# Berechnung effizienter Datenzusammenfassungen[1]

Sebastian Mair[2]

**Abstract:**

Das Extrahieren sinnvoller Repräsentationen von Daten ist ein grundlegendes Problem im maschinellen Lernen und kann aus zwei unterschiedlichen Perspektiven betrachtet werden: (i) im Bezug auf die Anzahl der Datenpunkte und (ii) hinsichtlich der Repräsentation eines jeden einzelnen Datenpunktes in Bezug auf seine Dimensionen. Diese Arbeit beschäftigt sich mit diesen Perspektiven zur Datenrepräsentation und leistet dazu verschiedene Beiträge. Der erste Teil behandelt die Berechnung repräsentativer Teilmengen für die Archetypenanalyse und die Problemstellung der optimalen Versuchsplanung. Dafür motivieren und untersuchen wir die Brauchbarkeit der Punkte am Rand der Daten als neuartige repräsentative Teilmenge. Basierend auf dem Coreset-Prinzip leiten wir eine weitere repräsentative Teilmenge für die Archetypenanalyse her, welche zusätzliche theoretische Garantien bietet. Der zweite Teil der Arbeit handelt von effizienten Datenrepräsentationen für Dichteschätzungsprobleme. Wir analysieren raum-zeitliche Probleme, die z.B. in der Analyse von Mannschaftssportarten auftreten, und zeigen, wie sich statistische Bewegungsmodelle anhand von Trajektoriendaten lernen lassen. Darüber hinaus untersuchen wir Probleme hinsichtlich der Interpolation von Daten mittels generativer Modelle.

## 1 Einführung

Im maschinellen Lernen geht es hauptsächlich um das Auffinden von Strukturen und Mustern in Daten, sowie um das Finden von funktionalen Zusammenhängen zwischen Eingabe- und Ausgabedaten. Oft werden dabei Objekte aus der realen Welt als Vektoren repräsentiert, wobei die Dimensionen, auch Attribute genannt, verschiedene Merkmale der Objekte beschreiben. Eine Menge dieser Vektoren, auch Datenpunkte genannt, bildet dann den Datensatz. Aus diesem lernt ein Verfahren des maschinellen Lernens dann Muster bzw. Zusammenhänge. Dieser Datensatz wird üblicherweise in Form einer Matrix dargestellt. In dieser Datenmatrix liegen die Datenpunkte aufeinander gestapelt.

Das am weitesten verbreitetste Lernszenario ist das des überwachten Lernens. Hierbei besitzt jeder einzelne Datenpunkt zusätzlich einen Zielwert. Das Ziel ist das Lernen einer Funktion, welche neuen Datenpunkten automatisch passende Zielwerte zuordnet. Man spricht von einem Regressionsproblem, wenn der Zielwert eine reelle Zahl ist. Ist die Menge an möglichen Zielwerten jedoch diskret, betrachtet man ein Klassifikationsproblem. Im Allgemeinen vereinfacht das Vorhandenseins eines Zielwertes sowohl das Lernen, als auch das Evaluieren eines Modells. Jedoch ist ein solcher Zielwert nicht immer gegeben, so z.B.

[1] Englischer Titel der Dissertation: „Computing Efficient Data Summaries“

[2] Universität Uppsala, Schweden, sebastian.mair@it.uu.se

beim unüberwachten Lernen. Hierbei ist das Ziel das Finden von Strukturen und Mustern in den Daten. Typische Beispiele von unüberwachten Lernverfahren sind Anomalieerkennung, Dimensionalitätsreduktion und das Lernen von Wahrscheinlichkeitsverteilungen.

Ein entscheidender Faktor für die Leistungsfähigkeit eines Modells ist mitunter die Darstellung der Daten. Diese Repräsentation, im Sinne des Datensatzes bzw. der Datenmatrix, ist jedoch oft suboptimal. So kann der Datensatz beispielsweise viel zu groß für die vorhandene Infrastruktur sein, da er z.B. zu viele unnötige Redundanzen beinhaltet. Ebenfalls können vorhandene Attribute wenig oder gar keine Vorhersagekraft besitzen. In beiden Fällen ist eine Repräsentationsänderung vorteilhaft. Diese Änderung kann entweder als Teil der Datenvorverarbeitung erfolgen, oder ein Teil des Lernalgorithmus sein. Im Allgemeinen gibt es zwei Sichtweisen auf die Daten, bei denen eine Änderung der Repräsentation erfolgen kann.

