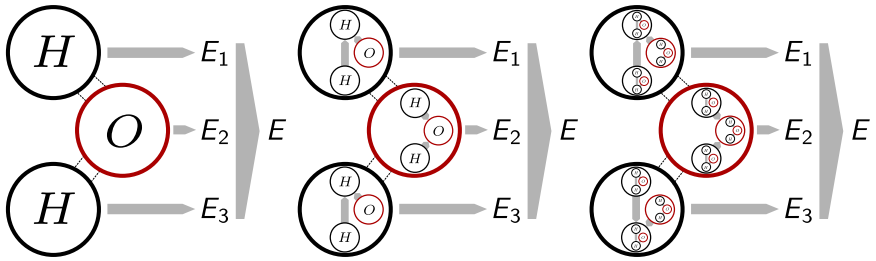
[2] Technische Universität Berlin, kristof.schuett@tu-berlin.de

Die Entdeckung und Erforschung neuartiger Moleküle und Materialien ist von entscheidender Bedeutung für eine große Bandbreite von Anwendungen. Diese finden sich z.B. in der Pharmazie und bei der Entwicklung effizienter Batterien oder Solarzellen. Um die chemischen Eigenschaften dieser Stoffe bestimmen zu können, werden quantenchemische Berechnungen benötigt. Deren Laufzeit skaliert mit der Anzahl der Elektronen des Systems, je nach Grad der Approximation, bis zu $O(n!)$ für die vollständige, numerische Lösung der Schrödinger Gleichung. Während gängige Näherungen wie Dichtefunktionaltheorie (DFT) mit $O(n^4)$ skalieren, kann dies bereits eine systematische Erkundung des chemischen Raumes undurchführbar machen. Das liegt u.a. an den vielen Möglichkeiten mit denen Molekülen oder Materialien aus Atomen gebildet werden können, wodurch der chemische Raum zu groß für eine erschöpfende Suche wird. Darüber hinaus können auch für Moleküldynamiksimulationen (MD-Simulationen) eines einzelnen Systems bereits mehrere Millionen Quantenchemierechnungen benötigt werden. Um zeitintensive Quantensimulationen zu umgehen, wurden in den vergangenen Jahren diverse Methoden des maschinellen Lernens (ML) eingesetzt, welche mithilfe eines Datensatzes von Referenzrechnungen den funktionalen Zusammenhang zwischen chemischer Zusammensetzung und Struktur eines atomistischen Systems sowie den quantenchemischen Eigenschaften erschließen [Ru12, Fa17, Ch17].

Ein atomistisches System kann allgemein als eine Menge von Atomen $S = \{(Z_i, \mathbf{r}_i) | i \in [1, n_{\text{atoms}}]\}$ geschrieben werden, wobei $Z_i \in \mathbb{N}$ das chemische Element und $\mathbf{r}_i \in \mathbb{R}^3$ die Position von Atom $i$ darstellen. Regressionsmodelle erfordern allerdings im Allgemeinen eine Einbettung in einen Vektorraum. Die Schwierigkeit besteht hier u.a. darin, dass die Systeme im Datensatz aus verschieden vielen Atomen bestehen können. Darüber hinaus soll die Repräsentation invariant gegenüber Rotation und Translation des Systems sowie der Ordnung der Atome sein, da diese auch die chemischen Eigenschaften unverändert lassen [Li15]. Aus diesen Gründen wurden in den letzten Jahren ein große Anzahl von Repräsentationen für Moleküle und Materialien entwickelt [BP07, BKC13, Sc14]. Dies hat allerdings den Nachteil, dass die Repräsentationen nicht an den vorhandenen Datensatz angepasst werden können während verschiedene chemische Eigenschaften unter Umständen auch verschiedene Repräsentationen erfordern um eine genaue Vorhersage zu ermöglichen. Dies gilt insbesondere für die Kodierung der chemischen Elemente, wobei oft auf Heuristiken zurückgegriffen wird [Br17, Fa18]. Tiefes Lernen kann hingegen während des Training die zu lernende Repräsentation bzgl. der Modalitäten der gegebenen Daten adaptieren. In Folgendem werden die in der Dissertation entwickelten neuronalen Netze zum Lernen von atomistischen Repräsentationen vorgestellt sowie auf deren Anwendung und Interpretierbarkeit eingegangen.

## 2 Atomistische neuronale Netze

Ein wichtiger Ansatz zur Repräsentation atomistischer Systemen ist die Partitionierung der Energie in Beiträge der individuellen Atome, welche den Einfluss der lokalen, chemischen Umgebungen einbeziehen. Auch wenn eine solche Partitionierung nicht eindeutig ist, so spiegeln diese latenten Variable doch den Einfluss lokaler Strukturen auf die Vorhersage wider. In der Dissertation wird dazu jedem Atom $i$ ein Merkmalsvektor $\mathbf{x}_i \in \mathbb{R}^F$ zugewie-




(a) T=0 (b) T=1 (c) T=2

Abb. 1: Illustration der iterativen Konstruktion von atomaren Repräsentationen im neuronalen Netz am Beispiel eines Wassermoleküls mit $T$ Korrekturschritten. Mit zunehmendem $T$ werden mehr Informationen über die Umgebung in den den atomaren Repräsentationen akkumuliert.

sen. Dieser entspricht zunächst einer vorläufigen atomaren Repräsentation $\mathbf{x}_i^{(0)} = A_{[Z_i,:]} \in \mathbb{R}^F$, welche die chemischen Elemente der Atome charakterisiert. Die Elementeinbettungen $A$ werden dabei während des Trainingsprozesses adaptiert. Davon ausgehend müssen nun den initialen Repräsentationen Informationen über die Atompositionen hinzugefügt werden. Dies geschieht mittels additiver Interaktionskorrekturen

$$\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t)} + \sum_{j \neq i} \mathbf{v}^{(t)}(\mathbf{x}_j^{(t)}, d_{ij}), \tag{1}$$

wobei die Interaktionsnetze $\mathbf{v}^{(t)}$ die Wechselwirkungen zwischen den Atomen modellieren. Abb. 1 illustriert wie auf diese Weise iterativ zunehmend komplexere Beschreibungen der atomaren Umgebungen konstruiert werden.

In der Dissertation werden zwei Ansätze zur Modellierung der Interaktionen $\mathbf{v}^{(t)}$ eingeführt. Eine Möglichkeit besteht in der Verbindung von der Repräsentation $\mathbf{x}_j$ eines benachbarten Atoms $j$ sowie der Distanz $d_{ij}$ zum Zentralatom $i$ durch ein faktorisiertes Tensorlayer

$$\mathbf{v}_{ij} = \tanh\left[W^{xf}\left((W^{fx}\mathbf{x}_j + \mathbf{b}^{f_1}) \circ (W^{fd}\hat{\mathbf{d}}_{\mathbf{ij}} + \mathbf{b}^{f_2})\right)\right], \tag{2}$$

darzustellen. Dabei werden die Eingaben zunächst mit den Gewichtsmatrizen $W^{fx}, W^{fd}$ in einen Faktorraum abgebildet werden, woraufhin ihr Hadamard-Produkt zurück in den Repräsentationsraum projiziert wird.

Schließlich können aus den Atomrepräsentationen die gewünschten, chemischen Eigenschaften vorhergesagt werden. Beispielsweise kann die Energie eines Systems als Summe von atomaren Energiebeiträgen

$$\hat{E}(S) = \sum_{i=1}^{n_{\text{atoms}}} \text{NN}_E(\mathbf{x}_i). \tag{3}$$

dargestellt werden, wobei $\text{NN}_E$ den Energiebeitrag eines Atoms in seiner Umgebung darstellt. Das gesamte Modell kann nun mit stochastischem Gradientenabstieg trainiert werden, wobei die Einbettungen $A$ sowie die Parameter des Interaktionsnetzes $\mathbf{v}^{(t)}$ sowie des

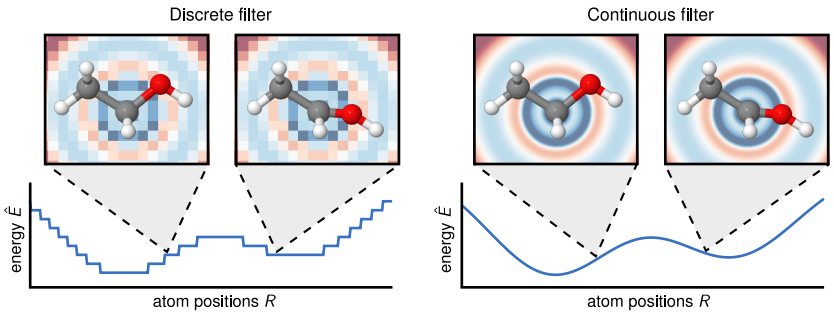



Abb. 2: Einfluss von diskretem vs. kontinuierlichem Filter im Faltungslayer: Unstetigkeiten im Filter übertragen sich auf die Energievorhersage, was der physikalischen Realität widerspricht.

Ausgabenetzes $\mathrm{NN}_E$ optimiert werden. Ein solches atomistisches, neuronales Netz das Atominteraktionen mit solchen faktorisierten Tensorlayern modelliert, bezeichnen wir als *Deep Tensor Neural Network (DTNN)* [Sc17a].

Eine Alternative um Atominteraktionen zu modellieren bieten Faltungslayer, ähnlich zu denen, die in neuronalen Netzen zur Bildklassifikation genutzt werden. So wie Bilder aus Pixeln aufgebaut sind, auf die eine diskrete Faltung angewandt wird, setzen sich atomistische Systeme aus Atomen zusammen. Diese sind allerdings, im Gegensatz zu Pixeln, nicht auf einem Gitter angeordnet, sondern können sich an beliebigen Positionen befinden. Aus diesem Grund kann eine Faltung mit diskretem Filter zu einer Energievorhersage mit Unstetigkeiten führen (siehe Abb. 2). Um das zu vermeiden wurden Faltungslayer der Form

$$\mathbf{x}_i^{l+1} = (X^l * W^l)_i = \sum_{j=1}^{n_{\text{atoms}}} \mathbf{x}_j^l \circ W(\mathbf{r}_i - \mathbf{r}_j), \tag{4}$$

mit der stetigen Filterfunktion $W$ eingeführt, welche wiederum durch ein neuronales Netz modelliert wird.

Aus der Kombination des oben beschriebenen DTNN-Ansatzes atomare Repräsentationen iterativ zu verfeinern, und Faltungslayer mit kontinuierlichen Filtern zur Modellierung der Interaktionen zu nutzen, wurde die *SchNet*-Architektur entwickelt (Abb. 3) [Sc17b, Sc18].

## 3 Analyse der gelernten Repräsentationen

Ein wichtiger Aspekt der eingeführten Modelle ist die Möglichkeit, die gelernte Repräsentation zu analysieren. Damit kann die Funktionsweise des neuronalen Netz studiert und sichergestellt werden, dass das Gelernte mit chemischer Intuition übereinstimmt. Dazu wurden in der Dissertation mehrere Methoden entwickelt, chemisch und räumlich aufgelöste Einblicke in die Repräsentation zu erhalten. Eine dieser Methoden ist das Visualisieren von lokalen, chemischen Potentialen der Moleküle. Dazu wird in das neuronale Netz neben dem zu analysierenden Molekül ein weiteres Atom als „Testladung“ gegeben, wobei dieses das Potential des Moleküls spürt, jedoch dessen Repräsentation nicht beeinflusst.

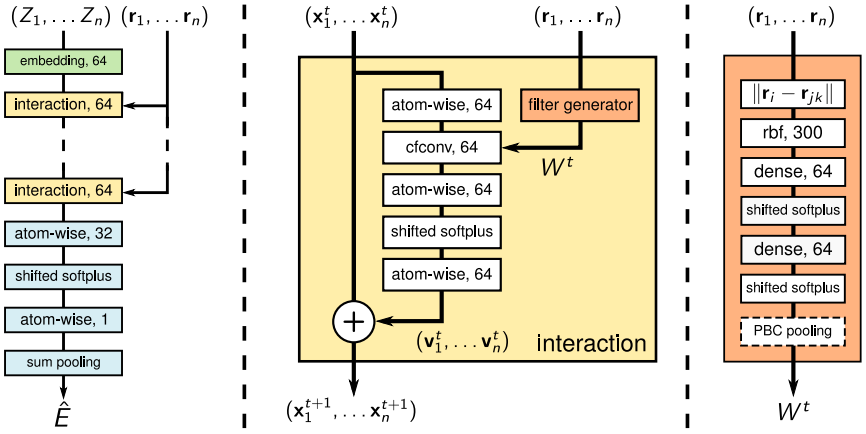



Abb. 3: Die Architektur von SchNet bestehend aus Elementeinbettungen (grün), Interaktions- (gelb) und Ausgabenetzen (blau) sowie Filternetzen (orange).

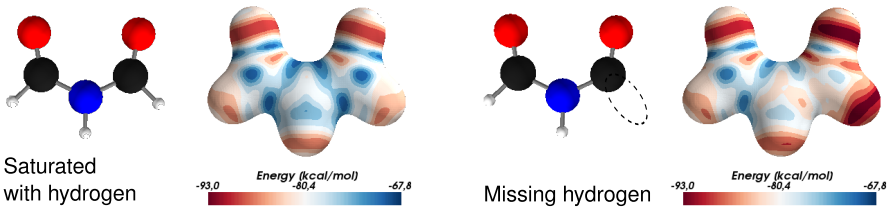



Abb. 4: Lokale, chemische Potentiale von N-Formylformamid mit einem Wasseratom als Testladung. Die Potentiale sind auf einer Isofläche mit $\sum_i \|\mathbf{r} - \mathbf{r}_i\| = 3.7$Å geplottet und wurden von einem auf dem QM9-Datensatz trainierten SchNet-Modell generiert.

Dies wird erreicht indem in den Interaktionslayern die Testladung von den Atomen des Moleküls nicht berücksichtigt wird.

Abb. 4 zeigt lokale, chemische Potentiale von N-Formylformamid. Links sieht man die vorhergesagte Energie des Wasserstoff-Testatoms auf einer Isofläche mit $\sum_i \|\mathbf{r} - \mathbf{r}_i\| = 3.7$Å. Die geringste Energie wird nahe der Sauerstoffatome (rot) erreicht, d.h. dass dort ein Wasserstoffatom (weiß) stärker vom Molekül angezogen wird. Dies entspricht der chemischen Intuition, da dort durch Umwandlung einer C-O-Doppelbindung in eine Einfachbindung ein weiteres Wasserstoffatom gebunden werden kann. Auf der rechten Seite sieht man die Auswirkung des Entfernen eines Wasserstoffes. Nun sinkt die Energie des Testatoms an der Fehlstelle, d.h. die offene Bindung wird in der Repräsentation reflektiert.

Ein entscheidender Nachteil von manuellen Deskriptoren ist, dass chemische Elemente entweder als orthogonal betrachtet werden oder Heuristiken zur Bestimmung ihrer Ähnlichkeit herangezogen werden müssen. Im DTNN-Ansatz werden jedoch Einbettungsvektoren für die Elemente direkt aus den Daten gelernt, so dass wir nun die Struktur dieses Vektorraums analysieren können. Abb. 5 zeigt die führenden zwei Hauptkomponenten der Elementeinbettungen für ein SchNet-Modell das auf dem Datensatz von 60,000 Ma-

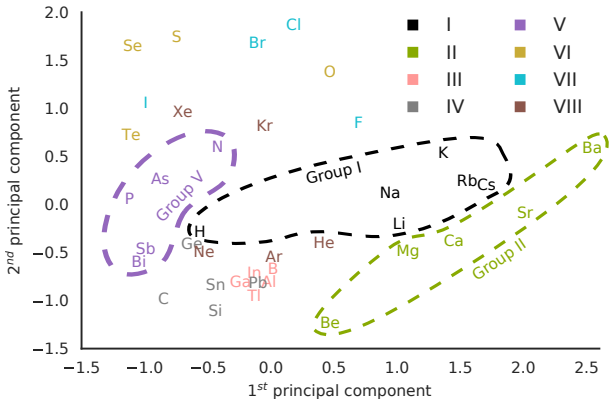



Abb. 5: Die ersten zwei Hauptkomponenten der von SchNet aus dem Materials-Project-Datensatz gelernten Elementeinbettungen $\mathbf{x}^0$. Im Einbettungsraum sind Cluster von Elementen zu erkennen, die den Hauptgruppen des Periodensystems entsprechen.

terialien aus dem *Materials Project*-Repository [Ja13] trainiert wurde. Bereits in dieser dimensionsreduzierten Darstellung zeigt sich, dass viele Elemente, die der selben Hauptgruppe (d.h. Spalte) des Periodensystems angehören, auch im Einbettungsraum gruppiert sind (siehe Farbkodierung in Abb. 5). Darüber hinaus zeigt sich innerhalb dieser Gruppen eine Ordnung welche den Perioden (d.h. den Reihen) entspricht. Beispielsweise ist die erste Hauptgruppe von links nach rechts von Wasserstoff (H) in der ersten Periode über Lithium (Li) und Natrium (Na) in der zweiten und dritten Periode zu den schweren Elementen (K, Rb, Cs) in der 4.-6. Periode. Ähnliche Ordnungen finden sich u.a. auch in der 2. Hauptgruppe (Be→Mg→Ca→Sr→Ba) sowie der 5. Hauptgruppe von Stickstoff (N) nach Bismut (Bi). Da das neuronale Netz keine explizite Information über die Struktur des Periodensystems erhalten hat, muss es dieses chemische Wissen aus den Geometrien und Energien der Materialien inferiert haben.

## 4 Anwendungen

### 4.1 Virtuelles Screening

In der Dissertation wird die Vorhersage verschiedener chemischer Eigenschaften auf diversen Molekül- und Materialdatensätzen demonstriert. Hier zeigen wir beispielhaft die Vorhersage der inneren Energie bei 0K $U_0$ auf dem QM9-Benchmark. Dieser besteht aus 133,885 kleinen, stabilen, organischen Molekülen, die aus bis zu 9 schweren Atomen aus den Elementen $\{C, O, N, F\}$ bestehen und mit Wasserstoff gesättigt wurden [Ra14, BR09, Re15]. Die mittleren, absoluten Fehler von DTNN und SchNet sind in Abb. 6 (links) abhängig von der Größe des Trainingssatzes dargestellt. Es zeigt sich, dass zum einen SchNet deutlich niedrigere Fehler als DTNN erreicht und zum anderen mehr Interaktionsblöcke $T$ ebenfalls zu einer Reduzierung des Fehlers beitragen. Bei Nutzung von 25,000

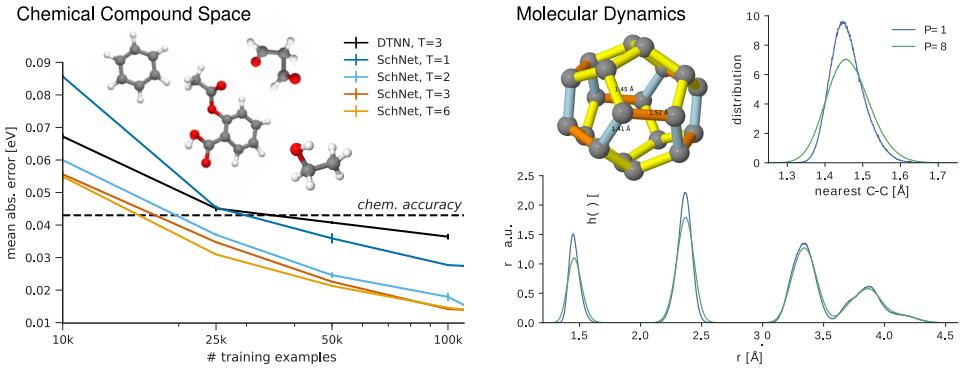



Abb. 6: Ergebnisse für Vorhersagen. Links: Die Lernkurven für DTNN und SchNet mit $T$ Interaktionsblöcken zeigen mittlere, absolute Fehler (MAE) in eV für die innere Energie ($U_0$) auf dem QM9-Datensatz. Rechts: Mit Energie- und Kraftvorhersagen von SchNet wurden Ringpolymer-MD-Simulationen für das Fulleren $C_{20}$ mit $P \in \{1,8\}$ Systemreplikas durchgeführt.

Referenzrechnungen, erreicht SchNet mit $T \geq 2$ hier bereits deutlich geringere Fehler als *chemische Genauigkeit*, welche im Allgemeinen als 0.043eV angenommen wird.

### 4.2 Moleküldynamiksimulationen

Neben der Vorhersage von chemischen Eigenschaften in stabilen Molekülkonfiguration, ist eine weitere Anwendung die Analyse des dynamischen Verhaltens eines Moleküls. Dabei wird die Bewegung eines Moleküls in einer durch ein Temperaturbad approximierten Umgebung simuliert. Dazu ist es nötig die auf die Atome wirkenden Kräfte zu berechnen, welche gentutzt werden um die neuen Atompositionen zu berechnen. Die Kräfte sind dabei die negative Ableitung der Energie nach den Atompositionen

$$\mathbf{F}_i(\mathbf{r}_1,\ldots,\mathbf{r}_n) = -\frac{\partial E}{\partial \mathbf{r}_i}(\mathbf{r}_1,\ldots,\mathbf{r}_n). \tag{5}$$

Daher können die atomaren Kräfte direkt von einem der oben vorgestellten Energiemodelle analytisch berechnet werden, was einer Backpropagation der Energievorhersage entspricht. Da viele Quantenchemiedaten Kräfte enthalten, können diese ebenfalls zum Training des Modells genutzt werden, indem ein kombinierter Energie-Kraft-Loss

$$\rho \|E - \hat{E}\|^2 + \frac{1}{n_{\text{atoms}}} \sum_{i=0}^{n_{\text{atoms}}} \left\| \mathbf{F}_i - \left( -\frac{\partial \hat{E}}{\partial \mathbf{r}_i} \right) \right\|^2, \tag{6}$$

genutzt wird, wobei $\rho$ den Einfluss des Energiefehlers skaliert. Das zusätzliche Training mit den atomaren Kräften stellt hier dem neuronalen Netz Information über die unmittelbare Umgebung des Datenpunktes zur Verfügung, was die benötigte Menge von Trainingspunkten drastisch reduziert.

