Gesellschaft für Informatik e.V. (GI) GI. (Hrsg.): INFORMATIK 2021,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2021 1729

Where is the Science in Data Science Projects?

Online-Workshop über die Wissenschaftlichkeit von Vorgehensmodellen für Data-Science-Projekte (WISDAP)

Jens Kaufmann[1], Stephan Kühnel[2], René Theuerkauf[3], Emal M. Alekosai[4], Sayed Hoseini[5], Uwe Neuhaus[6], Michael Schulz[7]

Abstract: Als Ergebnis einer virtuellen Arbeitsgruppe aus Wissenschaftler:innen und Praktiker:innen entstand zwischen April 2019 und Februar 2020 das Data-Science-Vorgehensmodell DASC-PM, dessen Ziel es ist, vorhandenes Wissen über die Durchführung von Data-Science-Projekten für alle Interessensgruppen in geeigneter Form zu strukturieren. Unter Berücksichtigung dieses Modells, aber nicht exklusiv darauf gestützt, soll der Aspekt der Wissenschaftlichkeit bei der Umsetzung von Data-Science-Projekten im Rahmen des Workshops besonders betrachtet werden.

Keywords: Data Science, DASC-PM, Wissenschaftlichkeit, Rigor, Relevance

1 Einleitung

Daten und ihre Auswertung haben in der modernen Wirtschaftsgesellschaft einen so großen Platz eingenommen, dass sich nicht nur der Begriff der „Datenökonomie“ entwickelt hat, sondern Rankings über die Leistungsfähigkeit selbiger in den Ländern der Welt erstellt werden [BCS19]. Die Nachfrage nach datenzentrierten Leistungen hat entsprechend zugenommen und mir ihr gestiegen ist auch die Komplexität der Aufgaben in den Bereichen Datenaufbereitung und Datenanalyse. Diese Begriffe und Aufgaben können zusammen mit weiteren verbundenen Tätigkeiten als *Data Science* zusammengefasst werden [Sc20a].

[1] Hochschule Niederrhein, Fachbereich Wirtschaftswissenschaften, Webschulstr. 41-43, 41065 Mönchengladbach, Deutschland jens.kaufmann@hs-niederrhein.de
[2] Martin-Luther-Universität Halle-Wittenberg, Lehrstuhl für Wirtschaftsinformatik, insb. Betriebliches Informationsmanagement, Universitätsring 3, 06108 Halle (Saale), Deutschland, stephan.kuehnel@wiwi.uni-halle.de
[3] Martin-Luther-Universität Halle-Wittenberg, Lehrstuhl für Wirtschaftsinformatik, insb. E-Business, Universitätsring 3, 06108 Halle (Saale), Deutschland, rene.theuerkauf@wiwi.uni-halle.de
[4] Robert Bosch GmbH, Borsigstr. 4, 70469 Stuttgart, Deutschland,mohammademal.alekozai@de.bosch.com
[5] Hochschule Niederrhein, Fachbereich Elektrotechnik und Informatik, Reinarzstr. 49, 47805 Krefeld, Deutschland, sayed.hoseini@hs-niederrhein.de
[6] NORDAKADEMIE Hochschule der Wirtschaft, Fachbereich Informatik, Köllner Chaussee 11, 25337 Elmshorn, Deutschland, uwe.neuhaus@nordakademie.de
[7] NORDAKADEMIE Hochschule der Wirtschaft, Fachbereich Informatik, Köllner Chaussee 11, 25337 Elmshorn, Deutschland, michael.schulz@nordakademie.de

1.1 Data Science als Anwendungs- und Forschungsfeld

Traditionelle Datenanalysemethoden stoßen bei den beschriebenen Anforderungen häufig an die Grenze ihrer Möglichkeiten. Ausschlaggebend dafür sind mehrere Faktoren. So sind auf der einen Seite große Mengen an Daten frei verfügbar, was die Menge der Eingaben in die Verfahren stark erhöht, zum anderen sind diese aber häufig von heterogener Gestalt und Qualität sowie von unstrukturierter Form [He14]. Insbesondere durch diese häufig als *Big Data* bezeichnete Datenmenge ergeben sich bei der Integration von Daten neue Herausforderungen. Neben numerischen Daten auch Text, Audio und Video verarbeiten zu können, erfordert bei der Analyse eine Kombination aus Herangehensweisen verschiedener Disziplinen wie Mathematik und Informatik zur Bewältigung der reinen technischen und analytischen Schwierigkeiten sowie Wissen über die Anwendungsdomäne. Der so entstehende Dreiklang wird bei [Co10] beschrieben und häufig als Ausgangsdefinition der Data Science verwendet.