## 2 Die zwei Sichtweisen des Repräsentationslernens

Beim Repräsentationslernen geht es darum, eine bessere Repräsentation der Daten zu finden. So kann z.B. die Datenmatrix für ein bestimmtes Lernszenario spezifisch angepasst werden. Diese Änderung der Repräsentation kann auf zwei unterschiedliche Arten betrachtet werden. Bei der einen Sichtweise geht es um die Anzahl der Datenpunkte und somit um die Stichprobengröße des Datensatzes während die andere von einer Änderung der Repräsentation in Bezug auf die Attribute handelt.

### Die Repräsentation des Datensatzes im Bezug auf die Anzahl der Datenpunkte

In vielen Problemstellungen sind Daten in großen Mengen vorhanden und beinhalten viele redundante Informationen. Dabei kann die Größe des Datensatzes im Sinne der Anzahl der Datenpunkte zu groß sein, um bestimmte Lernverfahren effizient anzuwenden. Abgesehen davon kann es auch sein, dass Speicher- und Berechnungsbeschränkungen uns dazu zwingen, eine Approximation des Lernverfahrens zu verwenden. Üblicherweise werden dann mit erheblichem Mehraufwand aufwändige Approximationen der Lernverfahren verwendet, um diese auf große Datenmengen anwenden zu können. Alternativ kann statt des Modells auch der zugrundeliegende Datensatz approximiert werden, womit wir uns in dieser Arbeit näher beschäftigen.

Wir nennen eine Teilmenge des Datensatzes *repräsentative Teilmenge* oder *Datenzusammenfassung*, wenn diese die Hauptcharakteristiken der Ursprungsdaten gut approximiert. Trivialerweise kann man unabhängig und gleich-verteilte Datenpunkte nach dem Zufallsprinzip aus dem Datensatz auswählen. Der Vorteil dieses Ansatzes liegt in seiner Einfachheit. Darüber hinaus liefert der Ansatz einen unverzerrten Schätzer der zu optimierenden Zielfunktion. Im folgenden werden wir jedoch zeigen, dass nur mit minimalem Mehraufwand repräsentativere Teilmengen generiert werden können, welche bei gleicher Stichprobengröße, erheblich bessere Approximationen liefern.

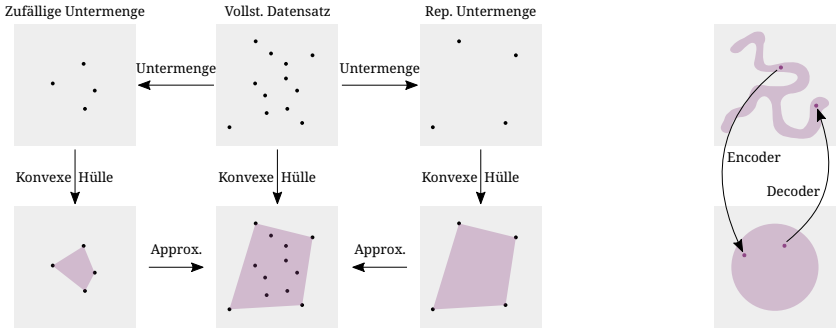


Abb. 1: Links: Eine repräsentative Untermenge für das Lernen einer konvexen Hülle. Rechts: Eine Datentransformation wie sie z.B. bei einem VAE vorkommt.

Abbildung 1 (links) zeigt Beispiele von repräsentativen Untermengen für die Berechnung einer konvexen Hülle. Eine Teilmenge bestehend aus unabhängig und gleich-verteilten Datenpunkten deckt einen kleineren Bereich ab als die gewünschte konvexe Hülle des vollständigen Datensatzes. Dies ist darauf zurückzuführen, dass die wichtigsten Punkte für dieses Problem am Rand der Daten liegen. Im Vergleich dazu enthält die repräsentative Teilmenge auf der rechten Seite der Abbildung mehr relevante Punkte und bietet somit eine bessere Zusammenfassung der vollständigen Daten. Infolgedessen liefert sie eine viel bessere Annäherung bei gleicher Größe der verwendeten Teilmenge.