Abb. 6 (rechts) zeigt die Ergebnisse einer Ringpolymer-MD-Simulation (RPMD) für das Fulleren $C_{20}$. In der üblichen Born-Oppenheimer-Näherung werden Atomkerne klassisch

behandelt, d.h. sie besitzen eine feste Position zu einem gegebenen Zeitpunkt. Im Gegensatz dazu wird bei einer RPMD die quantenmechanische Aufenthaltswahrscheinlichkeit durch einen Ring von gekoppelten Atomen dargestellt. Dies erfordert zusätzliche Quantenrechnungen für jedes Atomreplika $P$ im Ring, wobei $P = 1$ der klassischen Born-Oppenheimer-MD entspricht. In Abb. 4 (rechts) wurden Simulationen mit $P = 1$ und $P = 8$ bei einer Temperatur von 300K mithilfe eines SchNet-Modells durchgeführt, welches auf 20,000 $C_{20}$-Konfigurationen trainiert wurde. In den Plots ist zu sehen, wie sich die Distanzverteilungen aufgrund der nun berücksichtigten nuklearen Quanteneffekte für $P = 8$ verbreitern. Um diese Ergebnisse zu ermitteln wurde eine 1.25ns MD-Trajektorie mit einem Zeitschritt von 0.5fs erzeugt, womit für 8 Atomreplika insgesamt 20 Millionen Quantenberechnungen durchgeführt werden müssten. Da jede einzelne DFT-Rechnung auf 32 CPU-Kernen eine Rechenzeit von 11s benötigt, würde dies mit konventionellen Methoden eine Rechenzeit von ca. *7 Jahren* erfordern. SchNet verkürzt die Rechenzeit um Faktor $10^3$-$10^4$ auf ca. *7 Stunden*, wodurch vorher praktisch undurchführbare Simulationen möglich werden.

## 5 Schlussfolgerungen

In der zusammengefassten Dissertation wurden neuronale Netze zum Lernen von Repräsentationen entwickelt, wobei durch Nutzung von anwendungsspezifischem *a priori* Wissen die Dateneffizienz erhöht wird. Latente Variablen, die Konzepten aus dem Anwendungsgebiet entsprechen, ermöglichen dabei eine natürliche Interpretation des Modells. Bei der Anwendung auf Moleküle und Materialien wurde gezeigt, dass die entwickelten Modelle genaue Vorhersagen chemischer Eigenschaften treffen und Molekülдynamiksimulationen um mehrere Größenordnungen beschleunigen, da DTNN und SchNet linear mit der Anzahl der Atome skalieren. Während in der Arbeit DFT-Referenzrechnungen genutzt wurden, ist die Übertragung auf genauere Methoden (z.B. CCSD(T) mit $O(n^7)$) problemlos möglich [Ch18]. So kann SchNet selbst für kleine Moleküle Beschleunigungen um einen Faktor von mehreren Millionen erzielen. Die Analyse der Repräsentationen ermöglicht die Entdeckung chemischer Zusammenhänge, die nicht in den einzelnen Referenzrechnungen enthalten sind. U.a. wurden Einbettungen chemischer Elementen gelernt, die der Struktur des Periodensystems entsprechen sowie lokale Potentiale welche Bindungsmuster des Moleküls widerspiegeln. Damit sind die eingeführten, neuronalen Netze eine wichtige Grundlage zur Unterstützung naturwissenschaftlicher Forschung durch maschinelles Lernen, sowohl für die Beschleunigung von Simulationen als auch für den Gewinn wissenschaftlicher Erkenntnisse.

## Literaturverzeichnis

[BKC13] Bartók, Albert. P.; Kondor, Risi; Csányi, Gabor: On representing chemical environments. Phys. Rev. B, 87(18):184115, 2013.

[BP07] Behler, Jörg; Parrinello, Michele: Generalized neural-network representation of high-dimensional potential-energy surfaces. Phys. Rev. Lett., 98(14):146401, 2007.

[BR09] Blum, Lorenz C; Reymond, Jean-Louis: 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. J. Am. Chem. Soc., 131:8732, 2009.

[Br17] Brockherde, F.; Voigt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K.-R.: Bypassing the Kohn-Sham equations with machine learning. Nature Communications, 8:872, 2017.

[Ch17] Chmiela, Stefan; Tkatchenko, Alexandre; Sauceda, Huziel E; Poltavsky, Igor; Schütt, Kristof T; Müller, Klaus-Robert: Machine learning of accurate energy-conserving molecular force fields. Science advances, 3(5):e1603015, 2017.

[Ch18] Chmiela, Stefan; Sauceda, Huziel E; Müller, Klaus-Robert; Tkatchenko, Alexandre: Towards Exact Molecular Dynamics Simulations with Machine-Learned Force Fields. Nature Communications, 9, 2018.

[De09] Deng, Jia; Dong, Wei; Socher, Richard; Li, Li-Jia; Li, Kai; Fei-Fei, Li: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. Ieee, S. 248–255, 2009.

[Fa17] Faber, Felix A; Hutchison, Luke; Huang, Bing; Gilmer, Justin; Schoenholz, Samuel S; Dahl, George E; Vinyals, Oriol; Kearnes, Steven; Riley, Patrick F; von Lilienfeld, O Anatole: Prediction errors of molecular machine learning models lower than hybrid DFT error. J. Chem. Theory Comput., 13(11):5255–5264, 2017.

[Fa18] Faber, Felix A; Christensen, Anders S; Huang, Bing; von Lilienfeld, O Anatole: Alchemical and structural distribution based representation for universal quantum machine learning. The Journal of Chemical Physics, 148(24):241717, 2018.

[Ja13] Jain, Anubhav; Ong, Shyue Ping; Hautier, Geoffroy; Chen, Wei; Richards, William Davidson; Dacek, Stephen; Cholia, Shreyas; Gunter, Dan; Skinner, David; Ceder, Gerbrand; Persson, Kristin a.: The Materials Project: A materials genome approach to accelerating materials innovation. APL Materials, 1(1):011002, 2013.

[Ki18] Kindermans, Pieter-Jan; Schütt, Kristof T; Alber, Maximilian; Müller, Klaus-Robert; Erhan, Dumitru; Kim, Been; Dähne, Sven: Learning how to explain neural networks: PatternNet and PatternAttribution. In: International Conference on Learning Representations (ICLR). 2018.

[KSH12] Krizhevsky, A.; Sutskever, I.; Hinton, G. E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. S. 1097–1105, 2012.

[LBH15] LeCun, Y.; Bengio, Y.; Hinton, G.: Deep learning. Nature, 521(7553):436–444, 2015.

[Li15] von Lilienfeld, O Anatole; Ramakrishnan, Raghunathan; Rupp, Matthias; Knoll, Aaron: Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties. International Journal of Quantum Chemistry, 115(16):1084–1093, 2015.

[Mn15] Mnih, Volodymyr; Kavukcuoglu, Koray; Silver, David; Rusu, Andrei A; Veness, Joel; Bellemare, Marc G; Graves, Alex; Riedmiller, Martin; Fidjeland, Andreas K; Ostrovski, Georg et al.: Human-level control through deep reinforcement learning. Nature, 518(7540):529, 2015.

[Mo17] Montavon, Grégoire; Lapuschkin, Sebastian; Binder, Alexander; Samek, Wojciech; Müller, Klaus-Robert: Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition, 65:211–222, 2017.

[Ra14] Ramakrishnan, Raghunathan; Dral, Pavlo O; Rupp, Matthias; Von Lilienfeld, O Anatole: Quantum chemistry structures and properties of 134 kilo molecules. Scientific data, 1:140022, 2014.

[Re15] Reymond, J.-L.: The chemical space project. Acc. Chem. Res., 48(3):722–730, 2015.

[Ru12] Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A.: Fast and accurate modeling of molecular atomization energies with machine learning. Phys. Rev. Lett., 108(5):058301, 2012.

[Sc14] Schütt, Kristof T; Glawe, Henning; Brockherde, Felix; Sanna, Antonio; Müller, Klaus-Robert; Gross, Eberhard KU: How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. Phys. Rev. B, 89(20):205118, 2014.

[Sc15] Schmidhuber, Jürgen: Deep learning in neural networks: An overview. Neural networks, 61:85–117, 2015.

[Sc17a] Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K.-R.; Tkatchenko, A.: Quantum-chemical insights from deep tensor neural networks. Nature Communications, 8, 2017.

[Sc17b] Schütt, K. T.; Kindermans, P.-J.; Sauceda, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R.: SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In: Advances in Neural Information Processing Systems 30. S. 992–1002, 2017.

[Sc18] Schütt, Kristof T; Sauceda, Huziel E.; Kindermans, Pieter-Jan; Tkatchenko, Alexandre; Müller, Klaus-Robert: SchNet - a deep learning architecture for molecules and materials. The Journal of Chemical Physics, 148(24), 2018.

[SVL14] Sutskever, Ilya; Vinyals, Oriol; Le, Quoc V: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. S. 3104–3112, 2014.

[Vi15] Vinyals, Oriol; Toshev, Alexander; Bengio, Samy; Erhan, Dumitru: Show and tell: A neural image caption generator. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, S. 3156–3164, 2015.

**Kristof T. Schütt** ist wissenschaftlicher Mitarbeiter am Berliner Zentrum für Maschinelles Lernen. Er studierte Informatik an der TU Berlin und beschäftigte sich in seiner Masterarbeit mit der frühen Erkennung von JavaScript-Schadcode mit maschinellem Lernen. Im Mai 2018 schloss er seine Promotion am Fachbereich Maschinelles Lernen der TU Berlin zum maschinellen Lernen von Repräsentationen atomistischer Systeme ab. Sein aktueller Forschungsschwerpunkt liegt auf der Entwicklung interpretierbarer, neuronaler Netze und insbesondere deren Anwendung in der Quantenchemie.

# Logiksynthese für In-Memory-Computing mit resistiven Speichern[1]

Saeideh Shirinzadeh[2]

**Abstract:** Die wachsende Notwendigkeit, das Problem des Speicherengpasses in den gegenwärtigen Computerarchitekturen zu umgehen, hat zu einer großen Aufmerksamkeit für das In-Memory-Computing geführt, welches durch aufkommende Speichertechnologien wie RRAM (Resistive Random Access Memory) ermöglicht wird. Diese Dissertation untersucht In-Memory-Computing aus zwei Perspektiven, nämlich "customized" und befehlsbasiert. Der customized-Ansatz nutzt logische Repräsentationen, um In-Memory-Computing-Schaltungen zu realisieren. Der Ansatz schlägt Designansätze und Optimierungsalgorithmen für jede Repräsentation in Bezug auf Fläche und Latenz bei der Realisierung ihrer logischen Grundelemente vor. Beim befehlsbasierten Ansatz wird ein automatischer Compiler eingesetzt, um Logic-in-Memory-Computerarchitektur zu verwenden und resultierende Programme zu optimieren. Die experimentellen Ergebnisse für beide Ansätze zeigen erhebliche Verbesserungen gegenüber dem Stand der Technik.

## 1 Einführung

Die Entwicklungsfortschritte bei den Prozessoren moderner Computer übertreffen die des Arbeitsspeichers in Bezug auf Zugriffzeit bei Weitem. Die wesentlich höhere Latenzzeit des Speichers im Vergleich zu einem Prozessor einerseits und die Kommunikation zwischen diesen beiden in der von-Neumann-Architektur andererseits schränken die Gesamtleistung aktueller Computersysteme ein, was als *memory wall* bezeichnet wird. Neue Anwendungen wie *Internet der Dinge* (*engl*. Internet of Things: IoT) und *big data*, die mit großen Datenmengen umgehen, stellen hohe Anforderungen, was zu intensiver Forschung führt [So17]. Unter den möglichen Ansätzen erscheint *in-memory computing* als sehr vielversprechend. Die Integration der Speicher und Rechenparadigmen erlaubt es, die "memory wall" zu überwinden und die Leistung um ein Vielfaches zu steigern.

Eine der Kerntechnologien für In-Memory-Computing, ist RRAM (Resistive Random Access Memory). RRAM ist eine vielversprechende nicht-volatile Speichertechnologie mit hoher Skalierbarkeit und ohne Energieverlust im Standby, bei der der Innenwiderstand zwischen zwei Zuständen (hoch/niedrig) umgeschaltet werden kann. Es wurden verschiedene Ansätze vorgeschlagen, die diese resistive Schalteigenschaft nutzen, um logische Operationen innerhalb von RRAMs auszuführen. Der Gebrauch von *Material Implication* (IMP) wurde in großem Umfang für In-Memory Computing verwendet [Bo10]. In [Kv14] wurde eine Klasse logischer Operationen namens MAGIC vorgeschlagen, mit der Boolesche Funktionen mit NOR- und NOT-Gattern realisiert werden können. In [Ga16] wurde gezeigt, dass sich eine Majority-basierte Logik für In-Memory-Computing verwenden lässt, da RRAM Majority (MAJ) nativ implementiert werden kann und es somit

[1] Synthesis and Optimization for Logic-in-Memory Computing using Memristive Devices

[2] Universität Bremen, Fachbereich Mathematik und Informatik / Cyber-Physical Systems, DFKI GmbH
Bibliothekstr. 5, 28359 Bremen, Deutschland

möglich ist, Mehrheitsvergleiche als Logikbausteine für In-Memory-Computing zu verwenden wird.

In dieser Dissertation untersuchen wir die Synthese für In-Memory-Computing aus zwei verschiedenen Perspektiven, d.h. (i) basierend auf einem „customized" Ansatz auf Gatterebene, die logische Repräsentationen verwendet, und (ii) einem befehlsorientierten Ansatz für die effiziente Kompilierung und Ausführung von Programmen auf einer Logic-in-Memory-Architektur. Der vorgestellte customized-Syntheseansatz verwendet IMP und MAJ als grundlegende Operationen für RRAM-Arrays, während der befehlsbasierte Ansatz MAJ nur für die Ausführung von Programmen verwendet.

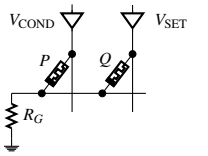
Der customized-Syntheseansatz beginnt mit der Suche nach effizienten Realisierungen für die logischen Grundelemente der verwendeten Repräsentationen, d.h. *Binary Decision Diagrams* (BDDs), *AND-Inverter-Graphen* (AIGs) und *Majority-Inverter-Graphen* (MIGs). Der Ansatz bietet Optimierungsalgorithmen und eine umfassende Entwurfsmethodik, um jede Darstellung in äquivalente Sequenzen von Operationen und RRAMs abzubilden, die auf einem resistiven Speicherarray ausgeführt werden sollen. Die Ergebnisse des vorgestellten customized-Ansatzes zeigen eine deutliche Verbesserung gegenüber herkömmlichen Ansätzen.

Durch den befehlsbasierten Ansatz automatisieren und optimieren wir vollständig eine vorhandene Computerarchitektur. Darüber hinaus befassen wir uns mit dem Problem der geringeren Schreibdauer von RRAM-Schaltern und schlagen Techniken zur Verschleißminderung vor, um die Lebensdauer der Architekturen zu erhöhen. Experimente, die mit großen Arithmetik- und Steuerfunktionen durchgeführt wurden, zeigen eine beträchtliche Verbesserung der Verteilung von Schreibvorgängen im gesamten Speicherfeld. Auch die Anzahl der Zyklen und der RRAM-Vorrichtungen, die die Latenz und den Bereich der resultierenden Implementierungen darstellen, konnte reduziert werden.

## 2 Logikoperationen innerhalb von RRAM

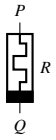
In [Bo10] wurde gezeigt, dass die Material Implication (IMP), d.h. $q' \leftarrow p \,\mathrm{IMP}\, q = \overline{p} + q$, aus der Interaktion zweier RRAM-Schalter ausgeführt werden kann; unter bestimmten Spannungspegeln, die durch $V_{\mathrm{SET}}$ und $V_{\mathrm{COND}}$ in Abb. 1 (a) angegeben sind. Die logischen Zustände der resistiven Schalter können auch einfach zwischen logischer 1 oder 0, d.h. FALSE-Betrieb, umgeschaltet werden, wenn sie an geeignete Spannungsimpulse angelegt werden. IMP und FALSE bilden zusammen einen universellen Satz von Logikoperationen, die ausreichen, um alle Boolesche Funktionen auf resistiven Arrays auszuführen.

In [Ga16] wurde eine von RRAM-Schaltern aktivierte intrinsische Majoritätsoperation eingeführt, die ausreicht, um jede Boolesche Funktion zu berechnen. Bezeichnen wir die obere und die untere Elektrode eines RRAM-Schalters mit $P$ und $Q$ (siehe Abb. 1 (b)). Angenommen, der aktuelle resistive Zustand des Schalters ($R$) kann durch Anlegen eines positiven oder negativen Spannungspegels $V_{PQ}$ auf 1 oder 0 umgeschaltet werden, dann ändert sich der nächste Zustand des RRAMs ($R'$)) basierend auf den in Abb. 1 (b) gezeigten Wahrheitstabellen. Durch Erweitern der Booleschen Relation in den Tabellen kann einfach gezeigt werden, dass der nächste Widerstandszustand des Schalters




| $p$ | $q$ | $q'$ |
|---|---|---|
| 0 | 0 | 1 |
| 0 | 1 | 1 |
| 1 | 0 | 0 |
| 1 | 1 | 1 |

(a)




| $P$ | $Q$ | $R$ | $R'$ |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 |
| 1 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 |

$R' = P \cdot \overline{Q}$

| $P$ | $Q$ | $R$ | $R'$ |
|---|---|---|---|
| 0 | 0 | 1 | 1 |
| 0 | 1 | 1 | 0 |
| 1 | 0 | 1 | 1 |
| 1 | 1 | 1 | 1 |

$R' = P + \overline{Q}$

(b)

Abb. 1: Die Booleschen Operationen, die innerhalb von RRAM ausgeführt werden können. (a) Implementierung des IMPs und seine Wahrheitstabelle [Bo10]. (b) Die intrinsische Majority Operation innerhalb eines RRAM-Schalters (MAJ) [Ga16].

$R' = M(P, \overline{Q}, R) = P \cdot \overline{Q} + P \cdot R + \overline{Q} \cdot R$ ist, d.h. das Ergebnis der Majoritätsfunktion mit drei Eingängen, die in diesem Dokument mit MAJ bezeichnet wird.

## 3 Customized Syntheseansatz

Der customized Syntheseansatz umfasst drei Stufen: (i) Finden effizienter Realisierungen mit RRAM-Schaltern für das logische Grundelement jeder Repräsentation BDD, AIG und MIG, (ii) Definieren der Entwurfsmethodik zum Abbilden der Repräsentationen auf das RRAM-Array (iii) und schließlich Optimieren der Repräsentationen mit Bezug auf die Anzahl der RRAMs und Rechenoperationen, die durch die Entwurfsmethodik bestimmt werden. Im Folgenden erläutern wir kurz den vorgeschlagenen customized Ansatz für die Logikdarstellung MIG und geben ein Implementierungsbeispiel. Aus Platzgründen verweisen wir Leserinnen und Leser auf [Sh18], wo Einzelheiten zu den Optimierungsalgorithmen für jede Darstellung beschrieben sind.

Die IMP-basierte Realisierung [Sh16, Sh18] für das Majority-Gatter ist im Folgenden dargestellt. Die Realisierung erfordert sechs RRAMs, Eingabeschalter $X$, $Y$ und $Z$ und zusätzliche Schalter $A$, $B$ und $C$, die zum Negieren oder Speichern der Operationsausgänge erforderlich sind. Die Majority-Funktion wird nach zehn Operationen ausgeführt. Bei der ersten Operation werden die erforderlichen Schalter mit den Eingabevariablen und Null geladen, und die restlichen Schritte umfassen die Operationen IMP und FALSE.

| | |
|---|---|
| **01:** $X = x, Y = y, Z = z$ $A = 0, B = 0, C = 0$ | **06:** $c \leftarrow y \,\text{IMP}\, c = \overline{x + y}$ |
| **02:** $a \leftarrow x \,\text{IMP}\, a = \bar{x}$ | **07:** $c \leftarrow z \,\text{IMP}\, c = \overline{x \cdot z + y \cdot z}$ |
| **03:** $b \leftarrow y \,\text{IMP}\, b = \bar{y}$ | **08:** $a = 0$ |
| **04:** $y \leftarrow a \,\text{IMP}\, y = x + y$ | **09:** $a \leftarrow b \,\text{IMP}\, a = x \cdot y$ |
| **05:** $b \leftarrow x \,\text{IMP}\, b = \bar{x} + \bar{y}$ | **10:** $a \leftarrow c \,\text{IMP}\, a = x \cdot y + y \cdot z + x \cdot z.$ |

Es ist offensichtlich, dass die MAJ-basierte Realisierung für die MIG-basierte Synthese durch die Nutzung der nativ implementierten Majority-Funktion in RRAM-Schaltern effizienter realisiert werden kann. Wie im Folgenden gezeigt, erfordert die Realisierung eines Majority-Gatters in diesem Fall maximal vier Schalter und drei Schritte.

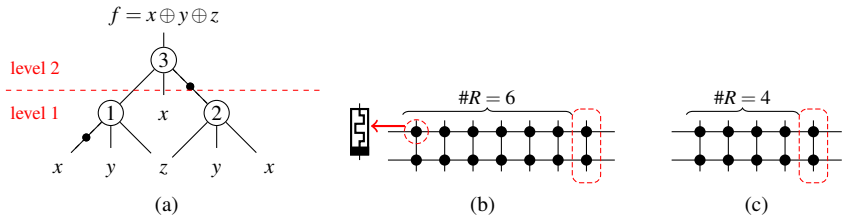



Abb. 2: (a) MIG, der ein Drei-Bit-XOR-Gatter darstellt, und eine obere Grenze der RRAM Crossbar für deren Implementierung mit (b) IMP-basierter and (c) MAJ-basierter Realisierung.

**1:** $X = x, Y = y, Z = z, A = 0$
**2:** $P_A = 1, Q_A = y, R_A = 0 \Rightarrow R'_A = \bar{y}$
**3:** $P_Z = x, Q_Z = \bar{y}, R_Z = z \Rightarrow R'_Z = M(x, y, z)$.