Während die bisher beschriebenen Aspekte noch als komplexere Formen etablierter Datenanalyse und ihrer Anwendung im betrieblichen Kontext aufgefasst werden können, zeigen weitere an diese angrenzende und Einfluss nehmende Entwicklungen, dass sich mit Data Science ein eigenes Forschungsfeld etabliert, das sich von bestehenden Methoden und Beschreibungen z. B. des Data-Mining abgrenzen lässt. Als ein Beispiel können Rechnersysteme genannt werden, die als Agenten eingesetzt werden, Daten automatisch auswerten und algorithmisch Entscheidungen treffen. Die dabei entstehenden Fragestellungen umfassen neben rein technischen und wirtschaftlichen Aspekten auch Betrachtungen der Ethik und des Datenschutzes [ABH17], welche eines umfassenden wissenschaftlichen Diskurses bedürfen. Ein weiterer Aspekt findet sich dann, wenn Techniken eingesetzt werden, die eine mangelnde Erklärbarkeit aufweisen und für den Nutzer selbst bei Kenntnis über die eingesetzte Methodik (z. B. neuronale Netze) der eigentliche Entscheidungsprozess auf den zugrunde liegenden Daten nicht nachvollziehbar ist. Dies ist insofern problematisch, da die Erklärung einer getroffenen Entscheidung ein Kernbestandteil menschlicher Kommunikation und Lernprozesse ist [Go18].

Data Science hat daher nicht nur die Aufgabe, Erkenntnisgewinn aus Daten zu ermöglichen, die notwendigen technischen wie analytischen Grundlagen zu schaffen oder bereitzustellen und die Analyseergebnisse mit fachlichem Wissen zu handlungsbefähigenden Interpretationen umzusetzen. Vielmehr muss sie auch dem Kern ihrer wörtlichen Bedeutung, der *Wissenschaft*, Respekt gezollt werden. Gemeint ist dabei allerdings nicht, dass ausschließlich Wissenschaftler die entsprechenden Methoden einsetzen sollen oder dass nur besonders zertifizierte Methodiken zum Einsatz kommen dürfen. Sie ist allerdings verpflichtet, nicht nur die reine Ergebnisproduktion in den Vordergrund zu stellen, sondern den Ansprüchen an Objektivität, Reliabilität und Validität nachzukommen, die auch von Praktikern gefordert wird und werden muss, um zum einen die Ergebnisse sicher und nachvollziehbar nutzen und präsentieren zu können und zum anderen die verwendeten Methoden korrekt auszuwählen oder bei Bedarf neu zu erstellen. Dies ist umso entscheidender, wenn beispielsweise die

Eignung bestimmter Verfahren für einen gegebenen Kontext geprüft werden soll und damit der Prozess des Erkenntnisgewinns im Vordergrund steht oder im Allgemeinen Verfahren aus dem akademischen Kontext in der Praxis für die konkrete Anwendung adaptiert werden.

1.2 Wissenschaftlichkeit der Data Science

Dem Workshop ging eine Studie der Workshop-Organisatoren voraus, die unter anderem vorhandene und gewünschte Kompetenzen sowie die aktuelle Nutzung von Vorgehensmodellen im Data-Science-Kontext untersuchte. Diese Studie motivierte konkret dazu, die Wissenschaftlichkeit der Data Science weiter aufzuarbeiten. So ist ein methodisches Vorgehen eine häufig von den Studienteilnehmer:innen verlangte Kompetenz von Data Scientitsts. Zudem waren Wissenschaftler:innen mit einem großen Anteil in der Studie vertreten und gaben gleichzeitig an, Projekte durchzuführen, sodass sich mindestens, aber, wie vorher beschrieben, nicht ausschließlich, für diese Anwendergruppe die Notwendigkeit ergibt, Projektergebnisse auch wissenschaftlich angemessen aufzubereiten und nachvollziehbar zu gestalten, was die Disziplin selbst beeinflusst. Abschließend sind Vorgehensmodelle ein tradiertes Element wissenschaftlicher Methodik (siehe Abschnitt 4 dieses Beitrags) und das hier verwendete das *Data Science Process Model* (DASC-PM) als Vorgehensmodell soll aufzeigen, wie sein Einsatz die Wissenschaftlichkeit der Data Science unterstützt.