### Die Repräsentation eines jeden einzelnen Datenpunktes

Üblicherweise wird Repräsentationslernen als eine Änderung der Repräsentation in Bezug auf die Dimensionen der Daten verstanden. So beinhalten z.B. nicht alle Dimensionen sinnvolle Informationen für die Lernaufgabe, oder die wesentlichen Informationen sind implizit in einem Unterraum eingebettet. Häufig kann eine Transformation der Datenrepräsentation das Lernverfahren an sich bzw. darauf aufbauende Verfahren vereinfachen.

Ein Beispiel für nicht-lineare Merkmalstransformationen sind Autoencoder, welche aus zwei neuronalen Netzen bestehen. Das erste Netz, der Encoder, bettet die Eingabedaten in einen sogenannten *latenten Raum* ein. Die Eingabedaten können dann durch die Repräsentationen im latenten Raum durch das zweite Netz, Decoder genannt, rekonstruiert werden. Die latente Repräsentation ist oft von geringerer Dimensionalität und zwingt den Autoencoder die wesentlichen Charakteristiken und Eigenschaften der Daten zu extrahieren.

Ein Autoencoder kann auch zur Generierung neuer Daten, welche ähnlich zu den Trainigsdaten sind, verwendet werden. Hierfür ist es nötig, dem latenten Raum eine bestimmte Struktur aufzuzwingen. Dies realisiert beispielsweise ein Variational Autoencoder (VAE), welcher üblicherweise eine Normalverteilung im latenten Raum erzwingt. Wir erhalten somit ein generatives Modell, in welchem wir Realisierungen der Normalverteilung mittels

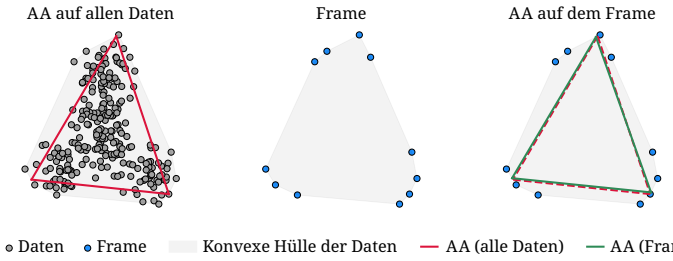


Abb. 2: Eine Archetypenanalyse mit drei Faktoren berechnet auf allen Daten (links), der Frame der Daten (mitte) und die Archetypenanalyse berechnet auf dem Frame (rechts).

des Decoders in neue, synthetische Datenpunkte transformieren können. Ein Beispiel ist in Abbildung 1 (rechts) dargestellt. Eine komplexe Datenverteilung wird in eine einfachere Verteilung umgewandelt. Dies ist nicht nur für die generative Modellierung, sondern auch für die Dichteschätzung vorteilhaft.

## 3 Beiträge dieser Arbeit

Die kummulative Dissertation [Ma22] beinhaltet sechs Einzelpublikationen und leistet verschiedene Beiträge im Bereich des unüberwachten Repräsentationslernens. Dabei werden insbesondere die zwei zuvor eingeführten Sichtweisen behandelt und pro Sichtweise eine Forschungsfrage abgeleitet. Der erste Teil der Arbeit beschäftigt sich mit der Frage, wie man effizient repräsentative Teilmengen für Lernaufgaben berechnen kann, welche es ermöglichen, dasselbe aus weniger Daten auf effizientere Weise zu lernen. Im zweiten Teil stellt sich die Frage, wie man Datenrepräsentationen erhält, welche die Anwendung spezifischer Operationen wie Dichteschätzung und Interpolation erleichtern. Im Folgenden wird pro Forschungsfrage eine Auswahl an Beiträgen der Dissertation vorgestellt.