Der MIG wird zunächst hinsichtlich der Anzahl der RRAMs und der Rechenschritte optimiert. Wir schlagen drei verschiedene Optimierungsalgorithmen im Bezug auf eine oder beide Kostenmetriken vor. Weitere Informationen können unter Optimierungsalgorithmen in [SSD16, Sh18] gefunden werden. Anschließend können sie auf RRAM-Arrays gemäß einer Level-by-Level-Entwurfsmethodik implementiert werden. Dies bedeutet, dass ausgehend vom unteren Rand des Diagramms alle Knoten in jeder MIG-Ebene gleichzeitig berechnet werden. Nach der Berechnung der einzelnen Ebenen werden die RRAM-Schalter freigegeben und können als Eingabeschalter für die Berechnung der nächsten Ebene verwendet werden. Diese Prozedur wird fortgesetzt, bis die Wurzelfunktion berechnet ist.

Zur Berechnung jedes Levels umfasst die Anzahl der erforderlichen RRAMs das Sechsfache (mit der IMP-basierte Realisierung) oder das Vierfache (mit der MAJ-basierte Realisierung) der Anzahl der Knoten in der Ebene. Für jede ergänzte Kante wird auch ein zusätzlicher RRAM-Schalter berücksichtigt. Da die RRAMs wiederverwendet werden, entspricht die Anzahl der Schalter, die zur Berechnung eines MIGs erforderlich sind, der maximalen Anzahl der erforderlichen Schalter über alle Ebenen. Die Anzahl der Berechnungsschritte zum Berechnen eines MIGs beträgt mindestens zehn oder dreimal die Anzahl der BDD-Stufen für IMP- bzw. MAJ-basierte Realisierungen. Dieser Wert muss jedoch zu der Anzahl der Ebenen hinzugefügt werden, die über komplementierte Kanten verfügen, da ihre Negation zusätzliche Operationen benötigt [Sh18].

Abbildung 2 (a) zeigt einen MIG, der ein XOR-Gatter mit drei Eingängen darstellt. Hier zeigen wir, wie dieses MIG auf einem RRAM-Array mit Schaltern implementiert werden kann, die durch $R_{ij}$ dargestellt werden, wobei $i$ und $j$ die Indizes der Zeile bzw. der Spalte bezeichnen.

**Synthese mit IMP-basierter Realisierung.** Die vorgestellte Entwurfsmethodik berechnet alle Knoten gleichzeitig auf einer Ebene und weist zu diesem Zweck jedem Knoten eine Zeile zu. Dies bedeutet, dass in jeder Zeile eine einzelne Operation pro Zyklus ausgeführt wird. Da das Beispiel für MIGs eine maximale Ebenengröße von zwei hat, benötigt die erforderliche RRAM crossbar mindestens zwei Reihen mit mindestens sechs Schaltern (siehe Abbildung 2 (b)). Ein zusätzlicher Schalter am Ende jeder Reihe wird für ergänzte

Kanten verwendet. Wie erwartet, ist die Anzahl der erforderlichen Schritte 22, d.h. das 10-fache der Tiefe 2 plus zwei zusätzliche Schritte für die ergänzten Kanten auf beiden Ebenen.

| | |
|---|---|
| Initialisierung | $R_{ij} = 0;$ |
| **1:** Laden für Ebene 1 | $R_{11} = x, R_{12} = y, R_{13} = z;$<br>$R_{21} = x, R_{22} = y, R_{23} = z;$ |
| **2:** Negation für Knoten 1 | $R_{17} \leftarrow x \,\text{IMP}\, R_{17} : R_{17} = \bar{x};$ |
| **3-11:** Berechnung der Ebene 1 | Knoten 1: $R_{14} = M(\bar{x}, y, z);$<br>Knoten 2: $R_{24} : M(x, y, z);$ |
| **12:** Laden für Ebene 2 | $R_{11} = x, R_{12} = M(x, y, z), R_{13} = M(\bar{x}, y, z)$<br>$R_{14} = R_{15} = R_{16} = R_{17} = 0;$ |
| **13:** Negation für Knoten 3 | $R_{17} \leftarrow R_{12} \,\text{IMP}\, R_{17} :$<br>$R_{17} = \overline{R_{12}} = \overline{M}(x, y, z);$ |
| **14-22:** Berechnung der Ebene 2 | $R_{14} = M(M(\bar{x}, y, z), x, \overline{M}(x, y, z));$ |

**Synthese mit** MAJ**-basierter Realisierung.** Die Schritte für die MAJ-basierte Implementierung des MIGs, dargestellt in Abb. 2 (a), werden im Folgenden gezeigt. Wie die Schritte zeigen, wird die XOR-Funktion mit nur drei RRAM Schaltern und innerhalb von nur vier Schritten ausgeführt, trotz der oberen Grenze von 8, die für ein MIG mit zwei Ebenen mit ergänzten Kanten erwartet werden. In der Tat können die komplementierten Kanten an den Knoten 1 und 3 direkt als zweite Eingabe von MAJ verwendet werden, ohne invertiert zu werden. Darüber hinaus können die aktualisierten RRAMs als Eingaben für den nächsten Zyklus verwendet werden, sodass der Schritt des Ladens nicht erforderlich ist. Es sollte beachtet werden, dass das Anwenden von Signalen auf die Zeilen und Spalten während der MAJ-basierten Implementierung eine Datenverzerrung vermeiden soll, indem zuvor berechnete Ergebnisse und die gleichzeitigen Operationen in anderen Zeilen beibehalten werden [Sh18]. Zum Beispiel werden in Schritt 2 die Werte von $R_{11}$ und $R_{21}$ beibehalten, indem die logischen Zustände ihrer Terminale angeglichen werden, wenn eine MAJ-Operation innerhalb von $R_{25}$ ausgeführt wird.

| | |
|---|---|
| Initialisierung | $R_{ij} = 0 : Q_{ij} = 1, P_{ij} = 0;$ |
| **1:** Laden | $Q_1 = Q_2 = 0, P_1 = P_2 = z;$<br>$R_{11} : RM_3(z, 0, 0) = M(z, 1, 0) = z;$<br>$R_{21} : RM_3(z, 0, 0) = M(z, 1, 0) = z;$ |
| **2:** Negation für Knoten 2 | $Q_1 = Q_2 = x, P_1 = x, P_2 = 1;$<br>$R_{25} : RM_3(1, x, 0) = M(1, \bar{x}, 0) = \bar{x};$ |
| **3:** Berechnung der Ebene 1 | Knoten 1: $P_1 = y, Q_1 = x, R_{11} = z$<br>$R_{11} : RM_3(y, x, z) = M(y, \bar{x}, z);$<br>Knoten 2: $P_1 = y, Q_2 = \bar{x}\ (@R_{25})\ , R_{21} = z;$<br>$R_{21} : RM_3(y, \bar{x}, z) = M(y, x, z);$ |

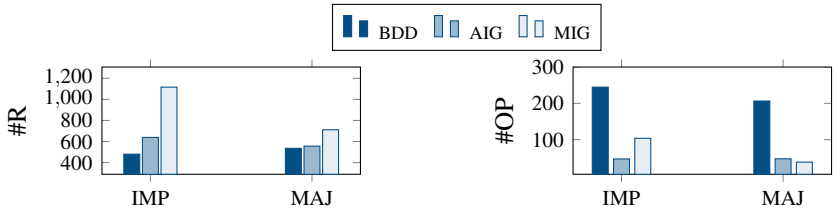



Abb. 3: Vergleich der Syntheseergebnisse anhand von logischen Darstellungen für In-Memory-Computing. (a) Die durchschnittliche Anzahl der RRAM-Schalter, (b) die durchschnittliche Anzahl der Operationen.

**4:** Berechnung der Ebene 2 $P_1 = x, Q_1 = @R_{21}, R_{11} = M(\bar{x}, y, z);$
$R_{11} : RM_3(x, @R_{21}, @R_{11}) = M(x, \overline{@R_{21}}, @R_{11}) :$
$M(M(\bar{x}, y, z), x, \overline{M}(x, y, z));$

Abb. 3 vergleicht die Syntheseergebnisse auf der Grundlage der drei Darstellungen, wobei sowohl IMP als auch MAJ für die Implementierung verwendet werden. Wie die Abbildung zeigt, benötigt der BDD-basierte Ansatz die geringste Anzahl von Schaltern, führt jedoch zu einer hohen Latenz. Andererseits erfordern Synthesemethoden, die auf AIG und MIG basieren, mehr RRAMs, verringern jedoch die Operationsdauer. Insbesondere der MIG-basierte Ansatz unter Verwendung von MAJ führt zu Implementierungen mit der geringsten Latenzzeit.

## 4 Befehlsbasierter Syntheseansatz

In [Ga16] wurde eine *Programmable Logic-in-Memory* (PLiM) Computerarchitektur vorgeschlagen, die es erlaubt, logische Operationen an einem regulären RRAM-Array durchzuführen. PLiM verfügt über einen Controller, der aus einer einfachen Zustandsmaschine und einigen Arbeitsregistern besteht, um MAJ-Operationen auszuführen. Es ist nur eine einzige MAJ-Anweisung pro Zyklus zulässig. Eine Anweisung hat das Format $M(P, \bar{Q}, R)$ mit drei zuzuweisenden Operanden. Der erste Operand $P$ ist das an die obere Elektrode der RRAM-Vorrichtung angelegte Signal, d.h. der Zeilentreiber, und der zweite Operand $Q$ ist das an die untere Elektrode angelegte Signal, d.h. der Spaltentreiber. Der dritte Operand $R$ ist der aktuelle Status des zu berechnenden Schalters, der automatisch aktualisiert wird, wenn die Anweisung ausgeführt wird.

Der PLiM-Computer leitet den Befehlssatz zum Berechnen einer Booleschen Funktion aus seiner MIG-Darstellung ab. Die Eigenschaften eines MIGs, dessen Knotenreihenfolge für Berechnungen und die resultierende Zuweisung der Operanden zu jeder Instruktion beeinflussen die Anzahl der benötigten RRAMs und Instruktionen, welche sich wiederum auf die Fläche und die Verzögerungszeit der resultierenden PLiM-Implementierungen auswirken [So17, So16, Sh17b]. In dieser Arbeit schlagen wir einen automatischen Compiler vor, der PLiM-Programme zur Berechnung beliebiger Funktionen generiert und dabei die oben genannten Kostenmetriken berücksichtigt.

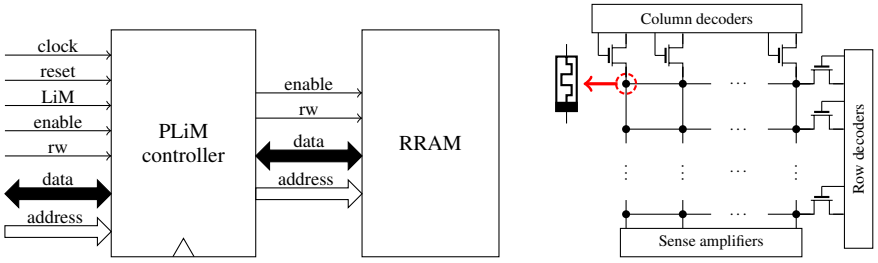
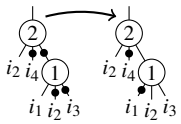



Abb. 4: (a) Die PLiM-Computerarchitektur [Ga16]. (b) Eine Speicherbank eines RRAM Crossbar-Arrays, auf dem ein PLiM-Programm ausgeführt wird.

Da PLiM Berechnungen vollständig seriell durchführt, ist die Anzahl der Knoten in einem MIG, d.h. die Größe des MIGs, ein bestimmender Faktor für die Länge der resultierenden Sequenz von Befehlen. Daher kann das Optimieren der MIGs in Bezug auf die Anzahl von Knoten die Latenz von PLiM-Implementierungen erheblich verbessern. Bei der MIG-Optimierung sollte jedoch nicht nur die Anzahl der Knoten berücksichtigt werden. MAJ benötigt im Idealfall eine komplementäre Kante, um einen Knoten innerhalb einer einzelnen Anweisung zu berechnen. Andernfalls sind ein zusätzlicher Schalter und eine Anweisung erforderlich, um einen Operanden zu invertieren, bevor der Knoten berechnet wird, wodurch die Anzahl und Position der komplementierten Kanten beeinflusst wird. Abb. 5 zeigt einen Beispiel-MIG mit zwei Knoten vor und nach der Optimierung, bei dem der Graph nur bezüglich der komplementierten Kanten geändert wurde. Wie die Abbildung zeigt, reduziert die MIG-Optimierung sowohl die Anzahl der RRAMs als auch die Anzahl der Anweisungen. Wir verweisen Leserinnen und Leser auf [So16] für den MIG-Optimierungsalgorithmus für PLiM.




| **Vor Optimierung** | | **Nach Optimierung** | |
|---|---|---|---|
| **1:** 0, 1, $@R_1$ | $R_1 \leftarrow 0$ | **1:** 0, 1, $@R_1$ | $R_1 \leftarrow 0$ |
| **2:** 1, $i_3$, $@R_1$ | $R_1 \leftarrow \bar{i}_3$ | **2:** $i_3$, 0, $@R_1$ | $R_1 \leftarrow i_3$ |
| **3:** $i_1$, $i_2$, $@R_1$ | $R_1 \leftarrow N_1$ | **3:** $i_2$, $i_1$, $@R_1$ | $R_1 \leftarrow N_1$ |
| **4:** 0, 1, $@R_2$ | $R_2 \leftarrow 0$ | **4:** $i_4$, $i_2$, $@R_1$ | $R_1 \leftarrow N_2$ |
| **5:** 1, $@R_1$, $@R_2$ | $R_2 \leftarrow \overline{N}_1$ | | |
| **6:** $i_2$, $i_4$, $@R_2$ | $R_2 \leftarrow N_2$ | | |

Abb. 5: Die Auswirkungen der MIG-Optimierung auf die Anzahl der RRAMs und Anweisungen.

Unser vorgeschlagener Compiler findet zuerst eine effiziente Reihenfolge von Kandidatenknoten für die Berechnung und übersetzt dann die Knoten in einen MAJ-Befehl, indem den Operanden die kleinste Anzahl von RRAM-Schaltern und -Befehlen zugewiesen wird. Der Compiler berücksichtigt auch die Ursachen für ungleichmäßigen Schreibverkehr, der einige Schalter langfristig viel früher als andere verschleißen lässt, was die Lebensdauer erhöht, indem die Schreibvorgänge über den gesamten Arbeitsspeicher verteilt werden [Sh17a]. Dies ist besonders wichtig, da RRAM-Schalter eine begrenzte Schreibdauer haben, die im Entwurfsprozess berücksichtigt werden sollte.

Abb. 6 zeigt die Anzahl der erforderlichen RRAMs und Anweisungen des vorgeschlagenen befehlsbasierten Ansatzes für EPFL-Benchmarks[1]. Die Ergebnisse sind für naive PLiM-Implementierungen und Implementierungen nach nur der MIG-Optimierung und Kompilierung gezeigt. Die Ergebnisse zeigen Verbesserungen von 19,95% bzw. 61,4% in Bezug auf die Anzahl der Befehle und der RRAM-Schalter. Die Standardabweichung der Schreibvorgänge über die RRAM-Schalter ist in Abb. 7 dargestellt. Der Vergleich der Ergebnisse mit den naiven Implementierungen zeigt eine Verbesserung von 72,17% gegenüber allen Benchmarks.

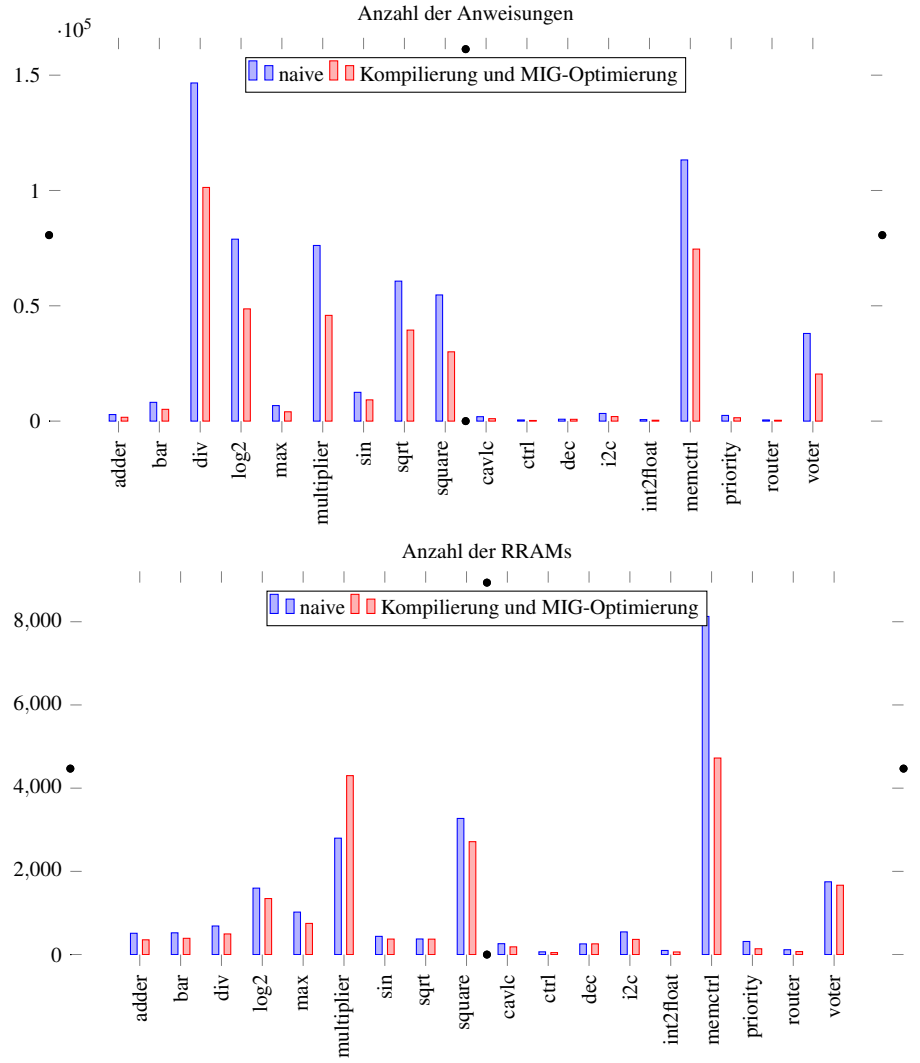



Abb. 6: Die Anzahl der Anweisungen und RRAMs, die von der PLiM-Computerarchitektur benötigt werden.

[1] http://lsi.epfl.ch/benchmarks

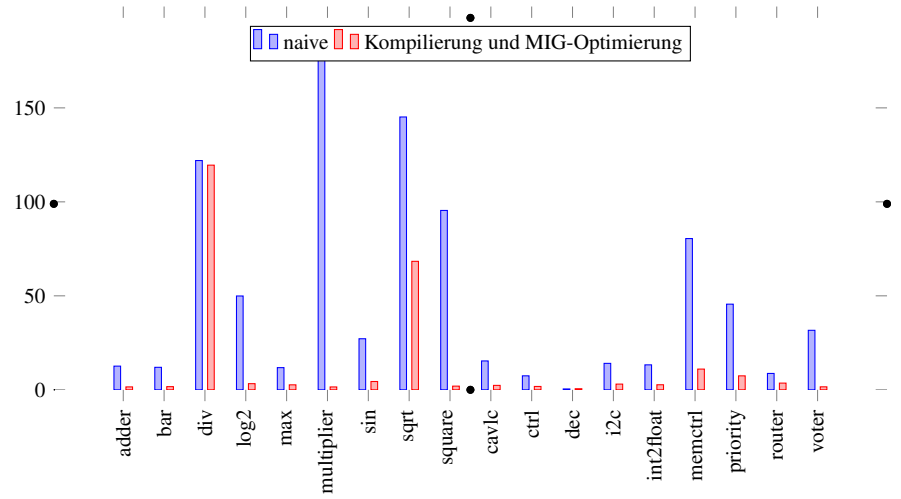



Abb. 7: Standardabweichung der Schreibvorgänge für die PLiM-Architektur.

## 5 Zusammenfassung

Diese Dissertation präsentiert einen umfassenden Ansatz für die Synthese von Logic-in-Memory-Schaltungen unter Verwendung der logischen Repräsentationen BDD, AIG und MIG. Der vorgestellte Ansatz führt die Realisierung der Logik-Grundelemente mit zwei grundlegenden Operationen ein, die durch RRAM-Schalter ermöglicht werden, und stellt Optimierungsalgorithmen und Entwurfsmethodiken für die Crossbar-Implementierung bereit. Die Arbeit schlägt auch einen automatischen Compiler für eine reguläre Logic-in-Memory-Computerarchitektur vor und verbessert die Ausführungskosten in Bezug auf Latenz, Fläche und Schreibgleichgewicht erheblich. Die Beiträge dieser Dissertation entwickeln einen umfassenden Rahmen für das vielversprechende Gebiet des Logic-in-Memory-Computing. Damit bietet sie eine solide Grundlage für dieses aufstrebende Computer-Paradigma, welches verschiedene Anwendungen ermöglicht, und schafft Möglichkeiten für die weitere Forschung und Entwicklung in den verschiedenen Bereichen programmierbarer Hardware und Architekturen von resistiven Speichern.

## Danksagung

Diese Arbeit entstand im Graduiertenkolleg System Design (SyDe) der Universität Bremen, das im Rahmen der Exzellenzinitiative finanziert wird.

## Literatur

[Bo10] Borghetti, J.; Snider, G.S.; Kuekes, P.J.; Yang, J.J.; Stewart, D.R.; Williams, R.S.: Memristive switches enable stateful logic operations via material implication. Nature, 464(7290):873–876, 2010.

[Ga16] Gaillardon, Pierre-Emmanuel; Amarù, Luca Gaetano; Siemon, Anne; Linn, Eike; Waser, Rainer; Chattopadhyay, Anupam; De Micheli, Giovanni: The Programmable Logic-in-Memory (PLiM) computer. In: Design, Automation & Test in Europe. S. 427–432, 2016.

[Kv14] Kvatinsky, S.; Belousov, D.; Liman, S.; Satat, G.; Wald, N.; Friedman, E.G.; Kolodny, A.; Weiser, U.C.: MAGIC – Memristor-Aided Logic. IEEE Trans. Circuits Syst. II, 61(11):895–899, 2014.

[Sh16] Shirinzadeh, Saeideh; Soeken, Mathias; Gaillardon, P.-E.; Drechsler, Rolf: Fast logic synthesis for RRAM-based in-memory computing using Majority-Inverter Graphs. In: Design, Automation & Test in Europe. S. 948–953, 2016.