Es ist in diesem Zusammenhang ein viel diskutiertes Thema, Wissenschaft gegen die Praxis oder wissenschaftliche Strenge (Rigor) gegen die praktische Relevanz der Ergebnisse zu stellen. Tatsächlich schließen sich beide Thematiken nicht aus und sind auch mit der praktischen Relevanz, die die Anwender von Data Science mutmaßlich vordergründig interessiert, problemlos zu vereinbaren [Di12]. Wie also die Wissenschaft, die „Science“ in Data Science zu interpretieren ist, ist Gegenstand des in diesem Beitrag beschriebenen Workshops.

Dabei sollen aus Sicht von Wissenschaftler:innen und Praktiker:innen drei Kernfragen adressiert werden:

1. Was kennzeichnet die Wissenschaftlichkeit in der Disziplin Data Science?
2. Wie kann ein wissenschaftliches Vorgehen in datengetriebenen Projekten adäquat gehandhabt und sichergestellt werden?
3. Welche Kernaspekte sind bei der Durchführung, Dokumentation und Evaluation von Data-Science-Projekten aus wissenschaftlicher Sicht zu berücksichtigen?

Im Rahmen des Workshops sollen diese Fragen nicht einfach nur beantwortet werden, vielmehr ist es ein Kernziel, aus den gewonnenen Erkenntnissen Guidelines abzuleiten, die Wissenschaftler:innen und Praktiker:innen bei der Adressierung wissenschaftlicher Aspekte im Rahmen von Data-Science-Projekten angemessen unterstützen.

1.3 Aufbau des Beitrags

Neben den allgemein gültigen Überlegungen und Vorkenntnissen der Teilnehmer:innen des Workshops bildet das Vorgehensmodell DASC-PM eine weitere Grundlage der Diskussion. Dieses wird im zweiten Abschnitt des Beitrags nochmal zusammenfassend vorgestellt. Der dritte Abschnitt beschreibt eine Studie der Workshops-Organisatoren zu vorhandenen und gewünschten Kompetenzen sowie zur aktuellen Nutzung von Vorgehensmodellen im Data-Science-Kontext. Die Ergebnisse der Studie werden als rahmengebendes Element in die Diskussion eingebunden und liefern den Teilnehmer:innen wertvolle Erkenntnisse über die Anforderungen und Einschätzungen von Wissenschaftler:innen und Praktiker:innen zur aktuellen Anwendung und Durchführung von Data-Science-Projekten, sodass die Fragestellung des Workshops nach der Wissenschaftlichkeit im Vorgehen keine Betrachtung zum Selbstzweck fördert, sondern für die Überführung in praktische Handlungsanweisungen verwendbare Ergebnisse hervorbringt. Abschließend wird in Abschnitt vier das Workshopkonzept vorgestellt. Dieses wird methodisch aus den zuvor aufgearbeiteten Ergebnissen und Anforderungen an wissenschaftliches Vorgehen abgeleitet. Abschnitt und Beitrag schließen mit einem Ausblick auf weitere Schritte.