### 3.1 Der Frame als repräsentative Untermenge

Wir untersuchen zunächst die Idee den Rand der Daten, im folgenden *Frame* genannt, als repräsentative Untermenge zu verwenden. Die Hypothese ist, dass für bestimmte lineare Lernverfahren die Datenpunkte am Rand bereits genügend Information für das Lernproblem tragen, sodass eine Restriktion auf den Rand zufriedenstellende Ergebnisse liefert. Dazu betrachten wir zunächst das Problem der Archetypenanalyse (AA) [CB94]. Das Ziel ist eine Matrixfaktorisierung der Datenmatrix in eine Gewichts- und Faktormatrix. Die Idee ist jeden Datenpunkt als Konvexkombination der Faktoren, hier *Archetypen* genannt, darzustellen. Hierbei sind die Archetypen selbst Konvexkombinationen aus Datenpunkten. Die Gewichte können nun als Wahrscheinlichkeitsverteilung interpretiert werden, was eine zusätzliche Interpretierbarkeit der Faktorisierung ermöglicht. Es lässt sich zeigen, dass für diese Matrixfaktorisierung, die Archetypen immer am Rand der konvexen Hülle liegen [CB94].

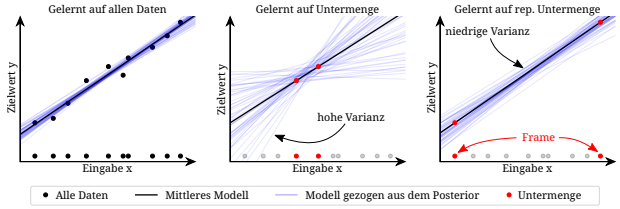


Abb. 3: Ein Beispiel für eine optimale Versuchsplanung in einer Dimension.

In [MBB17] zeigen wir, dass die Restriktion der Berechnung der Archetypenanalyse auf den Frame nahezu die selben Resultate liefert, wie die Berechnung der Faktorisierung auf allen Daten. Siehe dazu Abbildung 2. Zusätzlich wird durch die Restriktion die Berechnung der Archetypenanalyse erheblich beschleunigt.

Der Frame kann mit gängigen Algorithmen zur Bestimmung von konvexen Hüllen berechnet werden. Die Mehrheit dieser Algorithmen ist jedoch für nur zwei- bzw. dreidimensionale Probleme ausgelegt. Verfahren für höherdimensionale Datensätze sind ineffizient, da nicht nur die Randpunkte, sondern auch die überflüssigen Facetten berechnet werden. Als Abhilfe stellen wir in [MBB17] ein neues Verfahren zur Berechnung des Frames in Räumen mit höherer Dimensionalität vor. Unser Ansatz basiert auf quadratischer Programmierung und zeigt eine interessante Verbindung zum Optimierungsalgorithmus NNLS [LH95], welcher zur Lösung von Regressionsproblemen mit nicht-negativen Lösungsvektoren eingesetzt wird.

Ein zweites Lernproblem, welches von der Verwendung des Frames als repräsentative Untermenge profitiert, ist die optimale Versuchsplanung [Fe72]. Hierbei ist eine Menge von Datenpunkten gegeben, welche Experimente beschreiben. Zusätzlich wird ein linearer Zusammenhang zwischen einem parametrisierten Experiment und dem Resultat des durchgeführten Experimentes angenommen. Das Ziel ist es, bei vorgegebener Versuchsanzahl, jene Experimente auszuwählen und diese tatsächlich durchzuführen, sodass mit den Resultaten eine lineare Regression die best möglichste Vorhersage der nicht ausgeführten Experimente liefert.

Wir zeigen in [Ma18], dass auch hier die Einschränkung der Berechnung auf den Rand der Daten (Frame) den Berechnungsaufwand, bei nahezu gleichbleibender Qualität der Vorhersagequalität, deutlich verringert. Abbildung 3 liefert eine Intuition weshalb der Frame eine sinnvolle repräsentative Untermenge liefert. Hier wird die Varianz der Parameterschätzung durch die Wahl von Randpunkten minimiert. Im selben Forschungspapier zeigen wir Verbindungen des Frames zu geometrischen Interpretationen verschiedener Optimalitätskriterien, welche in der optimalen Versuchsplanung vorwiegend verwendet werden. Darüber hinaus, beschäftigen wir uns mit der nicht-linearen Erweiterung der Problemstellung. Die Regression kann durch Verwendung von Kern-Funktionen [SS02] auch nicht-lineare Zusammenhänge abbilden. Diesbezüglich adaptieren wir die Berechnung des Frames und zeigen wie dieser auch in kern-induzierten Merkmalsräumen bestimmt werden kann. Ebenfalls analysieren wir die erwartete Anzahl an Randpunkten für gängige Kerne theoretisch und diskutieren eine Verbindung der Frame-Berechnung mit dem LASSO-Verfahren [Ti96] zur Merkmalsselektion.