[Sh17a] Shirinzadeh, Saeideh; Soeken, Mathias; Gaillardon, P.-E.; De Micheli, Giovanni; Drechsler, Rolf: Endurance management for resistive Logic-In-Memory computing architectures. In: Design, Automation & Test in Europ. S. 1092–1097, 2017.

[Sh17b] Shirinzadeh, Saeideh; Soeken, Mathias; Gaillardon, Pierre-Emmanuel; Drechsler, Rolf: Logic Synthesis for Majority Based In-Memory Computing. In (Vaidyanathan, Sundarapandian; Volos, Christos, Hrsg.): Advances in Memristors, Memristive Devices and Systems, S. 425–448. Springer International Publishing, 2017.

[Sh18] Shirinzadeh, S.; Soeken, M.; Gaillardon, P.-E.; Drechsler, R.: Logic Synthesis for RRAM-based In-Memory Computing. IEEE Trans. on CAD of Integrated Circuits and Systems, 37(7):1422–1435, 2018.

[So16] Soeken, Mathias; Shirinzadeh, Saeideh; Gaillardon, P.-E.; Amarù, Luca Gaetano; Drechsler, Rolf; De Micheli, Giovanni: An MIG-based compiler for programmable logic-in-memory architectures. In: Design Automation Conference. S. 117:1–117:6, 2016.

[So17] Soeken, Mathias; Gaillardon, P.-E.; Shirinzadeh, Saeideh; Drechsler, Rolf; De Micheli, Giovanni: A PLiM Computer for the Internet of Things. IEEE Computer, 50(6):35–40, 2017.

[SSD16] Shirinzadeh, Saeideh; Soeken, Mathias; Drechsler, Rolf: Multi-objective BDD optimization for RRAM based circuit design. In: IEEE International Symposium on Design and Diagnostics of Electronic Circuits and Systems. S. 46–51, 2016.

**Saeideh Shirinzadeh** erhielt ihren B.Sc. (2010) und M.Sc. (2012) in Elektrotechnik an der University of Guilan, Iran. Nach einer Tätigkeit als Dozentin war sie von 2014 bis 2018 Doktorandin an der Universität Bremen, Fachbereich Mathematik und Informatik, in der Arbeitsgruppe Rechnerarchitektur (Leitung: Prof. Dr. Rolf Drechsler). Sie verteidigte ihre Doktorarbeit mit dem Prädikat *summa cum laude* im Oktober 2018. Seit 2019 arbeitet sie als Postdoc-Forscherin am Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI). Ihre Forschungsinteressen konzentrieren sich auf Logiksynthese, In-Memory-Computing, Mehrziel-Optimierung und evolutionäre Berechnungen.

# Auf dem Weg zu bedeutungsvoller Teilhabe in der Technikgestaltung – Notizen zur Evaluation von Technologieerfahrungen autistischer Kinder[1]

Katta Spiel[2]

**Abstract:** Viele Technologien für autistische Kinder fokussieren ausschließlich auf medizinisch als solche definierten Defiziten. In deren Evaluation zählen dann auch vorwiegend extrinsische Ziele: zentrale Frage ist nicht, welche Erfahrungen ein Kind macht, sondern was es danach 'besser' kann. Durch sich langsam vermehrende Ansätze, die autistische Kinder in die Designprozesse von Technologien mit einbinden, die sich den Interessen der Kinder zuwenden, werden andere Evaluierungszugänge von Nöten. Im Bereich der Mensch-Maschine Interaktion gibt es qualitative Ansätze, die die Erfahrungen von Menschen in den Mittelpunkt stellen. Jedoch bauen diese wiederum auf die Empathie von Forscher*innen mit ihren Partizipant*innen auf. Durch die unterschiedlichen Wahrnehmungsprozesse von autistischen wie nicht-autistischen Menschen ist ein solches Vorgehen hier allerdings unzureichend. Diese Dissertation stellt sich der Problematik durch drei Herangehensweisen: Erstens werden die Erfahrungen von autistischen Kindern mit Technologien kritisch kontextualisiert und evaluiert; Zweitens wird dafür die notwendige Methodologie zu einer partizipativen Vorgehensweise erarbeitet; Drittens wird durch eine tiefgehende Analyse der mikro-ethischen Zusammenhänge kritisch auf partizipative Technologie-Forschung reflektiert. Der Beitrag argumentiert dafür, die Bedürfnisse von (insbesondere gesellschaftlich marginalisierten) Menschen im Zusammenhang mit Technologieentwicklung individualisiert und differenziert zu betrachten.

## 1 Einführung

In der Evaluation von Technologien für autistische Kinder legen Forscher*innen bisher selten Wert darauf, die Erfahrungen der Kinder mit einzubeziehen. Ebenso wie beim Design reduzieren sich Evaluierungen hier typischerweise auf ein medizinisch definiertes Interesse. Dahingehend finden sich Beispiele wie diagnostische Werkzeuge (z.B. [We12]), assistive Technologien für den täglichen Bedarf (z.B. Kommunikationsstützen [To12] oder visuelle Strukturhilfen [Hi10]). Andere zielen spezifisch auf Interventionen ab (bspw. [BPPS14] nach SCERTS[3]) oder untersuchen die potentiell therapeutisch wertvollen Effekte spielerischer Technologien (bspw. [FYR10] oder [VMJ12] für Topobos bzw. Reactable). Jedoch ist es eher ungewöhnlich, dass Technologien ausschließlich dafür gestaltet werden, dass autistische Kinder bedeutungsvolle Erfahrungen machen oder schlichtweg Spaß und Freude an Interaktionen haben[4].

[1] Evaluating Experiences of Autistic Children with Technologies in Co-Design

[2] TU Wien, katta@igw.tuwien.ac.at

[3] SC - Social Communication, ER - Emotional Regulation, TS - Transactional Support, see http://www.scerts.com

[4] Wobei eines der wenigen Ausnahmen in der Arbeit von [Pa05] liegt, welche besonderen Wert auf die positiven sensorischen Eindrücke von autistischen Kindern legt.

Die Mehrheit dieser Technologien folgt demnach in ihrer Evaluation den extrinsischen Motivationen, die schon das Design angetrieben haben[5]. In den letzten Jahren finden sich zwar mehr und mehr Projekte, die partizipativ mit autistischen Kindern Technologien gestalten, diese verbleiben jedoch oftmals ebenso in einem rein medizinisch orientierten Zugang zu Autismus. **Daher lässt sich eine Forschungslücke darin, partizipativ Technologien zu entwickeln und zu evaluieren, die die intrischen Interessen autistischer Kinder, deren ganzheitliches Wohlbefinden und ihre verkörperten Erfahrungen mit den Technologien berücksichtigen.**

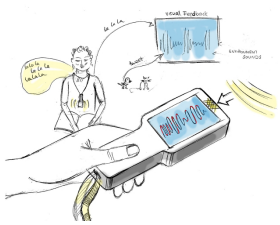
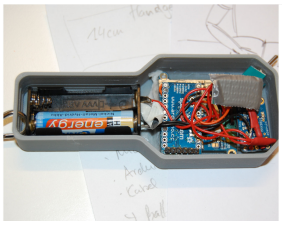
Methoden und Herangehensweisen aus dem Bereich des partizipativen Gestaltens lassen sich jedoch nur bedingt auf Evaluierungskontexte übertragen. Konstruktive Ansätze zu Technologieerfahrungen im Feld der Mensch-Maschine Interaktion verlassen sich auf Empathie als treibende Kraft für derartige Evaluationen (vgl. [MW07]). Da autistische Menschen jedoch die Welt inherent anders wahrnehmen als allistische[6] Forscher*innen [DJ13], ist es nicht so einfach möglich, sich 'in die Schuhe einer anderen Person' zu versetzen. Dies wird durch den üblicherweise gegebenen Altersunterschied zwischen erwachsenen Forscher*innen und autistischen Kindern zusätzlich verstärkt. **Diese Dissertation bietet dahingehend einen Rahmen dafür, strukturierte Prozesse durchzuführen, die unterschiedliche Blickwinkel auf die Erfahrung von autistischen Kindern mit partizipativ gestalteten Technologien untersuchen.**

Dabei ist es notwendig, auch die Perspektive der Kinder selbst direkt zu eruieren. Während für die Gestaltungsprozesse durchaus methodische Ansätze vorhanden sind (z.B. [BJ15, Ma17]), gibt es bedeutend weniger Forschung dazu, wie autistische Kinder in die Evaluation von Technologien aktiv mit eingebunden werden können. Auch wenn die Kommunikation zwischen allistischen und autistischen Menschen für beide Seiten von erhöhter Komplexität geprägt ist, werden die Erkenntnisse aus Evaluierungsprozessen robuster, je vielfältiger die eingeholten Perspektiven sind. **Dahingehend erarbeitet diese Dissertation einen methodischen Zugang für die Einbindung von neurodivergenten[7] Bevölkerungsgruppen in Evaluationsprozesse.**

Die Dissertation positioniert sich in den Bereichen Mensch-Maschine Interaktion sowie Partizipative Gestaltung mit Referenzen zu Critical Disability Studies. Folgende Ziele wurden verfolgt:
→ Erarbeitung eines Konzeptes zur Evaluierung der Erfahrungen von autistischen Kindern mit partizipativ gestalteten Technologien in ganzheitlicher, qualitativer Vorgehensweise
→ Methodische Untermauerung der Handlungsmacht autistischer Kinder und strukturelle Einbindung ihrer Perspektiven
→ Schaffung eines Rahmens in dem autistische Kinder aktiv die Interaktion mit Technologien mit definieren sowie deren Bedeutungen mit-konstruieren können

[5] Fr eine detaillierte Analyse dieses Sachverhaltes, siehe [Sp19].
[6] Der Begriff 'allistisch' steht für 'nicht autistisch' (nach [Ma03].
[7] Das Konzept von Neurodiversität geht von unterschiedlichen kognitiven Stilen aus anstatt manche von ihnen als medizinisch behandlungsbedürftig zu qualifizieren. Neurodivergente Menschen haben demnach einen anderen kognitiven Stil als neurotypische.

Diese Ziele wurden durch folgende Forschungsfragen weiter artikuliert:
1) Wie kann die Erfahrung von autistischen Kindern konzeptuell erfasst werden?
2) Wie können autistische Kinder aktiv in Evaluationsprozesse eingebunden werden und Deutungsmacht über ihre Interaktion mit Technologien zugesprochen bekommen?
3) Was sind die qualitativen Aspekte der Erfahrungen, die autistische Kinder in diesem Kontext machen?

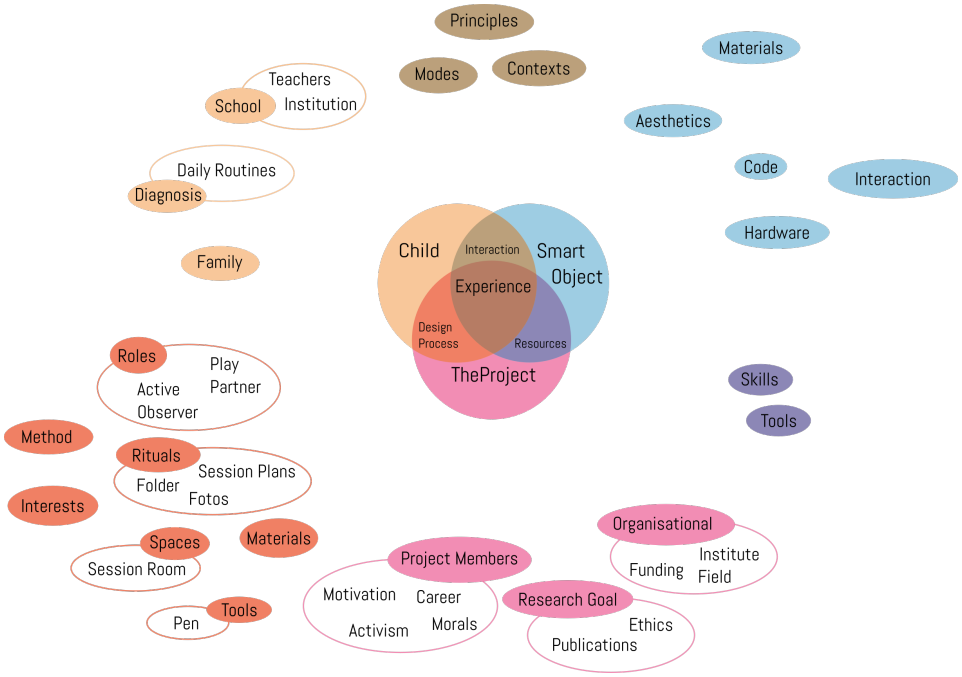

In der vollständigen Dissertation findet sich auch eine **kritische Übersicht über vorhandene Technologien** für autistische Kinder sowie deren Evaluation. Diese Zusammenfassung fokussiert sich auf das Konzept der *Critical Experience* sowie die *PEACE – Participatory Evaluation with Autistic ChildrEn* Methodik. Die Arbeit wird durch acht Fallstudien, die durch iterative, partizipative Design- und Evaluierungsprozesse im Rahmen des OutsideTheBox[8] Projektes zeigen, welche Implikationen das vorgeschlagene Vorgehen mit sich bringt. Diese Fallstudien bieten einzigartige, direkte Einblicke in die Aspekte, die für die Kinder von Bedeutung sind. Durch die Analyse von mikro-ethischen Entscheidungen innerhalb dieser Prozesse[9] erweitert die Arbeit auch das grundlegende Verständnis davon, was es heißt, in der partizipativen Gestaltung von Technologien mit marginalisierten Bevölkerungsgruppen ethisch zu handeln.

Die Arbeit wendet sich an Forscher*innen die mit Bevölkerungsgruppen arbeiten, die sehr verschieden von ihnen selbst sind, ebenso wie Entwickler*innen und Gestalter*innen von assistiven Technologien. Es wird für einen umsichtigen und kritisch informierten Zugang in der Zusammenarbeit mit marginalisierten Menschen argumentiert und gezeigt, wie dies erfolgreich gelingen kann.

## 2 Critical Experience

Ein kritisch eingegebettetes Evaluierungskonzept für die Erfahrungen von autistischen Kindern mit Technologien sollte *verschiedene Standpunkte* interpolieren und eine *flexible Datenakquise* ermöglichen. Diese Flexibilität sollte zudem zweifach sein: einerseits im Bezug auf die Datenquellen, um unterschiedliche Kommunikationsplattformen nach Präferenzen unterschiedlicher autistischer Kinder anzubieten, sowie andererseits im Bezug auf Aussagen, die im ersten Moment für Forscher*innen wenig Sinn geben [Fr12]. Zusätzlich sollte das Konzept den Lebenskontext der Kinder strukturell mit einbinden (ähnlich zu [FDG04]). Die sozialen und physischen Kontexte der Kinder, wie sie durch

[8] Gefördert vom österreichischen FWF. Siehe www.outsidethebox.at.
[9] Siehe hierzu auch [Sp18].




Eltern, Institutionen, Konzepte von Behinderung sowie wichtige Gegenstände abgebildet werden, müssen sorgsam in die Erkenntnisgewinnung miteingebunden werden.

Wenn dezidiert unterschiedliche Standpunkte gesammelt werden, vergrößert sich gleichzeitig das Risiko, widersprüchliche Positionen zu erhalten. Durch *Critical Experience* können Forscher*innen diesen diversen Zugängen offen entgegentreten (siehe auch [SG06]), indem sie alle, zumindest zu Anfangs, als gleichwertig betrachten. Um weiter die Robustheit der Erkenntnisgewinnung zu stärken, müssen Forscher*innen explizit argumentieren, warum sie bestimmten Standpunkten weniger Bedeutung beimessen als anderen.

Ausgehend vom konzeptuellen Standpunkt der Dissertation ergibt sich demnach eine theoretische Herausforderung: Die Integration unterschiedlicher Blickwinkel und Datenquellen in ein kohärentes Ganzes welches es erlaubt, Wissen über die Erfahrungen von autistischen Kindern mit Technologien zu konstruieren. Daher muss auch ein besonderer Augenmerk darauf gelegt werden *wer* welche Beiträge liefert. Akteur-Netzwerk Theorie (ANT) und Kritische Diskursanalyse (KDA) unterstützen diese Ziele.

Das oben abgebildete Schema illustriert einen Startpunkt für die Evaluation der technologischen Erfahrungen von autistischen Kindern nach *Critical Experience*. Die Beispielvorlage kann auf individuelle Umstände adaptiert werden. Die Kernpunkte jeglicher Evaluation bestehen aus Kind, technologischem Artefakt und den Forscher*innen, die die Erfahrung evaluieren, da deren Betrachtung per se fundamental die Erfahrung mit Technologien beeinflusst (siehe auch für die Beziehung zwischen Betrachtenden und Betrachtetem

[Ch01]). Die alleinige Erstellung eines Akteur-Netzwerkes reicht jedoch nicht dafür aus, zu verstehen wie sich einzelne Komponenten im Kontext zueinander verhalten. Für eine kritische Einbettung des Netzwerkes müssen Machtbeziehungen explizit analysiert und einzelne Akteure bedeutsam innerhalb des Netzwerkes verortet werden, wofür die Dissertation die Verwendung von KDA vorschlägt.

KDA wurde bereits erfolgreich angewandt um die Einbindung von Kinder in Interaktionsdesign zu beschreiben [IKK15]. Es konnte konkret festgestellt werden, dass Diskurse um Technologien und Kinder voller widersprüchlicher Aussagen sind, die gleichzeitig Kinder als Designpartner konzeptualisieren und ihre effektive Teilnahme an einem partizipativen Prozess verhindern. Ihre Einbindung ist einerseits geprägt durch Idealismus, andererseits problematisch in der Ausführung. Deswegen wird mit *Critical Experience* aktiv die Handlungsmacht der Kinder reflektiert, um ihren Beiträgen im Prozess der Bedeutungsgewinnung für Technologien Platz einzuräumen (siehe auch, für mehr Details [SFF17, Sp17a]). Dabei wird deren Teilnahme nicht erwartet oder angefordert, sondern jegliche Mitarbeit oder Verweigerung als valide interpretiert.

Nach der Erstellung eines schematischen Netzwerks, kann die weitere Analyse angegangen werden. Der Prozess für *Critical Experience* besteht aus fünf iterativen Schritten, welche auf einem initialen Akteur-Netzwerk (siehe oben) basieren:

1) *Diskurs und Kontext definieren*
Durch eine Kontextanalyse können Forscher*innen Akteure und Beziehungen identifizieren, die relevant für die Beantwortung ihrer Forschungsfragen sind. Beispielsweise könnte dies durch eine erweiterte Stakeholder-Analyse (vgl. [CF08]) geschehen, die auch nichtmenschliche Akteure einbezieht.

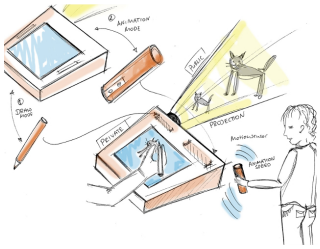
2) *Daten generieren, um Akteure und Beziehungen zu qualifizieren*
In diesem Schritt etablieren Forscher*innen geeignete Methoden zur Datengenerierung für individuelle Akteure und Beziehungen. Hierbei ist es wichtig, zuerst die Hauptakteure zu identifizieren und dann systematisch zu kontrollieren, dass diese in angemessener Weise in die Datenakquise eingebunden sind. Dadurch ist es wahrscheinlicher, dass ansonsten hinderliche Machtbeziehungen etwas umgangen werden können. Beispielsweise in Fällen in denen Eltern dazu tendieren, für ihr Kind zu antworten, können Forscher*innen zusätzliche Daten (wie etwa Zeichnungen) generieren, die die Meinung des Kindes direkter einfangen.

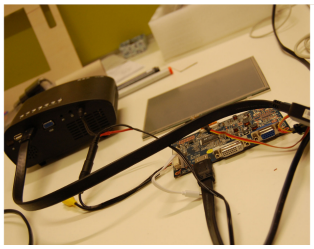
3) *Daten analysieren und Aussagen identifizieren*
Je nach Datenquelle führen Forscher*innen eine geeignete Datenanalyse durch und zu den Verbindungen sowie individuellen Akteuren im Akteur-Netzwerk Aussagen zu. In der Analyse bleibt das Konzept bewusst abstrakt, da unterschiedliche Daten verschieden analysiert werden können. Eine Aussage ist datengetrieben, wird aber interpretiert und abstrahiert. Beispielsweise kann eine leere Logdatei qualitativ verstanden und in die Aussage “Ich wurde nicht verwendet.” übersetzt werden. Für die Diskursanalyse im nächsten Schritt ist jede Aussage in Textform, kann allerdings auf eine Vielfalt auf Datenquellen wie Beobachtungen, Prototypenhistorien oder Interviewzitate zurückgeführt werden. Da es keine dezidierte Richtline dafür gibt, wann genug Aussagen identifiziert wurden, liegt

es im Ermessen der Forscher*innen, den Sättigungspunkt zu erkennen an dem neue Aussagen keinen weiteren Erkenntnisgewinn liefern.

4) *Aussagen kontextualisieren*
Anhand der gesammelten Aussagen analysieren Forscher*innen Kontext, Inhalt, diskursive Position und andere Eigenschaften individuell und im Gesamtzusammenhang. Dadurch können Widersprüche und fehlende Positionen identifizert werden. Beispielsweise kann es mittels Aussagen von Elternteilen deutlich werden, dass ein Geschwisterkind eine Technologie mehr nutzt als das autistische Kind für welches es intendiert war. Dies kann zur Folge haben, dass Forscher*innen nach Möglichkeit auch das Geschwisterkind in die Evaluation mit einbeziehen.

5) *Iteration über vorherige Schritte*
In diesem Schritt ergänzen Forscher*innen fehlende Akteure und Verbindungen im Akteur-Netzwerk (beispielsweise oben erwähntes Geschwisterkind). Danach werden die obigen vier Schritte iterativ wiederholt bis keine neuen Erkenntnisse mehr gewonnen werden können.






Der Ansatz hat das Potential, zu einer breiteren Diskussion über die Beschaffenheit von Erfahrungen in der Mensch-Maschine Interaktion beizutragen. Die pragmatische Perspektive wie sie von [MW07] vertreten wird bedeutete einen Schritt vorwärts im Bezug auf ein situiertes und nuanciertes Verständnis von Erfahrungen. *Critical Experience* geht jedoch bedeutend weiter indem es diese Erfahrungen als facettenreich und durch verschiedne Datenquellen als gleichzeitig extrinsisch und intrisisch motiviert konzipiert.