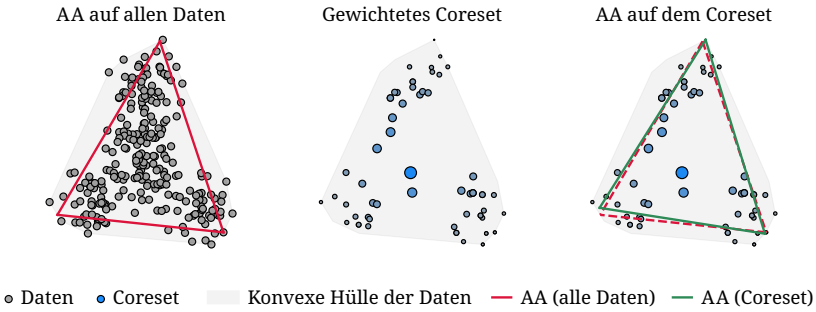


Abb. 4: Eine Archetypenanalyse mit drei Faktoren berechnet auf allen Daten (links), ein gewichtetes Coreset (mitte) und die Archetypenanalyse berechnet auf dem Coreset (rechts).

### 3.2 Coresets für die Archetypenanalyse

Obwohl der Frame als repräsentative Teilmenge den Rechenaufwand der Archetypenanalyse reduziert und dennoch eine kompetitive Lösung geliefert hat, besitzt diese Teilmenge einige Nachteile. Zunächst ist die Frame-Berechnung für hochdimensionale Datensätze problematisch, da die Berechnung polynomiell in der Dimensionalität der Daten skaliert. Zweitens ist die Größe der repräsentativen Teilmenge eine inhärente Eigenschaft des Datensatzes und kann somit nicht frei gewählt werden. Schließlich gibt es keine theoretische Fehlerabschätzung, sowie eine Garantie, dass die Verwendung des Frames tatsächlich eine kompetitive Lösung liefert.

Diesem Problem widmen wir uns speziell für die Archetypenanalyse in [MB19] unter Verwendung des Coreset-Frameworks. Hierbei bezeichnet man eine (gewichtete) Untermenge eines Datensatzes als *Coreset*, wenn ein Modell welches auf der Untermenge trainiert wurde, nachweislich mit dem Modell welches auf allen Daten trainiert wurde, konkurrieren kann. Die Idee ist, dass ein Datenpunkt stellvertretend für mehrere Datenpunkte steht und dementsprechend ein Gewicht zugeordnet wird. Ebenfalls sollen wichtige Bereiche des Eingaberaums mit mehr Punkten abgedeckt werden, als unwichtige. Üblicherweise sind Coresets effizient in linearer Zeit berechenbar und basieren auf probabilistischen Verfahren, welche eine Stichprobenziehung nach Wichtigkeit ermöglichen.

Die Mitte von Abbildung 4 zeigt ein Beispiel eines Coresets mit gewichteten Datenpunkten. Für das Problem der Archetypenanalyse haben wir eine Wahrscheinlichkeitsverteilung vorgeschlagen, welche Datenpunkte am Rand mit höherer Wahrscheinlichkeit, jedoch mit einem kleineren Gewicht auswählt. Im Gegensatz dazu werden Datenpunkte im Zentrum des Datensatzes seltener ausgewählt, bekommen aber ein größeres Gewicht. Für die Berechnung dieser Punkte muss der Algorithmus lediglich zweimal über den Datensatz iterieren. Somit ist die Berechnung der repräsentativen Untermenge sehr effizient. Des Weiteren haben wir eine Abschätzung hergeleitet, welche die Größe des Coresets in Verbindung mit dem Fehler der Approximation und der Wahrscheinlichkeit des Einhaltes dieser Abschätzung setzt.

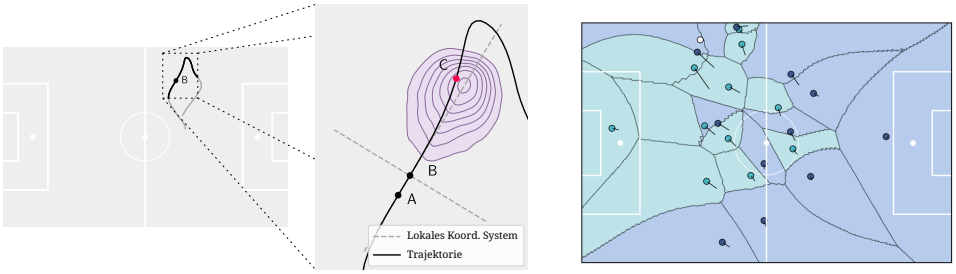


Abb. 5: Ein statistisches Bewegungsmodell (links) und von den Spielern kontrollierte Zonen (rechts).

Neben zahlreichen Experimenten zur empirischen Verifikation der Güte der repräsentativen Teilmenge zeigen wir ebenfalls, dass sich die Zielfunktion der Archetypenanalyse mit der Zielfunktion des $k$-Means Verfahren zur Clusteranalyse [Ll82] abschätzen lässt. Dies zeigt nicht nur eine interessante Verbindung beider unüberwachten Lernverfahren, sondern ermöglicht auch die Verwendung von Coresets, welche explizit für das $k$-Means Verfahren erstellt wurden, für die Archetypenanalyse.

### 3.3 Statistische Bewegungsmodelle

Der zweite Teil der Dissertation behandelt eine Repräsentationsänderung im Bezug auf die Dimensionalität der Daten mit dem Ziel, spezifische Operationen, wie z.B. die Dichteschätzung oder die Interpolation in generativen Modellen, zu verbessern.

Diesbezüglich befassen wir uns nun mit der Dichteschätzung auf Trajektorien in einem raum-zeitlichen Kontext. Speziell geht es um die Erstellung von statistischen Bewegungsmodellen von Objekten, also um die Quantifizierung der Wahrscheinlichkeit für die nächste Position mit einem gewissen Zeithorizont, gegeben verschiedener Kontextinformationen wie z.B. die Bewegungsrichtung, die momentane Geschwindigkeit sowie die Position anderer Objekte. Solche Probleme sind beispielsweise in der Koordination von Spielern in Mannschaftssportarten, beim Studieren von Migrationsmustern oder bei Tierwanderungen anzutreffen. In dieser Arbeit beschäftigen wir uns speziell mit Fußballdaten.

Die linke Seite von Abbildung 5 zeigt ein Beispiel für ein statistisches Bewegungsmodell. Ein Spieler läuft von Position A zu Position B. Das Bewegungsmodell liefert nun, gegeben der Bewegungsrichtung und Geschwindigkeit eines Spielers, eine Verteilung für die Position, welche der Spieler in einer vorgegebenen Zeit erreichen wird. Die tatsächliche Position C ist rot hervorgehoben. Traditionelle Modelle basierend auf physikalischen Annahmen, haben oft unrealistische Nebeneffekte und sind selten personalisiert. In [BLM19] haben wir das in Abbildung 5 (links) gezeigte Bewegungsmodell vorgeschlagen. Das Modell basiert auf Kerndichteschätzern und kann aus Bewegungsdaten personalisiert gelernt werden. Hierbei werden Trajektoriendaten in ein lokales Koordinatensystem überführt, welches die Dichteschätzung drastisch vereinfacht.

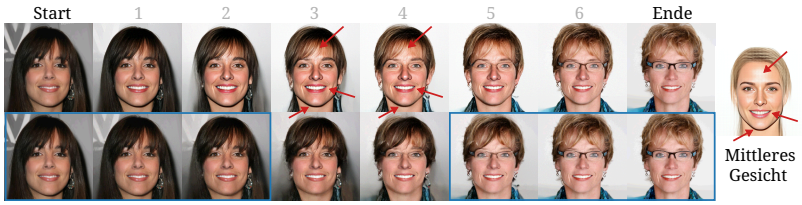


Abb. 6: Zwei Interpolationspfade eines generativen Modells am Beispiel von Gesichtern.