In der Dissertation wird die Vorgehensweise nach *Critical Experience* anhand von acht Fallstudien aus dem OutsideTheBox Projekt illustriert. Diese stellen anhand der Kombination von ANT und KDA detailliert dar, wie sich die Erfahrungen der Kinder konstituieren. Der Ansatz führt zu wertvollen Einblicken, welche sich aus unterschiedlichen Datenquellen speisen — vor allem aber aus den Perspektiven der Kinder selbst. Dadurch basiert die Evaluation der Erfahrungen von autistischen Kindern mit Technologien nicht nur auf der Empathie der Forscher*innen. Stattdessen wird eine Vielzahl an Blickwinkeln berücksichtigt und die aktive Einbindung der autistischen Kinder angestrebt.

## 3 PEACE – Participatory Evaluation with Autistic ChildrEn

Partizipative Evaluation als konzeptuelle Herangehensweise ist relativ unbekannt im Feld der Mensch-Maschine Interaktion. Die wenigen existierende Fallbeispiele beschäftigen sich mit kooperativen Arbeitskontexten [RRR95] und den Perspektiven von Patient*innen [KS12]. Während Bossen et al. feststellen, dass partizipative Designprojekte generell wenig evaluiert werden [BDI16], geht PEACE noch weiter, indem es die Partizipation mit in die Evaluation ausdehnt[10].

Hierfür gibt es jedoch kaum Methoden, die es erlauben, dass Partizipant*innen bedeutungsvoll an der Konstruktion der Bedeutung von Technologien teilhaben können. Insbesondere fr die Einbindung von Kindern existieren zwar Ansätze dafür, ihre Meinungen einzuholen, aber nicht dafür, sie direkt in die Entscheidungen über die Ziele und Methoden der Evaluation zu involvieren. Sie entscheiden weder wo noch wie Daten gesammelt werden und ihre Interessen werden nur selten berücksichtigt. Durch partizipativere Ansätze können Kinder dazu ermutigt werden über die gesammelten Daten (quantitativer wie qualitativer Form) aktiv zu reflektieren. Umgekehrt eröffnet dies zudem eine Möglichkeit für Forscher*innen, weiteres Wissen darüber zu erlangen, was die Kinder für essentiell halten.

Angefangen mit der Festlegung der Ziele für die Evaluation können autistische Kinder dadurch die vorgefertigten Erwartungen von Forscher*innen potentiell in Frage stellen insbesondere im Kontext von intendierten Nutzungsgruppen, Evaluationskriterien und methodischer Herangehensweise. In klassischen Evaluationen werden Methoden anhand einer Kombination von wissenschaftlicher Fragestellung und epistemologischer Grundlage der Forscher*innen ausgewählt. In partizipativen Evaluationen jedoch werden Methoden so ausgewählt, dass sie den Fähigkeiten der Partizipant*innen entsprechen und es möglich machen, dass die resultierenden Daten für alle Beteiligten bedeutungsvoll sind [Ni14]. Durch die Trennung von Zielen und Methoden der partizipativen Evaluation, können Partizipant*innen und Forscher*innen ihre Fragestellungen davon trennen, wie sie beantwortet werden können, da beide unterschiedliche Implikationen in Hinsicht auf Erkenntnisgewinn, Handlungsmacht und Teilhabe haben.

Die Methodologie von *PEACE* setzt sich grob aus drei Stufen zusammen, welche den Phasen für partizipative Evaluation der kanadischen Regierung folgen [CD16]. In der *Planungsphase* werden Ziele und Methoden festgelegt, in der *Implementierungsphase* Daten gesammelt und in der *Abschlussphase* die Resultate zusammen interpretiert und diskutiert.

Die Herangehensweise wird in der Dissertation anhand von vier Fallstudien mit autistischen Kindern in ihrer Realisierbarkeit kritisch überprüft. Besonders aufschlussreich sind hier die Fällen, in denen sich Kinder enthusiastisch die Methoden, die sie im Prozess erlernt haben, für ihre Alltagskommunikation aneignen. Dabei werden aber Erkenntnisse ebenso in Fällen erlangt, in denen eine partizipative Evaluation seitens der Kinder verweigert wird. Durch diese Fallstudien wird die Machbarkeit der aktiven Einbindung von autistischen Kindern in die Evaluation von Technologien, die sie partizipativ mitgestaltet haben, exemplarisch nachgewiesen. *PEACE* liefert in diesem Kontext als erstes Konzept

[10] Siehe für mehr Details auch [Sp17b].

die Rahmenbedingungen die es Forscher*innen erlauben, autistische Kinder so einzubinden, dass ihre Handlungsmacht, Bedürfnisse, Wünsche und Fähigkeiten in angemessener Weise berücksichtigt werden.

## 4 Zusammenfassung

Am Anfang der Dissertation stand es, drei grundsätzliche Beiträge zum Wissen über die Erfahrungen von autistischen Kindern mit Technologien zu liefern. Es wird nun überprüft, ob diese Anspruchshaltung gerechtfertigt ist.

In der Thesis wird grundlegend das Verständnis von technologisch getriebenen Erfahrungen durch den Augenmerk auf eine Gruppe von autistischen Kindern neu konzeptualisiert. Dadurch wird detailliert dargestellt, wie diese Erfahrungen über empathisches Verständnis hinaus, evaluiert werden können. Mit Akteur-Netzwerk Theorie und Kritischer Diskursanalyse als Grundlage bietet das *Critical Experience* Framework systematisch Einblick in die facettenreichen Perspektiven, die solche Erfahrungen erleben und beeinflussen. Die Thesis präsentiert die dazugehörigen Fallstudien, die die Erarbeitung von *Critical Experience* gestützt haben. Dahingehend wird nicht nur Wissen darüber generiert, wie autistische Kinder Technologien erfahren sondern auch eine Herangehensweise entwickelt, die es Forscher*innen generell erlaubt, Erfahrungen durch verschiedene menschliche wie nicht-menschliche Perspektiven kritisch und diskursiv zu analysieren.

In der Anwendung von *Critical Experience* wurde eine methodologische Lücke hinsichtlich der direkten Perspektive von autistischen Kindern deutlich. Folglich wird in der Thesis die Herangehensweise *PEACE* für partizipative Evalution mit autistischen Kindern entwickelt. *PEACE* ermöglicht es, die Bedeutung von Technologien, welche aus partizipativen Design Prozessen resultieren, zusammen mit Partizipant*innen zu eruieren. Mit den individuellen Fähigkeiten, Kenntnissen und Interessen eines Kindes als Startpunkt können Forscher*innen sich in ihrer privilegierten Position im Bezug auf Erkenntnisgewinn zurücknehmen und stattdessen den partizipativen Prozessen (und damit den Partizipant*innen selbst) Deutungshoheit zukommen lassen. Durch Fallstudien konnte gezeigt werden, dass somit nicht nur die Kindesperspektive besser erfasst werden konnte, sondern auch, dass sogar in der Veweigerung der Methode seitens eines Kindes wertvolle Einsichten erlangt werden können. Dies gilt insbesondere dann, wenn sie in die umfassende Evaluation mit *Critical Experience* eingebunden werden. **Dadurch wurde erfolgreich der Versuch unternommen, in der Evaluation von Technologien rigoros Platz für autistische Kinder und deren Perspektiven zu schaffen.**

## Literaturverzeichnis

[BDI16] Bossen, Claus; Dindler, Christian; Iversen, Ole Sejer: Evaluation in Participatory Design: A Literature Survey. In: PDC '16. ACM, New York, NY, USA, S. 151–160, 2016.

[BJ15] Benton, Laura; Johnson, Hilary: Widening participation in technology design: A review of the involvement of children with special educational needs and disabilities. International Journal of Child-Computer Interaction, 34:23–40, 2015.

[BPPS14] Bernardini, Sara; Porayska-Pomsta, Kaka; Smith, Tim J.: ECHOES: An intelligent serious game for fostering social communication in children with autism. Information Sciences, 264:41–60, 2014.

[CD16] CDC Agency for Toxic Substances and Disease Registry – Committee on Community Engagement: , Principles of Community Engagement, 2011/Accessed July 14, 2016.

[CF08] Carmien, Stefan Parry; Fischer, Gerhard: Design, Adoption, and Assessment of a Socio-technical Environment Supporting Independence for Persons with Cognitive Disabilities. In: CHI'08. ACM, New York, NY, USA, S. 597–606, 2008.

[Ch01] Chataway, Cynthia J: Negotiating the Observer-observed relationship. From subjects to subjectivities: A handbook of interpretive and participatory methods, S. 239–255, 2001.

[DJ13] De Jaegher, Hanne: Embodiment and sense-making in autism. Frontiers in Integrative Neuroscience, 7:15, 2013.

[FDG04] Forlizzi, Jodi; DiSalvo, Carl; Gemperle, Francine: Assistive Robotics and an Ecology of Elders Living Independently in Their Homes. Hum.-Comput. Interact., 19(1):25–59, Juni 2004.

[Fr12] Frauenberger, Christopher; Good, Judith; Keay-Bright, Wendy; Pain, Helen: Interpreting Input from Children: A Designerly Approach. In: CHI '12. ACM, New York, NY, USA, S. 2377–2386, 2012.

[FYR10] Farr, William; Yuill, Nicola; Raffle, Hayes: Social benefits of a tangible user interface for children with Autistic Spectrum Conditions. Autism, 14(3):237–252, 2010.

[Hi10] Hirano, Sen H.; Yeganyan, Michael T.; Marcu, Gabriela; Nguyen, David H.; Boyd, Lou Anne; Hayes, Gillian R.: vSked: evaluation of a system to support classroom activities for children with autism. In: CHI '10. ACM, Atlanta, USA, S. 1633–1642, 2010.

[IKK15] Iivari, Netta; Kinnula, Marianne; Kuure, Leena: With best intentionsa Foucauldian examination on childrens genuine participation in ICT design. Information Technology & People, 2015.

[KS12] Kusunoki, Diana; Sarcevic, Aleksandra: Applying Participatory Design Theory to Designing Evaluation Methods. In: CHI '12 Extended Abstracts on Human Factors in Computing Systems. CHI EA '12, ACM, New York, NY, USA, S. 1895–1900, 2012.

[Ma03] Main, Andrew: , allism: an introduction to a little-known condition, 2003.

[Ma17] Malinverni, Laura; Mora-Guiard, Joan; Padillo, Vanesa; Valero, Lilia; Hervs, Amaia; Pares, Narcis: An inclusive design approach for developing video games for children with Autism Spectrum Disorder. Computers in Human Behavior, 71:535–549, 2017.

[MW07] McCarthy, John; Wright, Peter: Technology as Experience. MIT Press, August 2007.

[Ni14] Nind, Melanie: What is inclusive research? A&C Black, 2014.

[Pa05] Pares, Narcis; Masri, Paul; van Wolferen, Gerard; Creed, Chris: Achieving dialogue with children with severe autism in an adaptive multisensory interaction: the "MEDIATE"project. IEEE Trans. on Visualization and Computer Graphics, 11(6):734–743, 2005.

[RRR95] Ross, Susi; Ramage, Magnus; Rogers, Yvonne: PETRA: participatory evaluation through redesign and analysis. Interacting with Computers, 7(4):335–360, 1995.

[SFF17] Spiel, Katta; Frauenberger, Christopher; Fitzpatrick, Geraldine: Experiences of autistic children with technologies. IJCCI, 11:50–61, 2017.

[SG06] Sengers, Phoebe; Gaver, Bill: Staying Open to Interpretation: Engaging Multiple Meanings in Design and Evaluation. In: DIS '06. ACM, New York, USA, S. 99–108, 2006.

[Sp17a] Spiel, Katta; Frauenberger, Christopher; Hornecker, Eva; Fitzpatrick, Geraldine: When Empathy Is Not Enough: Assessing the Experiences of Autistic Children with Technologies. In: CHI '17. ACM, New York, NY, USA, S. 2853–2864, 2017.

[Sp17b] Spiel, Katta; Malinverni, Laura; Good, Judith; Frauenberger, Christopher: Participatory Evaluation with Autistic Children. In: CHI '17. ACM, New York, NY, USA, S. 5755–5766, 2017.

[Sp18] Spiel, Katta; Brulé, Emeline; Frauenberger, Christopher; Bailly, Gilles; Fitzpatrick, Geraldine: Micro-ethics for Participatory Design with Marginalised Children. In: Proceedings of the 15th Participatory Design Conference: Full Papers - Volume 1. PDC '18, ACM, New York, NY, USA, S. 17:1–17:12, 2018.

[Sp19] Spiel, Katta; Frauenberger, Christopher; Keyes, Os; Fitzpatrick, Geraldine: Agency of Autistic Children in Technology Research. Under Review., 2019.

[To12] Torii, I; Ohtani, K.; Shirahama, N.; Niwa, T.; Ishii, N.: Voice output communication aid application for personal digital assistant for autistic children. In: 2012 IEEE/ACIS 11th Conference on Computer and Information Science (ICIS). S. 329–333, Mai 2012.

[VMJ12] Villafuerte, Lilia; Markova, Milena; Jorda, Sergi: Acquisition of Social Abilities Through Musical Tangible User Interface: Children with Autism Spectrum Condition and the Reactable. In: CHI EA '12. ACM, New York, NY, USA, S. 745–760, 2012.

[We12] Westeyn, Tracy L.; Abowd, Gregory D.; Starner, Thad E.; Johnson, Jeremy M.; Presti, Peter W.; Weaver, Kimberly A.: Monitoring childrens developmental progress using augmented toys and activity recognition. Pers. and Ubiqu. Comp., 16(2):169–191, 2012.

**Katta Spiel** ist derzeit Post-Doc Forscher*in an der KU Leuven und der Universität Wien. Die Dissertation entstand im Rahmen von zwei FWF-geförderten Projekten an der TU Wien: OutsideTheBox sowie Social Play Technologies. In beiden wurde mit autistischen (und allistischen) Kindern partizipativ die Gestaltung von Technologien umgesetzt. Davor hat Katta zwei grundständige Studien (Bachelor) in Medienkultur und Mediensysteme sowie einen Master im Bereich Medieninformatik an der Bauhaus-Universität Weimar abgeschlossen. In aktuellen Forschungszusammenhängen profiliert sich Katta durch kritische, teilhabende aber auch spielerische Technologie-Forschung mit marginalisierten Bevölkerungsgruppen. Durch diese Arbeit entstehen nicht nur neuartige Innovationen sondern es werden auch die Machtzusammenhänge aktueller Technologien praktisch wie kritisch hinterfragt. Neben der Forschung ist Katta auch Textilgestalterin (v.a. im Bereich Strickwaren), spielt Roller Derby und war von 2009-2014 Stadträt*in in Weimar.

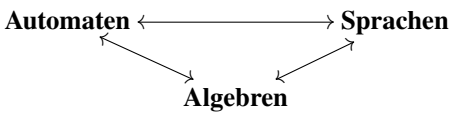
# Ein kategorientheoretischer Zugang zur algebraischen Sprachtheorie

Henning Urbat[1]

**Abstract:** In der algebraischen Theorie der formalen Sprachen werden Automaten und ihre Sprachen klassifiziert, indem man ihnen algebraische Strukturen zuordnet. Der algebraische Ansatz wurde für verschiedene Typen von Sprachen erfolgreich umgesetzt und führte zu einer Vielzahl von wichtigen Ergebnissen (z.B. Entscheidbarkeitsaussagen), die häufig auf strukturell sehr ähnlichen Ideen basieren. In der Dissertation [Ur18] wird nachgewiesen, dass zahlreiche Schlüsselkonzepte der algebraischen Sprachtheorie einen kategorientheoretischen Hintergrund haben. Die Hauptidee des kategoriellen Ansatzes besteht zum einen darin, formale Sprachen und die ihnen zugeordeten algebraischen Strukturen durch Monaden auf einer algebraischen Kategorie zu modellieren, und zum anderen in einer Interpretation der algebraischen Spracherkennung durch Betrachtung von prädualen Kategorien. Auf diese Weise kann gezeigt werden, dass zentrale Elemente der algebraischen Sprachtheorie einen generischen Charakter haben und strukturell unabhängig vom Sprachtyp sind.

## 1 Einleitung

Automaten als formale Modelle zustandsbasierter Systeme gehören zu den etabliertesten Werkzeugen der Theoretischen Informatik. Sie existieren in Hunderten Varianten und haben vielfältige praktische Anwendungen, etwa im Compilerbau, in der Künstlichen Intelligenz und in der Verifikation sicherheitskritischer Systeme. Die Untersuchung von Automaten als mathematische Strukturen führt zu tiefliegenden Ergebnissen und überraschenden Querverbindungen, die häufig auf der Interpretation von automatentheoretischen Fragestellungen aus einem algebraischen oder topologischen Blickwinkel basieren. Die Grundidee des algebraischen Zugangs zur Automatentheorie besteht darin, mathematische Beziehungen zwischen Automaten, den von ihnen repräsentierten Sprachen, und geeigneten algebraischen Strukturen herzustellen, wie im folgenden Diagramm angedeutet:




Auf diese Weise erhalten mächtige algebraische Methoden, etwa aus der Halbgruppen- und Gruppentheorie, Einzug in die Welt der Automaten und formalen Sprachen. Das im obigen Diagramm beschriebene Prinzip ist sehr allgemein und auf viele Klassen von Sprachen (z.B. reguläre Sprachen, $\omega$-reguläre Sprachen oder Baumsprachen) anwendbar.

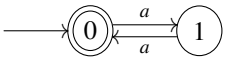
[1] Friedrich-Alexander-Universität Erlangen-Nürnberg, Lehrstuhl Informatik 8, henning.urbat@fau.de
Englischer Titel der Dissertation: “A Categorical Approach to Algebraic Language Theory”

In der Dissertation [Ur18] und den zugrunde liegenden Arbeiten [Ad14, Ad15, AMU15, Ch16, CU16, Ur17] wird ein kategorientheoretischer Zugang zur algebraischen Sprachtheorie entwickelt, der zahlreiche Elemente dieser Theorie *generisch* – also unabhängig vom konkreten Sprachtyp – präsentiert. Ziel dabei ist (i) eine vereinheitlichende Sicht auf Konzepte, die zuvor für verschiedene Sprachtypen separat untersucht wurden, und (ii) eine modulares und robustes Rahmenwerk, das generische und anwendungsspezifische Aspekte der Theorie voneinander isoliert und damit die Herleitung neuer Ergebnisse vereinfacht. Zur Motivation werden im folgenden Abschnitt zunächst einige klassische Ergebnisse der algebraischen Sprachtheorie diskutiert. Anschließend werden die in der Arbeit entwickelten kategoriellen Methoden mit einigen ihrer Anwendungen vorgestellt.

## 2 Hintergrund: Algebraische Spracherkennung

Der Ausgangspunkt der algebraischen Sprachtheorie liegt in der Beobachtung, dass sich reguläre Sprachen nicht nur durch die klassischen endlichen Automaten, sondern auch rein algebraisch durch Monoide beschreiben lassen. Diese Korrespondenz basiert auf dem fundamentalen Konzept der *algebraischen Erkennung* von Sprachen: Eine Sprache $L \subseteq \Sigma^*$ über einem Alphabet $\Sigma$ wird von einem Monoid $M$ erkannt, wenn ein Monoidmorphismus $h\colon \Sigma^* \to M$ und eine Teilmenge $S \subseteq M$ mit $L = h^{-1}[S]$ existiert. Dabei bezeichnet $h^{-1}[S] = \{\, w \in \Sigma^* : h(w) \in S \,\}$ das Urbild von $S$ bzgl. $h$.

**Beispiel.** Zur Illustration betrachten wir die Sprache $L = (aa)^*$ aller Wörter gerader Länge über dem Alphabet $\Sigma = \{a\}$. Sie wird von dem abgebildeten endlichen Automaten erkannt. Um diese Sprache algebraisch zu erfassen, verwendet man das additive Monoid $\mathbb{Z}_2 = \{0,1\}$ der ganzen Zahlen modulo 2 und den Monoidmorphismus $h\colon \{a\}^* \to \mathbb{Z}_2$, der ein Wort auf seine Länge modulo 2 abbildet. Dann ist $L = h^{-1}[\{0\}]$, d.h. $L$ wird vom Monoid $\mathbb{Z}_2$ erkannt.




Endliche Automaten und endliche Monoide sind in Bezug auf die Erkennung von Sprachen äquivalente Strukturen, wie das folgende klassische Ergebnis zeigt:

**Satz** (Myhill-Nerode)**.** *Eine Sprache ist genau dann regulär, wenn sie von einem endlichen Monoid erkannt wird.*

Das in der Einleitung beschriebene Diagramm lässt sich also wie folgt instanziieren:


Die Charakterisierung der regulären Sprachen durch Monoide ermöglicht es, für wichtige Konzepte aus der Theorie der endlichen Automaten ein algebraisches Gegenstück zu identifizieren. Das betrifft etwa die Existenz minimaler Strukturen: Genau wie jede reguläre Sprache einen eindeutigen Minimalautomaten hat, gibt es auch ein eindeutiges minimales Monoid, das sie erkennt – das *syntaktische Monoid*. Im obigen Beispiel ist der abgebildete Automat der Minimalautomat der Sprache $(aa)^*$, und $\mathbb{Z}_2$ ist ihr syntaktisches Monoid.