Kerndichteschätzer als nicht-parametrische Methoden haben den Vorteil, dass die Güte der Vorhersage mit jedem zusätzlichen Datenpunkt zunimmt. Der Nachteil ist jedoch, dass der Berechnungsaufwand einer Vorhersage ebenfalls ansteigt. Deshalb haben wir in [Fa21a] das lokale Koordinatensystem mittels invertierbarer neuronaler Netze in geeignete Repräsentationen überführt. Dies hat mehrere Vorteile: die Komplexität einer Vorhersage ist statisch und nicht länger abhängig von der Größe des Datensatzes, die Vorhersagequalität steigt, und es sind komplexere Kontextinformationen möglich. Während das Modell basierend auf Kerndichteschätzern nur die Bewegungsrichtung und Geschwindigkeit berücksichtigt hat, kann das neue Modell in [Fa21a] auch auf die Positionen der anderen Spieler konditioniert werden. Empirische Experimente auf Fußballdaten haben gezeigt, dass durch die Anreicherung an Kontextinformation die Vorhersagegüte drastisch steigt.

In [BLM19] haben wir ebenfalls gezeigt, wie sich statistische Bewegungsmodelle zur Berechnung von kontrollierten Zonen auf dem Spielfeld verwenden lassen. Hierbei ist eine, von einem Spieler kontrollierte Zone, jener Bereich, welcher von diesem Spieler am wahrscheinlichsten zuerst erreicht wird. Ein Beispiel ist in Abbildung 5 auf der rechten Seite abgebildet. Jeder Spieler hat einen eigenen kontrollierten Bereich. Die Striche geben an, aus welcher Richtung der Spieler kommt. Solche kontrollierten Zonen ermöglichen eine neuartige, datengetriebene Analyse von Spielen.

### 3.4 Interpolieren in generativen Modellen

Im Beitrag [Fa21b] beschäftigen wir uns mit generativen Modellen basierend auf tiefen neuronalen Netzen. Wie in Abbildung 1 (rechts) angedeutet, ist die Idee Daten, welche einer komplizierten Verteilung folgen, in eine andere Repräsentation zu überführen, welche vorgegebenen Strukturen folgt. Abbildung 6 zeigt zwei Interpolationspfade eines generativen Modells, welches eine Standardnormalverteilung im latenten Raum annimmt. Dabei sind die Gesichter am Rand normale Datenpunkte während die sechs Gesichter dazwischen synthetische Bilder entlang der Interpolationspfade sind. Rechts der Interpolation ist der dekodierte Erwartungswert des generativen Modells abgebildet. Es zeigt somit das „mittlere Gesicht“. Der obere Interpolationspfad resultiert aus einer einfachen linearen Interploation. Wie mit den roten Pfeilen hervorgehoben, nehmen die Gesichter in der Mitte Eigenschaften des Erwartungswertes an. So ist z.B. eine glänzende Stirn in keinem der Ursprungsbilder,

jedoch im „mittlere Gesicht“ vorhanden. Der Grund hierfür ist, dass in hochdimensionalen Räumen standardnormalverteilte Datenrepräsentationen einen gewissen Abstand zum Zentrum des Koordinatensystems haben. Interpolanten einer linearen Interpolation haben jedoch einen wesentlich geringeren Abstand und sind somit dem Zentrum näher. Dieses Zentrum ist allerdings der Erwartungswert der Standardnormalverteilung. Im unteren Interpolationspfad in Abbildung 6 wird zusätzlich der Abstand der Interpolanten zum Zentrum interpoliert. Damit behält der Interpolationspfad den erwarteten Abstand zum Zentrum und das vorherige Problem tritt nicht länger auf. Allerdings gibt es ein neues Problem. Da sich die Interpolationsgeschwindigkeit ändert, gibt es eine Verzerrung hin zu den Randbildern (in blau hervorgehoben) und die Interpolation wirkt ungleichmäßig.

Diesem Problem widmen wir uns in [Fa21b]. Wir schlagen für den latenten Raum die Verwendung einer Einheitssphäre vor, in welchem die Interpolationen als geodätische Wege effizient berechnet werden können. Dabei verwenden wir eine stereografische Projektion um die Daten aus einem euklidischen Raum in die Einheitssphäre zu transformieren. Auf der Einheitssphäre verwenden wir eine von Mises-Fisher Verteilung als Pendant einer Normalverteilung. Zahlreiche quantitative und qualitative Experimente zeigen, dass die somit gewonnenen Interpolationspfade natürlicher Wirken und nicht von den in Abbildung 6 gezeigten Problemen betroffen sind. Dabei wird durch die Verbesserung der Interpolationsqualität die Leistungsfähigkeit des generativen Modells nicht beeinflusst. Die Qualität der Erstellung beliebiger neuer Datenpunkte bleibt erhalten.