Eines der Hauptziele des algebraischen Zugangs zur Automatentheorie ist die Klassifizierung von Eigenschaften regulärer Sprachen durch Eigenschaften ihrer syntaktischen Monoide. Das bekannteste Ergebnis dieser Art ist der Satz von Schützenberger [Sc65]. Er betrachtet die Klasse der *sternfreien Sprachen*, also Sprachen, die sich durch einen regulären Ausdruck wie $(\overline{a}+b)\overline{ab}$ beschreiben lassen, der den Komplementoperator $\overline{(-)}$, aber keinen Kleene-Stern $(-)^*$ verwenden darf. Schützenbergers Satz besagt, dass eine reguläre Sprache genau dann sternfrei ist, wenn ihr syntaktisches Monoid *aperiodisch* ist, d.h. wenn es die Gleichung $x^{n+1} = x^n$ für hinreichend großes $n$ erfüllt. Äquivalent dazu ist die Aussage, dass die eindeutige idempotente Potenz $x^\omega$ jedes Elements $x$ die Gleichung $x^\omega = x \cdot x^\omega$ erfüllt. In unserem obigen Beispiel sieht man leicht, dass das Monoid $\mathbb{Z}_2$ nicht aperiodisch ist. Folglich ist die Sprache $(aa)^*$ nicht sternfrei. Da das syntaktische Monoid effektiv aus dem Minimalautomaten einer gegebenen Sprache berechnet werden kann, impliziert der Satz von Schützenberger, dass Sternfreiheit eine entscheidbare Eigenschaft regulärer Sprachen ist. Diese Entscheidbarkeitsaussage wäre unter alleiniger Verwendung von Standardergebnissen über Automaten und reguläre Ausdrücke schwierig zu beweisen, und illustriert daher die Relevanz algebraischer Methoden in der Automatentheorie.

Neben den sternfreien Sprachen gibt es unzählige weitere Klassen von regulären Sprachen, die im Stil von Schützenbergers Satz durch Eigenschaften ihrer syntaktischen Monoide charakterisierbar sind. Um einen systematischen Zugang zu derartigen Korrespondenzergebnissen zu erhalten, führte Eilenberg [Ei76] zwei fundamentale Konzepte ein: *Varietäten von regulären Sprachen* und *Pseudovarietäten von Monoiden*. Eine Varietät von regulären Sprachen ist eine Klasse $\mathscr{V}$ von regulären Sprachen, die unter den booleschen Operationen, Ableitungen und homomorphen Urbildern abgeschlossen ist. Das heißt, (i) für alle Sprachen $K, L \subseteq \Sigma^*$ in $\mathscr{V}$ ist $\emptyset, \Sigma^*, K \cup L, K \cap L, \Sigma^* \setminus L \in \mathscr{V}$, (ii) für alle Sprachen $L \subseteq \Sigma^*$ in $\mathscr{V}$ und alle Wörter $y \in \Sigma^*$ ist

$$y^{-1}L = \{x \in \Sigma^* : yx \in L\} \in \mathscr{V} \quad \text{und} \quad Ly^{-1} = \{x \in \Sigma^* : xy \in L\} \in \mathscr{V},$$

und (iii) für alle Sprachen $L \subseteq \Sigma^*$ in $\mathscr{V}$ und alle Monoidmorphismen $h\colon \Delta^* \to \Sigma^*$ ist $h^{-1}[L] \in \mathscr{V}$. Eine Pseudovarietät von Monoiden ist eine Klasse $\mathbb{V}$ von endlichen Monoiden, die unter homomorphen Bildern, Untermonoiden und endlichen Produkten abgeschlossen ist. Die Beziehung zwischen diesen Konzepten wird durch den folgenden Satz hergestellt:

**Satz** (Eilenberg-Korrespondenz)**.** *Varietäten von regulären Sprachen stehen in bijektiver Korrespondenz zu Pseudovarietäten von Monoiden.*

Eilenbergs Satz liefert eine generische Beziehung zwischen Eigenschaften von Sprachen und Eigenschaften von Monoiden. Beispielsweise bilden die sternfreien Sprachen eine Varietät von regulären Sprachen, und die aperiodischen endlichen Monoide bilden eine Pseudovarietät von Monoiden, und somit kann der Satz von Schützenberger (ebenso wie viele verwandte Ergebnisse) als Instanz der Eilenberg-Korrespondenz interpretiert werden.

Eine wichtige Ergänzung zu Eilenbergs Satz ist die von Reiterman [Re82] bewiesene modelltheoretische Charakterisierung von Pseudovarietäten: diese entsprechen genau den Klassen von endlichen Monoiden, die durch *proendliche Gleichungen* (eine topologische Verallgemeinerung von klassischen Gleichungen zwischen Termen) axiomatisierbar sind.

Die Identität $x^\omega = x \cdot x^\omega$, die aperiodische endliche Monoide definiert, ist ein Beispiel für eine proendliche Gleichung. Reitermans Satz ist nicht spezifisch für Monoide, sondern gilt allgemeiner für Pseudovarietäten endlicher Algebren über einer finitären Signatur.

Die fundamentalen Ergebnisse von Eilenberg und Reiterman stellen eine enge Verbindung zwischen formalen Sprachen, algebraischen Strukturen und proendlichen Gleichungstheorien her, und gehören zu den Grundpfeilern der algebraischen Theorie regulärer Sprachen. In den vergangenen Jahrzehnten wurden zahlreiche Verallgemeinerungen und Erweiterungen des Konzeptes der algebraischen Spracherkennung und insbesondere der Eilenberg-Reiterman-Theorie erforscht, die sich in zwei Richtungen klassifizieren lassen:

(I) Trotz der Allgemeinheit von Eilenbergs Satz wurde schnell realisiert, dass viele interessante Klassen von regulären Sprachen keine Varietäten bilden, weil einige der erforderlichen Abschlusseigenschaften nicht erfüllt sind. Daher wurden diverse Abschwächungen des Varietätenkonzeptes untersucht. Auf der algebraischen Seite erfordert das die Betrachtung von Monoiden mit zusätzlicher Struktur. Beispielsweise betrachtete Pin [Pi95] die Erkennung von Sprachen durch *geordnete* Monoide und bewies eine Korrespondenz zwischen *positiven Varietäten von regulären Sprachen* (die nicht notwendig unter Komplement abgeschlossenen sind) und Pseudovarietäten von geordneten Monoiden.

(II) Eine weitere wichtige Forschungsrichtung befasst sich mit der Erweiterung der klassischen algebraischen Sprachtheorie, die sich auf reguläre Sprachen endlicher Wörter bezieht, auf andere Typen von formalen Sprachen. Hierfür ist es nicht länger ausreichend, Monoide als erkennende algebraische Strukturen zu verwenden. Beispielsweise können $\omega$-reguläre Sprachen (d.h. die von Büchi-Automaten repräsentierten Sprachen unendlicher Wörter) algebraisch durch *$\omega$-Halbgruppen* beschrieben werden, eine Verallgemeinerung von Monoiden mit einer unendlichen Multiplikation. Darüber hinaus existieren algebraische Zugänge für zahlreiche weitere Sprachtypen, darunter rationale Potenzreihen, Sprachen von Wörtern über linearen Ordnungen, Baumsprachen und Kostenfunktionen.

Alle oben beschriebenen Erweiterungen der Eilenberg-Reiterman-Theorie folgen dem Pfad der klassischen algebraischen Theorie regulärer Sprachen und operieren in fünf Schritten:

(1) Identifiziere eine algebraische Theorie T, so dass die betrachteten Sprachen genau den von endlichen T-Algebren erkannten Sprachen entsprechen.

(2) Untersuche die Existenz und Konstruktion von *syntaktischen T-Algebren*, also den minimalen algebraischen Erkennern von Sprachen.

(3) Führe das Konzept einer *Varietät von Sprachen* ein, also einer Klasse von Sprachen, die unter einer Teilmenge der booleschen Operationen, einem geeigneten Konzept von *Ableitungen* und Urbildern unter T-Algebra-Morphismen abgeschlossen ist.

(4) Führe das Konzept einer *Pseudovarietät von T-Algebren* ein, also einer Klasse von endlichen T-Algebren mit geeigneten Abschlusseigenschaften.

(5) Beweise eine Eilenberg-Korrespondenz zwischen Varietäten von Sprachen und Pseudovarietäten von T-Algebren.

Die Implementierung dieser Schritte erfordert jeweils die Adaption von Definitionen, Konstruktionen und Beweisen aus dem klassischen Fall der regulären Sprachen und Monoide. In der Folge erhält man eine Vielzahl von Ergebnissen, die strukturell sehr ähnliche Aussagen für verschiedene Sprachtypen etablieren und oft einen Ad-hoc-Charakter haben. Zum Beispiel existieren in der Literatur mehr als 20 Varianten von Eilenbergs Varietätensatz. Diese Situation motiviert die Suche nach einem allgemeineren Zugang zur algebraischen Sprachtheorie, mit der Zielsetzung, zuvor separate Ergebnisse unter ein gemeinsames Dach zu stellen und als Spezialfälle generischer Prinzipien zu interpretieren. In unserer Arbeit wird dies durch Verwendung kategorientheoretischer Methoden erreicht.

## 3 Kategorientheoretische Perspektive

In diesem Abschnitt stellen wir die Kernelemente unseres kategorientheoretischen Zugangs zur algebraischen Sprachtheorie vor. Seine zentrale Erkenntnis besteht in der Beobachtung, dass die im vorherigen Abschnitt beschriebenen Schritte (1)–(5) wesentlich vereinfacht oder sogar vollständig automatisiert werden können. Der kategorielle Ansatz wird durch die folgende "Gleichung" zusammengefasst:

**Algebraische Sprachtheorie = Monaden + Unäre Präsentationen + Dualität.**

Im Folgenden beschreiben wir genauer, wie diese Konzepte verwendet werden.

### 3.1 Monaden und Sprachen

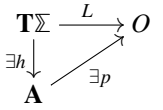
Die Voraussetzung für einen algebraischen Zugang zu einem beliebigen Typ von Sprachen ist eine Beschreibung dieser Sprachen durch geeignete algebraische Strukturen. In unserem Rahmenwerk verwenden wir hierfür das Konzept einer *Monade*, ein etablierter Formalismus der Kategorientheorie, mit dem man algebraische Strukturen abstrakt (unter Weglassung syntaktischer Konzepte wie Terme und Operationen) erfassen kann. Beispielsweise sind Monoide genau die Algebren der Monade $\mathbf{T}\Sigma = \Sigma^*$ auf **Set**, der Kategorie der Mengen, die jeder Menge $\Sigma$ das freie Monoid $\Sigma^*$ zuordnet. Allgemeiner betrachten wir eine beliebige Grundkategorie $\mathscr{D}$ von (evtl. geordneten, mehrsortigen) algebraischen Strukturen wie Mengen, halbordneten Mengen oder Vektorräumen, sowie eine Monade $\mathbf{T}$ auf der Kategorie $\mathscr{D}$. Sprachen werden als Morphismen $L\colon \mathbf{T}\mathbb{\Sigma} \to O$ in $\mathscr{D}$ modelliert, wobei $\mathbb{\Sigma}$ und $O$ Objekte von $\mathscr{D}$ sind, die Eingaben und Ausgaben repräsentieren. Auf diese Weise lassen sich zahlreiche Typen von formalen Sprachen kategoriell beschreiben, z.B.:

(a) Für die klassischen regulären Sprachen wählt man die Monoid-Monade $\mathbf{T}\Sigma = \Sigma^*$ auf $\mathscr{D} = \mathbf{Set}$ und das Ausgabeobjekt $O = \{0,1\}$.

(b) Rationale Potenzreihen über einem Körper $\mathbb{K}$ werden durch die Monade $\mathbf{T}$ auf der Kategorie $\mathscr{D}$ aller $\mathbb{K}$-Vektorräume repräsentiert, die freie $\mathbb{K}$-Algebren konstruiert. Als Ein- bzw. Ausgabeobjekt wählt man $\mathbb{\Sigma} = \mathbb{K}^{\Sigma}$ für ein endliches Alphabet $\Sigma$ und $O = \mathbb{K}$. Weil die freie $\mathbb{K}$-Algebra $\mathbf{T}\mathbb{\Sigma}$ vom Vektorraum mit Basis $\Sigma^*$ getragen wird, entspricht eine lineare Abbildung $L\colon \mathbf{T}\mathbb{\Sigma} \to \mathbb{K}$ einer Potenzreihe $L_0\colon \Sigma^* \to \mathbb{K}$.

(c) Für $\omega$-reguläre Sprachen betrachtet man die $\omega$-Halbgruppen-Monade $\mathbf{T}(\Sigma,\Gamma) = (\Sigma^+, \Sigma^\omega + \Sigma^* \times \Gamma)$ auf der Kategorie $\mathscr{D} = \mathbf{Set}^2$ der zweisortigen Mengen, zusammen mit $\mathbb{\Sigma} = (\Sigma, \emptyset)$ für ein endliches Alphabet $\Sigma$ und $O = (\{0,1\}, \{0,1\})$.

Die Interpretation von Sprachen als Morphismen erlaubt die Einführung eines allgemeinen Konzeptes algebraischer Spracherkennung: eine Sprache $L\colon \mathbf{T}\mathbb{\Sigma} \to O$ ist $\mathbf{T}$-*erkennbar*, wenn sie durch eine endliche $\mathbf{T}$-Algebra $\mathbf{A}$ faktorisiert (siehe Abbildung rechts). Dieses Konzept umfasst u.a. die Erkennung regulärer Sprachen durch Monoide, die Erkennung $\omega$-regulärer Sprachen durch $\omega$-Halbgruppen, und die Erkennung von Potenzreihen durch $\mathbb{K}$-Algebren.




Die Verwendung von Monaden zur generischen Modellierung algebraischer Spracherkennung wurde zuerst von Bojańczyk [Bo15] für den sehr eingeschränkten Fall $\mathscr{D} = \mathbf{Set}$ untersucht. Der Ansatz von *op. cit.* basiert auf speziellen Eigenschaften dieser Grundkategorie, insbesondere der Tatsache, dass alle Monaden durch Operationen und Gleichungen präsentierbar sind. Die Verallgemeinerung auf beliebige Grundkategorien $\mathscr{D}$ wie im oben beschriebenen Rahmenwerk ist daher nichttrivial und erfordert neue Techniken.

### 3.2 Proendliche Monaden

In der algebraischen Theorie regulärer Sprachen ist es oft hilfreich, eine topologische Perspektive einzunehmen. Das wichtigste Werkzeug in diesem Zusammenhang ist der Stone-Raum $\widehat{\Sigma^*}$ aller *proendlichen Wörter* über dem Alphabet $\Sigma$, der als inverser Limes aller endlichen Quotientenmonoide des freien Monoids $\Sigma^*$ konstruiert wird. Reguläre Sprachen entsprechen genau den zugleich abgeschlossenen und offenen Teilmengen von $\widehat{\Sigma^*}$. Eines der Kernkonzepte unseres kategorientheoretischen Ansatzes ist die *proendliche Monade* $\widehat{\mathbf{T}}$, die der Monade $\mathbf{T}$ zugeordnet wird und die Konstruktion des Raumes der proendlichen Wörter verallgemeinert. Die Monade $\widehat{\mathbf{T}}$ "lebt" in der Kategorie $\widehat{\mathscr{D}}$, die als freie Vervollständigung der Kategorie der endlichen $\mathscr{D}$-Algebren unter inversen Limites entsteht. Beispielsweise ist $\widehat{\mathbf{Set}}$ die Kategorie **Stone** der Stone-Räume, und für die Monoid-Monade $\mathbf{T}\Sigma = \Sigma^*$ auf **Set** ist die proendliche Monade gegeben durch $\widehat{\mathbf{T}}\Sigma = \widehat{\Sigma^*}$ auf **Stone**.

Als erste Anwendung der proendlichen Monade ergibt sich eine kategorielle Verallgemeinerung des Satzes von Reiterman. Hierfür wird das Konzept einer *Pseudovarietät von* $\mathbf{T}$*-Algebren* sowie einer *proendlichen Theorie* über $\widehat{\mathbf{T}}$ eingeführt. Pseudovarietäten von $\mathbf{T}$-Algebren verallgemeinern Pseudovarietäten von Monoiden, und proendliche Theorien bilden eine kategorielle Abstraktion von proendlichen Gleichungen. Wir erhalten:

**Satz** (Verallgemeinerte Reiterman-Korrespondenz [Ch16, Ur17])**.** *Proendliche Theorien über* $\widehat{\mathbf{T}}$ *stehen in bijektiver Korrespondenz zu Pseudovarietäten von* $\mathbf{T}$*-Algebren.*

Wählt man als $\mathbf{T}$ eine Monade auf der Kategorie der Mengen, die finitäre algebraische Strukturen (z.B. Monoide) repräsentiert, so erhält man den klassischen Reiterman-Satz als Spezialfall dieses Ergebnisses. Darüber hinaus lässt sich der Satz auf Situationen anwenden, für die bisher keine Reiterman-Korrespondenz bekannt war, etwa auf nicht finitäre algebraische Strukturen wie $\omega$-Halbgruppen.

### 3.3 Unäre Präsentationen und syntaktische Algebren

Der Schlüssel zu einem kategoriellen Verständnis von syntaktischen Monoiden (und anderen synktaktischen Algebren) liegt im Konzept einer *unären Präsentation* einer **T**-Algebra. Anschaulich beschreibt eine solche Präsentation, wie sich eine **T**-Algebra durch unäre Operationen beschreiben lässt. Beispielsweise kann ein Monoid $M$ durch die Translationen $x \mapsto yx$ und $x \mapsto xy$ mit $y \in M$ präsentiert werden. Der Zusammenhang zwischen unären Präsentationen und syntaktischen Algebren ist durch den folgenden Satz gegeben:

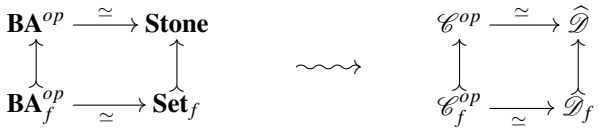
**Satz** ([Ur17])**.** *Wenn die freie* **T***-Algebra* $\mathbf{T}\mathbb{\Sigma}$ *eine unäre Präsentation hat, dann hat jede* **T***-erkennbare Sprache* $L\colon \mathbf{T}\mathbb{\Sigma} \to O$ *eine syntaktische* **T***-Algebra.*

Dieses Ergebnis vereinheitlicht nicht nur die Konstruktion zahlreicher bekannter Typen von syntaktischen Strukturen, sondern es erklärt auch die Rolle, die diese in der algebraischen Sprachtheorie einnehmen: aus kategorieller Sicht bedeutet das Arbeiten mit syntaktischen Algebren in Wirklichkeit das Arbeiten mit unären Präsentationen. Es ergeben sich viele Vorteile, das Konzept einer unären Präsentation explizit herauszustellen; unter anderem liefert es eine klarere Sicht auf viele Definitionen und Beweise in der Literatur.

### 3.4 Dualität

Die finale Zutat unseres kategoriellen Zugangs zur algebraischen Sprachtheorie liegt in der Interpretation algebraischer Spracherkennung durch *Dualisierung*. Wie bereits erwähnt, lassen sich die regulären Sprachen topologisch als abgeschlossene offene Mengen im Stone-Raum $\widehat{\Sigma^*}$ aller proendlichen Wörter beschreiben. Pippenger [Pi97] konnte nachweisen, dass sich dieses Ergebnis als Instanz der klassischen *Stone-Dualität* zwischen der Kategorie **BA** der booleschen Algebren und der Kategorie **Stone** der Stone-Räume auffassen lässt: Die boolesche Algebra aller regulären Sprachen über $\Sigma$ (mit den mengentheoretischen booleschen Operationen) ist dual zum Stone-Raum $\widehat{\Sigma^*}$. Die Korrektheit von Pippengers Ergebnis fußt implizit auf der Tatsache, dass sich die Stone-Dualität zu einer dualen Äquivalenz zwischen der Kategorie $\mathbf{BA}_f$ der endlichen booleschen Algebren und der Kategorie $\mathbf{Set}_f$ der endlichen Mengen (= endlichen Stone-Räume) einschränken lässt.

Eine der zentralen Erkenntnisse unserer Arbeit besteht darin, dass die topologischen Details der Stone-Dualität unerheblich sind: Man kann mit einem beliebigen (!) Paar $\mathscr{C}$ und $\mathscr{D}$ von algebraischen Kategorien arbeiten, die auf der Ebene von endlichen Algebren dual zueinander sind. Zwei solche Kategorien heißen *prädual*. Das bedeutet, dass man die auf der linken Seite des folgendes Diagramms abgebildete Stone-Dualität durch eine abstrakte duale Situation wie auf der rechten Seite ersetzen kann.



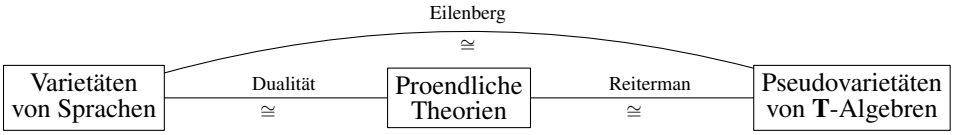
Neben der Stone-Dualität ($\mathscr{C} =$ boolesche Algebren und $\mathscr{D} =$ Mengen) gibt es viele weitere Beispiele von prädualen Kategorien, etwa die Birkhoff-Dualität ($\mathscr{C} =$ distributive Verbände und $\mathscr{D} =$ halbgeordnete Mengen) oder die Selbstdualität von endlichdimensionalen Vektorräumen ($\mathscr{C} = \mathscr{D} =$ Vektorräume). Für jedes präduale Paar $\mathscr{C}/\mathscr{D}$ kann man die Menge $\mathsf{Rec}(\mathbb{\Sigma})$ aller $\mathbf{T}$-erkennbaren Sprachen über dem Eingabeobjekt $\mathbb{\Sigma}$ in natürlicher Weise mit der Struktur einer $\mathscr{C}$-Algebra versehen. Das ermöglicht eine Verallgemeinerung von Pippengers dualer Charakterisierung der booleschen Algebra aller regulären Sprachen auf die Ebene einer beliebigen Monade $\mathbf{T}$:

**Satz** (Verallgemeinerte Pippenger-Korrespondenz [Ur17])**.** *Die Objekte* $\mathsf{Rec}(\mathbb{\Sigma}) \in \mathscr{C}$ *und* $\widehat{\mathbf{T}}\mathbb{\Sigma} \in \widehat{\mathscr{D}}$ *sind dual zueinander.*

Dieses Ergebnis ermöglicht es, das sehr allgemeine Konzept einer *Varietät von* $\mathbf{T}$*-erkennbaren Sprachen* einzuführen. Diese Varietäten werden in der Kategorie $\mathscr{C}$ gebildet, und ihre Definition beinhaltet den Begriff einer *Ableitung* von Sprachen, der auf einer unären Präsentation für $\mathbf{T}$-Algebren basiert. Im klassischen Fall (d.h. für die Monoid-Monade $\mathbf{T}\Sigma = \Sigma^*$ auf $\mathscr{D} = \mathbf{Set}$) reflektiert die Abgeschlossenheit von Varietäten von regulären Sprachen unter den booleschen Operationen die Tatsache, dass Varietäten in der Kategorie $\mathscr{C}$ der booleschen Algebren gebildet werden, und die Definition der Ableitungen $y^{-1}L$ und $Ly^{-1}$ ist der unären Präsentation von Monoiden durch die Operationen $x \mapsto yx$ and $x \mapsto xy$ geschuldet. Möchte man modifizierte Konzepte von Varietäten betrachten, die nur unter einer Teilmenge der booleschen Operationen abgeschlossen sind, so muss lediglich die unterliegende Dualität angepasst werden. Beispielsweise modelliert man Pins positive Varietäten von regulären Sprachen durch das präduale Paar $\mathscr{C} =$ distributive Verbände / $\mathscr{D} =$ halbgeordnete Mengen und die Monade $\mathbf{T}$ auf $\mathscr{D}$, die geordnete Monoide repräsentiert. Auf diese Weise erhält man zahlreiche Varietätenbegriffe aus der Literatur als Instanzen unserer allgemeinen Definition. Darüber hinaus ergibt sich eine starke Verallgemeinerung von Eilenbergs Varietätensatz, die eines der Hauptergebnisse unserer Dissertation darstellt:

**Satz** (Verallgemeinerte Eilenberg-Korrespondenz [Ur17])**.** *Varietäten von* $\mathbf{T}$*-erkennbaren Sprachen stehen in bijektiver Korrespondenz zu Pseudovarietäten von* $\mathbf{T}$*-Algebren.*

Dank der abstrakten kategoriellen Modellierung der verwendeten Begriffe ist der Beweis dieses Satzes konzeptionell erstaunlich einfach. Er basiert auf zwei Beobachtungen: (i) Varietäten von $\mathbf{T}$-erkennbaren Sprachen können, unter Verwendung der verallgemeinerten Pippenger-Korrespondenz, als das *duale* Konzept zu proendlichen Theorien interpretiert werden, und (ii) nach dem verallgemeinerten Reiterman-Satz stehen proendliche Theorien in bijektiver Korrespondenz zu Pseudovarietäten von $\mathbf{T}$-Algebren. Die zentralen Ergebnisse unserer Arbeit werden somit durch das folgende Diagramm in Beziehung gesetzt:




Die kategorielle Eilenberg-Reiterman-Theorie liefert ein abstraktes und parametrisches Rahmenwerk für die algebraische Theorie der formalen Sprachen. Die erzielten Ergebnisse demonstrieren, dass die Schritte (2) bis (5) des "klassischen" Fünf-Punkte-Plans (siehe

Abschnitt 2) vollständig generisch sind: Nach einer anwendungsspezifischen Wahl der Monade **T** und ihrer unären Präsentation erhält man die Konstruktion von syntaktischen Algebren, die Konzepte einer Varietät von Sprachen und einer Pseudovarietät von Algebren sowie den Varietätensatz direkt aus unseren allgemeinen Ergebnissen.

## 4 Anwendungen

Die entwickelten kategoriellen Methoden sind auf verschiedene Typen von Sprachen und auf vielfältige Fragestellungen anwendbar. Als Instanzen des verallgemeinerten Eilenberg-Satzes erhalten wir rund ein Dutzend wichtige Ergebnisse aus der Literatur, darunter fünf Eilenberg-Sätze für reguläre Sprachen, zwei Eilenberg-Sätze für $\omega$-reguläre Sprachen, zwei Eilenberg-Sätze für Sprachen über linearen Ordnungen, einen Eilenberg-Satz für Baumsprachen, und einen Eilenberg-Satz für Kostenfunktionen. Darüber hinaus konnten mehrere neue Eilenberg-Korrespondenzen abgeleitet werden, zum Beispiel eine Erweiterung des lokalen Varietätensatzes von Gehrke, Grigorieff und Pin [GGP08] von endlichen Wörtern auf unendliche Wörter, Bäume und Kostenfunktionen. Als weitere direkte Anwendung unserer Techniken haben sich mehrere neue Beiträge zur Theorie der klassischen regulären Sprachen ergeben, unter anderem eine Verallgemeinerung der Äquivalenz zwischen endlichen Automaten und Monoiden durch Interpretation dieser Konzepte über monoidalen Kategorien [AMU15], eine neue, rein automatentheoretische Interpretation des klassischen Eilenberg-Satzes via Algebra-Koalgebra-Dualität für endliche Automaten [Ad14, Ad15], sowie ein neuer algebraischer Zugang zur Konkatenation regulärer Sprachen durch Betrachtung monoidaler Adjunktionen [CU16]. Die kategorielle Perspektive hat somit auch zu neuen Einsichten über bereits umfangreich erforschte Strukturen geführt.

## 5 Fazit

Die in der Dissertation [Ur18] entwickelten Techniken tragen substanziell zu einem kategorientheoretischen Verständnis der algebraischen Theorie der formalen Sprachen bei. Durch die Allgemeinheit des eingeführten Rahmenwerks konnte demonstriert werden, dass die Schlüsselelemente dieser Theorie nicht inhärent algebraischer oder topologischer Natur sind, sondern durch abstrakte kategorielle Prinzipien beschrieben werden können. Diese Einsicht führt zu einer konzeptionellen Vereinfachung, Verallgemeinerung und Vereinheitlichung zahlreicher Ideen und Ergebnisse, die zuvor für verschiedene Sprachtypen separat betrachtet wurden. Die klare Trennung zwischen den generischen Aspekten der Theorie und ihrem anwendungsspezifischen Teil ermöglicht sowohl eine frische Sicht auf bekannte Strukturen als auch eine stark vereinfachte Herleitung neuer Ergebnisse.

## Literatur

[Ad14] Adámek, J.; Milius, S.; Myers, R.; Urbat, H.: Generalized Eilenberg theorem I: Local Varieties of languages. In: Proc. 17th International Conference on Foundations of Software Science and Computation Structures. Jgg. 8412 in LNCS. Springer, S. 366–380, 2014.

[Ad15] Adámek, J.; Myers, R.; Milius, S.; Urbat, H.: Varieties of languages in a category. In: Proc. 30th Annual ACM/IEEE Symposium on Logic in Computer Science. IEEE, S. 414–425, 2015.

[AMU15] Adámek, J.; Milius, S.; Urbat, H.: Syntactic Monoids in a Category. In: Proc. 6th Conference on Algebra and Coalgebra in Computer Science. LIPIcs. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2015. Best Paper Award.

[Bo15] Bojańczyk, M.: Recognisable languages over monads. In: Proc. 19th International Conference on Developments in Language Theory, Jgg. 9168 in LNCS, S. 1–13. Springer, 2015.

[Ch16] Chen, L.-T.; Adámek, J.; Milius, S.; Urbat, H.: Profinite Monads, Profinite Equations and Reiterman's Theorem. In: Proc. 19th International Conference on Foundations of Software Science and Computation Structures. Jgg. 9634 in LNCS. Springer, S. 531–547, 2016.

[CU16] Chen, L.-T.; Urbat, H.: Schützenberger Products in a Category. In: Proc. 20th International Conference on Developments in Language Theory. Jgg. 9840 in LNCS. Springer, S. 89–101, 2016.

[Ei76] Eilenberg, S.: Automata, Languages, and Machines Vol. B. Academic Press, 1976.

[GGP08] Gehrke, M.; Grigorieff, S.; Pin, J.-É.: Duality and equational theory of regular languages. In: Proc. 35th International Colloquium on Automata, Languages and Programming, Part II. Jgg. 5126 in LNCS. Springer, S. 246–257, 2008.

[Pi95] Pin, J.-É.: A variety theorem without complementation. Russ. Math., 39:80–90, 1995.

[Pi97] Pippenger, N.: Regular languages and Stone duality. Th. Comp. Sys., 30(2):121–134, 1997.

[Re82] Reiterman, J.: The Birkhoff theorem for finite algebras. Algebra Universalis, 14(1):1–10, 1982.

[Sc65] Schützenberger, M. P.: On finite monoids having only trivial subgroups. Inform. and Control, 8:190–194, 1965.

[Ur17] Urbat, H.; Adámek, J.; Chen, L.-T.; Milius, S.: Eilenberg Theorems for Free. In: Proc. 42nd International Symposium on Mathematical Foundations of Computer Science. LIPIcs. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, S. 43:1–43:14, 2017. EATCS Best Paper Award.

[Ur18] Urbat, H.: A Categorical Approach to Algebraic Language Theory. Dissertation, Technische Universität Braunschweig, 2018.

**Henning Urbat** ist seit 2017 Postdoc am Lehrstuhl für Theoretische Informatik der Friedrich-Alexander-Universität Erlangen-Nürnberg. Zuvor studierte er Mathematik und Informatik an der Technischen Universität Braunschweig und promovierte am Institut für Theoretische Informatik von Prof. Dr. Jiří Adámek. Sowohl seine Diplomarbeit als auch seine Dissertation wurden von der Carl-Friedrich-Gauß-Fakultät der TU Braunschweig ausgezeichnet. Sein Forschungsschwerpunkt liegt in der Entwicklung kategorieller und (ko-)algebraischer Methoden in der Informatik, insbesondere in der Automatentheorie und ihren Anwendungen.

# Missbrauchserkennung in kooperativen intelligenten Verkehrssystemen[1]

Rens W. van der Heijden[2]

**Abstract:** Für die Verbesserung der Verkehrssicherheit und Verkehrseffizienz wird an kooperativen intelligenten Verkehrssystemen (Cooperative Intelligent Transport Systems, C-ITS) gearbeitet. Diese infastrukturlosen Kommunikationsnetzwerke erlauben den Informationsaustausch zwischen Fahrzeugen, welches allerdings auch ein großes Spektrum neuer Angriffe ermöglicht. Besonders wichtig ist hier die Überprüfung der Integrität und Korrektheit der Datenflüsse, da fehlerhafte Informationen zu Unfällen führen können. Das Maat Fusionsframework, welches im Rahmen dieser Arbeit entstanden ist, nutzt Subjective Logic, um eine modulare und flexible Überprüfung dieser Daten zu ermöglichen. Zur Umsetzung des Maat genannten Frameworks wurden Beiträge zur Fusion mittels Subjective Logic und neue Missbraucherkennungsalgorithmen entwickelt. Außerdem wurden verschiedene Metriken erarbeitet, welche die Anforderungen der C-ITS besser abbilden als bisherige Ansätze, und es wurde der erste öffentlich verfügbare Datensatz zur Auswertung von Missbraucherkennungssystemen erstellt.

## 1 Einführung

Autounfälle sind eine der Haupttodesursachen in der westlichen Welt, wofür Automobilhersteller und Forschende ein breites Spektrum an Lösungen analysiert haben. Die Kommunikation zwischen Fahrzeuge gilt hierbei als besonders vielversprechend, da sie hochmoderne Fahrerassistenzsysteme ermöglicht, die bisher nur mithilfe von Sensoren mit begrenzter Reichweite arbeiten. Um dem Fahrzeug ein vollständigeres Bild seiner Umgebung zu vermitteln, wurden verschiedene Standards vorgeschlagen, die den Informationsaustausch zwischen den Fahrzeugen ermöglichen. Jüngste Entwicklungen in diesem Bereich, die mehr Komponenten in diese Kommunikationsarchitektur integrieren, führen zu kooperativen intelligenten Verkehrssystemen (C-ITS), welche Entscheidungen auf Grundlage von Sensorinformationen treffen, die aus teilweise nicht-vertrauenswürdigen Quellen empfangen werden.

Für einen erfolgreichen Einsatz von C-ITS ist die Absicherung gegen ungültiges Verhalten sowie gegen bösartige Angriffe essentiell. Ohne einen solchen Schutz kann die Gültigkeit der von anderen Fahrzeugen erhaltenen Informationen nicht garantiert werden, sodass die Zuverlässigkeit aller C-ITS-Anwendungen beeinträchtigt wird. Die Forschung hat erhebliche Ressourcen in die Entwicklung grundlegender Sicherheitsfunktionen wie Pseudonymisierung und Absenderauthentifizierung investiert, was auch zu Standardisierung

[1] Englischer Titel der Dissertation: “Misbehavior Detection in Cooperative Intelligent Transport Systems” [He18a]

[2] Institut für verteilte Systeme, Universität Ulm, rens.vanderheijden@uni-ulm.de

geführt hat. Ein Bereich, der hierbei bisher wenig beachte wurde, ist das Fehlverhalten authentisierter Entitäten im Netzwerk. Ein bösartiges Fahrzeug kann gezielt Fehlverhalten aufzeigen, welches aufgrund gültigen Schlüsselmaterials von anderen Entitäten im Netzwerk als authentisch akzeptiert wird. Zum Beispiel könnte so versucht werden, gefälschte Nachrichten zu übermitteln, die gezielt eine Notfallreaktion auslösen, was im schlimmsten Fall einen Unfall zwischen anderen Fahrzeugen verursachen könnte. Solche Angriffe können mit Standard-Sicherheitsfunktionen wie kryptographischen Signaturen nicht verhindert werden, weswegen Erkennungsmechanismen für solche Angriffe unersetzlich sind.

Die Beiträge dieser Arbeit umfassen (a) einen Überblick über bestehende Mechanismen für Misbehavior Detection, (b) Maat, einen Vorschlag für ein generisches Fusionsframework zur Misbehavior Detection in C-ITS, (c) Multi-Source-Fusionsoperationen für Subjective Logic, die das mathematische Fundament unseres Frameworks bilden, (d) mehrere neue Erkennungsmechanismen, (e) eine detaillierte Überprüfung der Bewertungsmethoden und Vorschläge für neue Metriken, (f) einen neuen, öffentlichen Datensatz, der als Grundlage für den Vergleich von Erkennungsmechanismen dienen kann, (g) eine detaillierte Bewertung der vorgeschlagenen Mechanismen und Fusionsverfahren und (h) einen Ausblick, wie diese Ergebnisse auf andere cyberphysikalische Systeme angewendet werden können.

## 2 Stand der Forschung

In [He18b] beschreiben und klassifizieren wir verschiedene Erkennungsansätze die bereits in der Literatur behandelt sind. Die übergeordnete Klassifikation ist eine zweidimensionale Taxonomie, in der feinkörnigere Erkennungstechniken eingeordnet werden. Außerdem wird klassifiziert, wo die Erkennung stattfindet, sowie eine Abschätzung, inwiefern dieser Mechanismus mit Datenschutzanforderungen vereinbar ist. Ein besonders auffälliges Ergebnis war, dass viele Erkennungsansätze unterschiedliche, teilweise disjunkte Mengen von Angriffen erkennen können. Außerdem gibt es viele Mechanismen, die nur in bestimmten C-ITS-Szenarien anwendbar sind, z.B. nur im städtischen Verkehr oder nur auf Autobahnen. Diese zwei Erkenntnisse zeigen, dass die Kombination von Erkennungsmechanismen für die Entscheidungsfindung ein wesentlicher Bestandteil sein sollte, um die Sicherheit von C-ITS entsprechend zu verbessern. Diese Arbeit nähert sich dem Thema durch die Entwicklung von Maat, welches die Gültigkeit der empfangenen Daten sicherstellt.

Die bisherige Forschung hat sich mit dem Thema der Misbehavior Detection Frameworks bereits auseinandergesetzt. Wir beschreiben hier kurz vier wesentliche Arbeiten, die prototypisch für die bisherigen Ansätze sind, und unterschiedliche Stärken und Schwächen mit sich bringen: VEBAS von Schmidt et al. [Sc08], den Ansatz basierend auf Kalman Filter von Stübing [St13], das Framework von Raya [Ra09] und die Ideen von Bißmeyer [Bi14].

VEBAS [Sc08] kombiniert eine große Anzahl von Plausibilitäts- und Konsistenzmechanismen, die jeweils verschiedene Arten von Angreifern erkennen können. Die verschiedenen Mechanismen werden nach einer gewissen Laufzeit statisch zusammengeführt, wonach

mittels einem Exponential Weighted Average (EWA) ein Vetrauenswert abgeleitet wird. Die große Stärke dieser Arbeit ist die Einfachheit und die daraus folgende Interpretierbarkeit des Outputs. Schwächen sind u.a. die Inflexibiliät der Fusion, insbesondere wenn es darum geht, Mechanismen für neu erkannte Angriffsmuster hinzuzufügen, und die Verzögerung bei der Erkennung. Raya [Ra09] verfolgt ähnliche Ansätze, setzt aber einen noch stärkeren Fokus auf die Vertrauensmechanismen. Die späteren Arbeiten von Stübing [St13] und Bißmeyer [Bi14] beschreiben konkrete Ansätze für die Erkennung einer bestimmten Art von Angriffen, nämlich die Fälschung der Positionsdaten, die periodisch von den Fahrzeugen verbreitet werden. Im Ansatz von Stübing wird ein dediziertes Framework gebaut, um trotz des Einsatzes von Pseudonymisierungsalgorithmen aus den Nachrichten benachbarter Fahrzeugen mittels eines Kalman-Filters abschätzen zu können, wo sich welches Fahrzeug befindet. Dieses Framework hat die große Stärke, dass die Ergebnisse direkt verfügbar sind und ggf. angewendet werden können; das Framework ist jedoch nicht für andere Arten von Daten geeignet. Der Erkennungsansatz von Bißmeyer hat ähnliche Ziele und basiert darauf, dass der Angreifer bei Fälschung der Position eine Überschneidung mit der Position eines benachbartes Fahrzeug riskiert. Was in diesen Arbeiten fehlt, ist die Flexibilität, wie dies bei Arbeiten wie VEBAS möglich ist, neue Erkennungsmechanismen hinzuzufügen.

## 3 Maat

Als Teil der Doktorarbeit wurde ein neues Framework entwickelt, Maat, dessen grundsätzlicher Ansatz es ist, die Ergebnisse verschiedener Erkennungsmechanismen mittels Subjective Logic zu fusionieren. Subjective Logic ist ein mathematisches Framework, welches den Ausdruck von Unsicherheit über Daten mittels sogenannten Opinions ermöglicht. Um dies zu ermöglichen, wurden Beiträge zur Logik selbst geliefert, sowie die Entwicklung neuer Erkennungsmechanismen, die Reproduktion und Verbesserung bestehender Erkennungsmechanismen, und der Aufbau eines Weltmodells, in dem die Erkennungsergebnisse verwaltet, ausgewertet und ggf. verbreitet werden können. Maat verwendet diese Logik, um ein flexibles Datenmanagement- und Fusionssystem aufzubauen, das die Vertrauenswürdigkeit der Daten bei jedem Zugriff durch Anwendungen bestimmt. Zur Unterstützung dieses Datenmanagements verwendet Maat einen gerichteten Graphen zur Speicherung der Daten und der zugehörigen Erkennungsergebnisse. Durch die getrennte Erfassung der Daten und der dazugehörigen Detektionsergebnisse kann eine Vielzahl von potenziellen neuen Detektoren untersucht werden. Darüber hinaus ermöglicht es den Austausch von Erkennungsergebnissen, zum Beispiel für die Revocation.

**Erkennungsmechanismen** Im Rahmen der Dissertation wurden verschiedene Erkennungsmechanismen entwickelt und erweitert [HLK18]. Außerdem haben wir versucht Ergebnisse bestehender Arbeiten zu reproduzieren, um diese an die neuen Anforderungen der C-ITS anzupassen. Im Folgenden wird beispielhaft der Acceptance Range Threshold (ART) [Sc08] eingeführt. Dieses Erkennungsverfahren prüft anhand des Abstands und

einer Abschätzung der Kommunikationsreichweite, ob die Positionsinformationen gefälscht wurden; ein Beispiel ist in Abbildung 1 zu sehen.

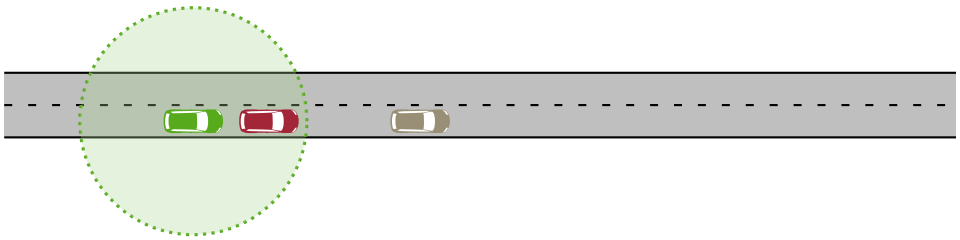


Abb. 1: Das grüne Fahrzeug nimmt für den ART die zirkelförmige Kommunikationsreichweite an, und erkennt hierdurch den roten Angreifer, welcher vorgibt, an der grauen Position zu sein.

Eine Schwäche dieses Mechanismus ist, dass die Empfangstreichweite als fest angenommen wird, was jedoch bei sich schnell fortbewegenden Fahrzeuge nicht der Fall ist. Aufgrund dessen wird vorgeschlagen, mittels Subjective Logic eine Unsicherheit auszudrücken, die in der Fusion mit anderen Mechanismen eine bessere Erkennungsrate ermöglicht [HLK18].

**Subjective Logic** Subjective Logic bietet eine Vielzahl von Fusionsoperatoren, um Opinions zu fusionieren [Jø16]. Eine Opinion ist definiert über ein Zufallsvariabele $X$ und besteht aus zwei Funktionen und einem Wert; der Belief Function $\mathbf{b}$, der Base Rate $\mathbf{a}$, und dem Unsicherheitswert $u$. Die Belief Function modelliert explizites Wissen, während die Base Rate A-priori-Wissen modelliert; durch die additive Eigenschaft $1 = u + \sum_{x \in X} \mathbf{b}(x)$ wird die Second-order Uncertainty mittels $u$ beschrieben. Um die Opinions verschiedener Entitäten fusionieren zu können, ist eine Verallgemeinerung der existierenden Fusionoperatoren nötig, da manche Fusionsoperatoren nicht kommutativ sind. Diese wurde im Rahmen der Dissertation [He18a] geliefert; als Beispiel zeigt Definition 1 die WBF, bei der die Evidenz weder unendlich noch null ist ($\forall A \in \mathbb{A} : u_X^A \neq 0$) $\wedge$ ($\exists A \in \mathbb{A} : u_X^A \neq 1$). Für die Base Rate wird aus Platzgründe auf [He18a] verwiesen.