## 4 Fazit und Ausblick

Diese Dissertation befasst sich mit dem unüberwachten Lernen von effizienten Datenrepräsentationen. Dabei wurde das Problem aus zwei orthogonalen Sichtweisen betrachtet: (i) in Bezug auf den Stichprobenumfang und (ii) hinsichtlich der Dimensionen eines jeden Datenpunktes. Zu beiden Sichtweisen wurden mehrere Beiträge geleistet und neue Ansätze, Methoden, Verbindungen zwischen unterschiedlichen Lernproblemen, sowie Probleme und deren Lösungen aufgezeigt. Die einzelnen Beiträge wurden durch Experimente und theoretischen Analysen untermauert.

Zukünftige Arbeiten widmen sich einer Kombination beider Sichtweisen, anstatt diese separat voneinander zu betrachten. So sind in Bezug auf die stetig wachsende Datenbasis effiziente Repräsentationen ebendieser notwendig, die sowohl in Anzahl, als auch in Dimensionalität kompakter sind, dabei aber nicht an Aussagekraft verlieren.

## Literatur

[BLM19] Brefeld, U.; Lasek, J.; Mair, S.: Probabilistic movement models and zones of control. Machine Learning 108/1, S. 127–147, 2019.

[CB94] Cutler, A.; Breiman, L.: Archetypal analysis. Technometrics 36/4, S. 338–347, 1994.

[Fa21a] Fadel, S. G.; Mair, S.; da Silva Torres, R.; Brefeld, U.: Contextual movement models based on normalizing flows. AStA Advances in Statistical Analysis/, S. 1–22, 2021.

[Fa21b] Fadel, S. G.; Mair, S.; da Silva Torres, R.; Brefeld, U.: Principled Interpolation in Normalizing Flows. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, S. 116–131, 2021.

[Fe72] Fedorov, V. V.: Theory of optimal experiments. Elsevier, 1972.

[LH95] Lawson, C. L.; Hanson, R. J.: Solving least squares problems. SIAM, 1995.

[Ll82] Lloyd, S.: Least squares quantization in PCM. IEEE transactions on information theory 28/2, S. 129–137, 1982.

[Ma18] Mair, S.; Rudolph, Y.; Closius, V.; Brefeld, U.: Frame-Based Optimal Design. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, S. 447–463, 2018.

[Ma22] Mair, S.: Computing Efficient Data Summaries, Diss., Leuphana Universität Lüneburg, 2022.

[MB19] Mair, S.; Brefeld, U.: Coresets for Archetypal Analysis. In: Advances in Neural Information Processing Systems. S. 7247–7255, 2019.

[MBB17] Mair, S.; Boubekki, A.; Brefeld, U.: Frame-based data factorizations. In: International Conference on Machine Learning. PMLR, S. 2305–2313, 2017.

[SS02] Schölkopf, B.; Smola, A. J.: Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2002.

[Ti96] Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological)/, S. 267–288, 1996.

**Sebastian Mair**, geboren 1988 in Dillingen an der Donau, studierte zunächst Informatik an der Hochschule Darmstadt, bevor er ein Masterstudium in Informatik sowie ein Bachelorstudium in Mathematik an der Technischen Universität Darmstadt absolvierte. Im Anschluss promovierte er im maschinellen Lernen unter der Betreuung von Ulf Brefeld an der Leuphana Universität in Lüneburg, an welcher er auch als Wissenschaftlicher Mitarbeiter beschäftigt war. Seine Promotion schloss er im September 2021 mit *summa cum laude* ab. Im Dezember 2021 startete er als Postdoktorand an der Universität Uppsala in Schweden. Seine Forschungsinteressen sind unüberwachtes Lernen, effiziente Datenrepräsentationen, generative Modellierung und statistisches maschinelles Lernen.