**Definition 1 (Multi-source WBF)** *Gegeben die Opinion $\omega_X^A = (\mathbf{b}_X^A, u_X^A, \mathbf{a}_X^A)$ von Aktor $A \in \mathbb{A}$ über die Zufallsvariable $X$ ist die gemeinsame Opinion $\omega_X^{\widehat{\diamond}\mathbb{A}} = (\mathbf{b}_X^{\widehat{\diamond}\mathbb{A}}, u_X^{\widehat{\diamond}\mathbb{A}}, \mathbf{a}_X^{\widehat{\diamond}\mathbb{A}})$:*

$$\mathbf{b}_X^{\widehat{\diamond}\mathbb{A}}(x) = \frac{\sum\limits_{A \in \mathbb{A}} \mathbf{b}_X^A(x)\left(1 - u_X^A\right) \prod\limits_{A' \in \mathbb{A}, A' \neq A} u_X^{A'}}{\left(\sum\limits_{A \in \mathbb{A}} \prod\limits_{A' \neq A} u_X^A\right) - |\mathbb{A}| \cdot \prod\limits_{A \in \mathbb{A}} u_X^A} \qquad u_X^{\widehat{\diamond}\mathbb{A}} = \frac{\left(|\mathbb{A}| - \sum\limits_{A \in \mathbb{A}} u_X^A\right) \cdot \prod\limits_{A \in \mathbb{A}} u_X^A}{\left(\sum\limits_{A \in \mathbb{A}} \prod\limits_{A' \neq A} u_X^A\right) - |\mathbb{A}| \cdot \prod\limits_{A \in \mathbb{A}} u_X^A}$$

Um bereits existierende Fusionsansätze modellieren zu können, die nicht in Subjective Logic vorhanden sind, wurden außerdem die Operatoren Minimum (MIN) und Majority

(MAJ) definiert. Um Vertrauensbeziehungen zwischen Fahrzeugen mit einbeziehen zu können, bietet Subjective Logic bereits transitive Vertrauensbeziehungen. Letztlich wurde im Rahmen der Dissertation eine Variante des EWA vorgeschlagen, welche über Opinions definiert ist, da dieser Ansatz häufig in der Literatur zur Anwendung kommt.

**Weltmodell** Maat nutzt die Subjective Logic um Beziehungen zwischen Informationen, Entitäten und Erkennungsmechanismen in einem gerichteten Graph zu modellieren, bei dem die Kanten mit Opinions beschriftet werden. In Abbildung 2 befindet sich ein Beispiel für einen solchen Graph. Die Semantik der Kanten von Maat zum Erkennungsmechanismus ist eine konfigurierbare Gewichtung (welche wir mit *Fusionability* bezeichnen), während die Kanten vom Erkennungsmechanismus zu Daten oder Entitäten den Erkennungsergebnissen entsprechen. Die Kanten von Entitäten zu anderen Entitäten oder zu Daten sind Opinions, die in einer Nachricht mitübertragen werden können.

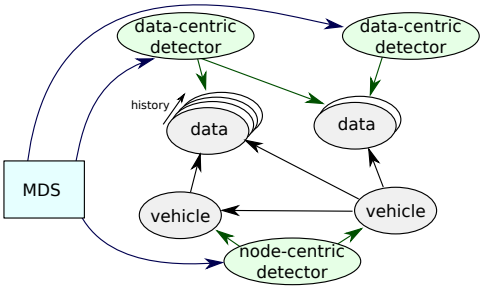



Abb. 2: Ein Weltmodell in Maat, in dem zwei Fahrzeuge Daten liefern, und innerhalb von Maat drei Erkennungsmechanismen genutzt werden. Alle Kanten werden mit Opinions beschriftet.

Wenn eine neue Nachricht im Fahrzeug eingeht, wird diese an Maat übermittelt und an alle angeschaltete Erkennungsmechanismen weitergegeben, die über diese (und ggf. ältere) Nachrichten jeweils Opinions erzeugen; diese werden anschließend zusammen mit den Informationen in der neuen Nachricht gleichtzeitig im Weltmodell gespeichert. Eine C-ITS Anwendung kann jederzeit Anfragen an das Weltmodell stellen; dieses wird zur Zeit der Anfrage den letzten konsistenten Zustand nutzen und eine Fusion durchführen. Die Laufzeit, bis ein neuer konsistenter Zustand vorliegt, hängt maßgeblich von der Laufzeit der einzelnen Erkennungsmechanismen ab, die modular wählbar sind.

Um eine Anfrage beantworten zu können, wird in Maat der Fusionsprozess in Abbildung 3 angestoßen und durchgeführt. Um das Vertrauen in bestimmte Daten zu bestimmen, wird zuerst eine Pfadsuche durchgeführt. Danach wird in jedem Pfad der Transitive-Trust-Operator genutzt, um das Pfad-Vertrauen zu berechnen; insbesondere wird hierdurch das Vertrauen in den Erkennungsmechanismus mit eingerechnet. Dies ist nützlich um die Erkennung nachträglich optimieren zu können (z.B. wenn Maat bereits in vielen Fahrzeugen implementiert ist). Danach wird der konfigurierte Fusionsoperator genutzt, um das Gesamtvertrauen zu berechnen. Das Ergebnis ist eine Opinion, die wahlweise zu einer

Entscheidung projiziert werden kann (wie in der Abbildung dargestellt) oder von einer Anwendung direkt verarbeitet werden kann. Letzteres ist insbesondere nützlich, wenn die Anwendung die Unsicherheit der Opinion direkt verarbeiten kann.

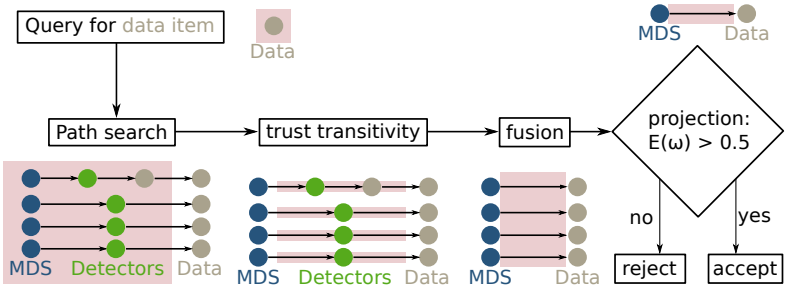



Abb. 3: Das Fusionsprozess von Maat

## 4 Auswertung

Da die meisten Arbeiten weder Quellcode des Detektors, noch Quellcode des Angreifers, noch das Auswertungsszenario veröffentlichen, ist die Reproduktion bestehender Ergebnisse erschwert; für einige Arbeiten konnten daher die Ergebnisse nur teilweise reproduziert werden. Um dieses Problem für die Zukunft zu vermeiden, wurde ein synthetischer Datensatz entwickelt, welcher auf Basis verbreiteter Open-Source Tools eine Grundlage für Angriffserkennung bildet. Dieser Datensatz, VeReMi, besteht neben dem Auswertungsszenario auch aus Angriffen; die Erkennungsmechanismen wurden als Teil von Maat veröffentlicht. Es ist allerdings wichtig zu betonen, dass dieser Datensatz immer als *Grundlage* dienen sollte; VeReMi bietet einen Rahmen, in dem neue Angriffe ergänzt werden können.

**Metriken** Im Rahmen dieser Arbeit wurden zwei Kategorien von Metriken analysiert: Anwendungsmetriken und Erkennungsqualitätsmetriken. Für die Anwendungsmetrik wurde die Anwendung Cooperative Adaptive Cruise Control (CACC) als Grundlage gewählt; da in der Literatur noch keine geeignete Metriken oder Angriffe vorhanden waren, wurde die Analyse solcher Angriffe im Detail durchgeführt. Im Rahmen von [HLK17] haben wir herausgestellt, wie die Anfälligkeit von CACC für Angriffe aussieht, sowohl in der Form von Jamming als auch in der Form von Datenfälschung. Für die Erkennungsqualität wurde Precision & Recall gewählt, auf Basis der korrekte Klassifikation der Nachrichten.

Der Datensatz, der im Rahmen der Arbeit entwickelt wurde, bietet ein realistisches Verkehrsszenario, auf dem CACC allerdings nicht direkt einsetzbar war. Dementspechend wurden hierfür die Erkennungsmetriken eingesetzt, die auch bei bestehenden Arbeiten zum Einsatz kommen. Es lässt sich allerdings noch ergänzen, dass viele bisherige Arbeiten die Metriken auf die Erkennung von Angreifern anwenden, statt auf die Erkennung von Angreifernachrichten. Der Unterschied zeigt sich darin, dass im ersten Fall nur am Ende

des Simulationsdurchlaufs analysiert wird, wie vielen Angreifern fälschlicherweise vertraut wurde. Viel interessanter ist jedoch, insbesondere für C-ITS Anwendungen, wie sich das Vertrauen pro Nachricht zusammenstellt; ob der Angriff am Ende erkannt wurde, reicht nicht aus, wenn der Angriff vorher schon einen Unfall ausgelöst hat. Aufgrund dessen wird in unseren Arbeiten über die Nachrichten aggregiert statt über die Fahrzeuge.

Neben der Standardmetrik (Precision & Recall) analysieren wir auch den Gini Index des False Positive (FPR) bzw False Negative rates (FNR) über die Fahrzeuge. Ziel hiervon ist festzustellen, wie gleichmäßig solche Fehler sind; je gleichmäßiger die Fehler, desto unwahrscheinlicher ist es, dass ein Austausch von Erkennungsergebnissen die Erkennung verbessern könnte. Außerdem dient dieser Ansatz zur Überprüfung der These, dass das Verhältnis zwischen Fahrzeug und Angreifer die Möglichkeit der korrekten Erkennung maßgeblich beeinflusst. Wenn ein Angreifer zum Beispiel die Position lediglich um wenige Millimeter fälscht, ist die Erkennung nicht möglich, weil dies wesentlich unterhalb der Messfehler von GPS liegt; um den maximalen Recall zu erreichen sollten jedoch auch diese Angriffe erkannt werden. Der Gini Index $G_{FPR}$ wird für eine Simulation mit $\mu$ gutartigen Fahrzeugen, die jeweils eine FPR von $x_i$ haben, wie folgt definiert: $G_{FPR} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n^2\mu}$

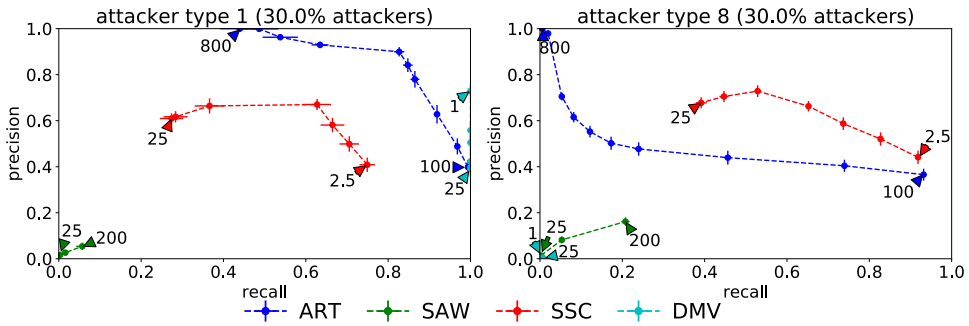



Abb. 4: Erkennungsergebnisse einzelner Erkennungsalgorithmen (aus [HLK18])

**Ergebnisse & Diskussion** Basierend auf der beschriebenen Auswertungsmethodik und dem VeReMi-Datensatz wird im Folgenden die Auswertung von Maat dagestellt. Hierzu wurde Maat mit verschiedenen Erkennungsmechanismen ausgestattet, welche bereits in unserer VeReMi-Arbeit vorgestellt wurden [HLK18]. Im folgenden werden beispielhaft Erkennungsergebnisse im Bezug auf Angreifer Typen 1 (immer die gleiche Position) und 8 (für jede Nachricht ein Random Offset der echten Position) in ein Szenario mit hoher Verkehrsdichte und Angreiferdichte (30%) abgebildet. In Abbildung 4 sind die Mechanismen mit unterschiedlichen Schwellwerten abgebildet. Hieraus lässt sich ablesen, dass sich die Erkennungsmechanismen stark in ihrer Performanz unterscheiden, in Abhängigkeit von der Art der Angreifer. Zum Beispiel erkennt der DMV-Detektor Angreifer 1 immer (Recall = 1), jedoch Angreifer 8 nie (Recall = 0). Hieraus lässt sich ableiten, dass die Fusion verschiedener Mechanismen nötig ist, denn ansonsten könnte der Angreifer den richtigen Angriff wählen,

um die Erkennung zu umgehen. In Abbildung 5 ist im gleichen Szenario die Fusion mit zwei unterschiedlichen Werten für den EWA und jeweils fünf unterschiedlichen Fusionsansätzen dargestellt. Die Zahlen in dem Plot stellen hier das Basisvertrauen der Mechanismen dar. Bei der Fusion wurde zusätzlich EWA eingesetzt, um die Qualität der fusionierten Ergebnisse vor der Fusion zu erhöhen. Da in unserem Szenario die Angreifer jede Nachricht manipulieren, sollte der EWA hier eine höhere Erkennungsrate ermöglichen. Allerdings zeigen die Ergebnisse, dass der EWA nur marginal beeinflusst, welche Angreifernachrichten erkannt werden (leichte Steigerung im Recall), aber eine bedeutende Reduktion der Precision zur Folge hat. Außerdem ist direkt erkennbar, dass MIN allen Nachrichten misstraut, während MAJ hauptsächlich vom Angreifer abhängt (gut gegen Angreifer 1, schlecht gegen Angreifer 8). Zwischen den einzelnen Subjective Logic Operatoren gibt es keinen Operator, der in jedem Fall gut funktioniert; in zukünftigen Studien soll herausgearbeitet werden, inwiefern Verbesserungen an den Erkennungsmechanismen hier Fortschritte ermöglichen. Eine Alternative hierzu wäre die programmatische Beschreibung des Fusionsprozesses, womit festgelegt werden kann, welcher Prozess wozu geeigneter ist; in diesen Fall würde für jede Anwendung ein jeweils passender Prozess definiert werden müssen.

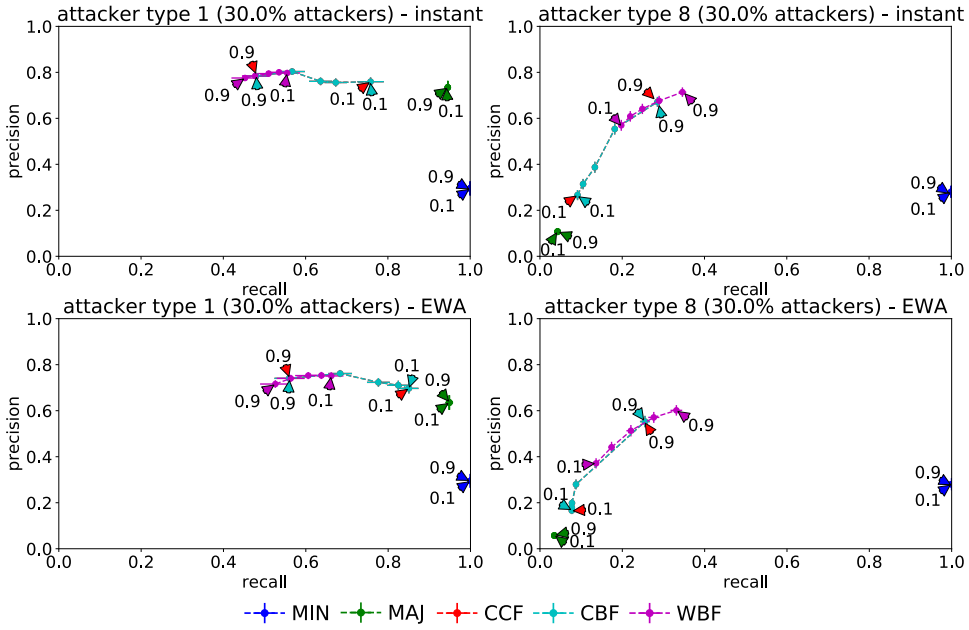



Abb. 5: Ergebnisse für verschiedene Fusionsoperatoren, mit einem EWA von 1 und 0.1.

## 5 Zusammenfassung

In dieser Arbeit wurde basierend auf einer ausführlichen Literaturanalyse für Misbehavior Detection ein neues Framework, Maat, beschrieben und implementiert. Dieses Framework

ermöglicht die Kombination verschiedener Erkennungsverfahren mittels Subjective Logic, sowie die Möglichkeit, das Framework mit neuen Erkennungsansätzen anzureichern. Um dies zu erreichen, wurden im Rahmen der Subjective Logic verschiedene Beiträge geleistet: insbesondere wurden die nicht-kommutativen Fusionsoperatoren verallgemeinert und verschiedene Operatoren entwickelt, die naive Fusionsansätze modellieren. Es wurden außerdem verschiedene neue Erkennungsmechanismen entwickelt und es wurden die Ergebnisse verschiedener bestehender Arbeiten reproduziert. Zur Auswertung wurden neue Mertiken entwickelt, die die neuen Anforderungen von C-ITS berücksichtigen, wobei insbesondere die Qualität der Erkennung ohne Vertrauensmechanismus betrachtet wird. Dies ist wichtig, da es bekannte Angriffstechniken gibt um Vertrauensmechanismen anzugreifen.

Über die konkreten Ergebnisse für Misbehavior Detection hinaus wurde in der Arbeit beleuchtet, wie sich diese Ansätze auf andere Cyber-Physical Systems übertragen lassen. Im Rahmen dieses Ausblicks wurde analysiert, inwiefern sich die Fusion direkt auf das Monitoring von Industriesteuersysteme übertragen lässt. Hierbei stellt sich klar heraus, dass die Erfolgswahrscheinlichkeit einer zielführenden Erkennung wesentlich von der Informationsgrundlage der Erkennungsmechanismen abhängt. Im Bereich der Industriesteuersysteme ist diese Informationshierarchie sehr heterogen, was dazu führt, dass die Erkennung verteilt umgesetzt werden sollte. Außerdem basieren viele reale Angriffe auf fehlender Authentisierung in bestehenden verdrahteten Netzwerken; hier ist es für einen Angreifer einfach, alle Informationsflüsse zum Erkennungssystem zu kontrollieren.

Zukünftige Arbeiten können sich mit der Anwendung von Maat für fahrzeuginterne Angriffserkennungszwecke, sowie die Verbesserung der Erkennungsrate mittels dynamischer Konfiguration beschäftigen. Für fahrzeuginterne Netzwerke stellt sich besonders die Frage, ob die Erkennung mit echtzeitfähigen Mechanismen ausreichend ist. Die Ergebnisse der Arbeit legen nahe, dass die dynamische Konfiguration und Gewichtung eine erhebliche Verbesserung der Erkennungsrate ermöglichen könnte, da die Fehlerrate der einzelnen Mechanismen einen starken Zusammenhang mit dem Szenario aufweist. Obwohl bisherige Experimente mit Maat auf Cyber-Physical Systems beschränkt sind, ist der Fusionsprozess im Prinzip lösungsneutral. Dies bedeutet, dass Maat theoretisch auch für die Fusion anderer Angriffserkennungsmechanismen eingesetzt werden könnte. Das Framework könnte auch für andere High-Level-Informationsfusionsanwendungen zum Einsatz kommen, jedoch müssten hierfür neue Auswertungsansätze entwickelt werden. Die größte Hürde hierbei ist jedoch die Opinions geeignet zu wählen, so dass eine Fusion auch einen nennenswerten Vorteil hat.

## Literatur

[Bi14] Bißmeyer, N.: Misbehavior Detection and Attacker Identification in Vehicular Ad-hoc Networks, Diss., Darmstadt Technical University, 2014.

[He18a] van der Heijden, R. W.: Misbehavior detection in cooperative intelligent transport systems, Diss., Universität Ulm, 2018.

[He18b] van der Heijden, R. W.; Dietzel, S.; Leinmüller, T.; Kargl, F.: Survey on Misbehavior Detection in Cooperative Intelligent Transportation Systems. IEEE Communication Surveys & Tutorials/, 2018.

[HLK17] van der Heijden, R. W.; Lukaseder, T.; Kargl, F.: Analyzing Attacks on Cooperative Adaptive Cruise Control (CACC). In: Vehicular Networking Conference. VNC, Best paper award, IEEE, S. 45–52, Nov. 2017.

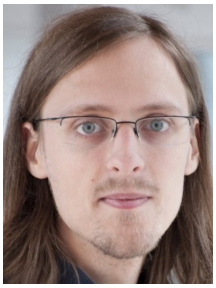
[HLK18] van der Heijden, R. W.; Lukaseder, T.; Kargl, F.: VeReMi: A Dataset for Comparable Evaluation of Misbehavior Detection in VANETs. In: 14th EAI SecureComm. Springer, Aug. 2018.

[Jø16] Jøsang, A.: Subjective Logic: A Formalism for Reasoning Under Uncertainty. Springer International Publishing Switzerland, 2016.

[Ra09] Raya, M.: Data-centric trust in ephemeral networks, Diss., Lausanne: EPFL, 2009.

[Sc08] Schmidt, R. K.; Leinmüller, T.; Schoch, E.; Held, A.; Schäfer, G.: Vehicle Behavior Analysis to Enhance Security in VANETs. In: Proceedings of the 4th Workshop on Vehicle to Vehicle Communications (V2VCOM 2008). IEEE, S. 1–8, 2008.

[St13] Stübing, H.: Multilayered Security and Privacy Protection in Car-to-X Networks, Diss., 2013.



**Rens W. van der Heijden** wurde am 19. September 1989 in den Niederlanden geboren. Er hat in August 2010 sein Bachelor in Informatik abgeschlossen und bekam im August 2012 einen *cum laude* Masterabschluss in *Computer Science* mit Spezialisierung in IT-Sicherheit. Am 9. November 2018 promovierte er *magna cum laude* an der Universität Ulm beim Institut für Verteilte Systeme bei Prof. Dr. Frank Kargl. Seit 2018 ist er wissenschaflticher Mitarbeiter am gleichen Institut und forscht dort schwerpunktmäßig an IT-Sicherheit in Fahrzeugen im Rahmen des BMBF SecForCARs Projekts. Seine weiteren Forschungsinteressen sind Subjective Logic, Intrusion Detection und Privacy. Er ist außerdem in der Lehre vertreten, wo er die Vorlesung Security & Privacy in Mobile Systems liest.