2 DASC-PM – Ein Modell für Data-Science-Aktivitäten

Data Science wird – wie ein wesentlicher Teil der Aktivitäten in Unternehmen – häufig in Projektform betrieben. Entsprechend ist das Umfeld dieser Projekte ebenso mit einzubeziehen wie die spezifische Ausgestaltung der Aktivitäten selbst. Da Data Science mehr umfasst als reine Analyseverfahren, benötigt sie auch ein eigenes, geeignetes Vorgehensmodell [Sc20b]. Um dieses Ziel zu erreichen, haben 22 Personen aus Wissenschaft und Wirtschaft in einer virtuellen Arbeitsgruppe ein Prozessmodell für Data-Science-Projekte entwickelt. Dort wurden umfangreiche Interviews zu den Aspekten der Disziplin sowie zu Kompetenz- und Rollendefinitionen geführt. Die Ergebnisse wurden im Anschluss von allen Mitgliedern der Arbeitsgruppe evaluiert. Dieser Prozess aus Interviews und Aufarbeitungen wurde in einem iterativen Verfahren zehn Mal wiederholt, sodass die Ergebnisse jedes Durchlaufs in die darauf folgenden mit eingeflossen sind. In einem abschließenden Aufarbeitungsprozess wurden die Arbeitsergebnisse zusammengeführt und angereichert. Das Resultat ist das *Data Science Process Modell* (DASC-PM) - ein Vorgehensmodell, welches Data-Science-Projekte anhand von Schlüsselbereichen beschreibt und die Kompetenzen sowie Rollen der Beteiligten darstellt. Der so dokumentierte Prozess zu Data-Science-Projekten sowie die Kompetenzbereische sind in Abbildung 1 dargestellt. Das Modell selbst inklusive der hier vorgestellten Auszüge wird in [Sc21] beschrieben.

Das Modell beschreibt in der Arbeitsgruppe ausgewählte Aspekte der Data Science in einer möglichst umfassenden, nicht aber abschließenden Form. Insbesondere da die Data Science ein sehr aktives Forschungsfeld darstellt, ist DASC-PM keineswegs ein statisches Modell, sondern ein dynamisches Konstrukt, das sich z. B. durch die Ergebnisse von

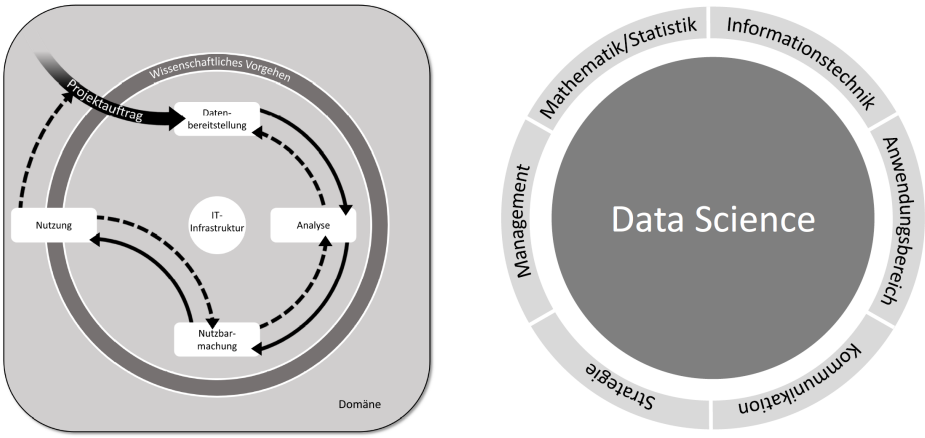

Abb. 1: (a) Data-Science-Vorgehensmodell DASC-PM (b) Nötige Kompetenzen in einem Data-Science-Projekt, beide aus [Sc21]

Workshops weiter verändern wird. Der in diesem Beitrag beschriebene Workshop soll seinen Anteil leisten, inhaltliche Lücken in Bezug auf den Aspekt der Wissenschaftlichkeit bei der Durchführung von Projekten zu schließen.

3 Studie

Im Zeitraum vom 30. November 2020 bis zum 21. Dezember 2020 wurde eine Umfrage zu Data-Science-Projekten durchgeführt. Adressiert wurden (unter Zuhilfenahme von E-Mail-Verteilerlisten und Social-Media-Kanälen) Personengruppen, die bereits Data-Science-Erfahrungen sammeln konnten. Durch dieses Vorgehen ist die genaue Zahl der Empfänger:innen des Aufrufes zur Teilnahme nicht zu benennen. Als Anreiz für die Beteiligung an der Umfrage wurde die Zurverfügungstellung der aufbereiteten Ergebnisse in Aussicht gestellt.

3.1 Informationen zu der Umfragedurchführung

Insgesamt kam es zu 107 Aufrufen der Umfrage. In 67 Fällen wurde die Bearbeitung unterbrochen. Ein Abbruchmuster war dabei nicht festzustellen. 40 Teilnehmer:innen haben die Umfrage beendet. Das arithmetische Mittel der Bearbeitungszeit lag bei 14,87 Minuten, der Median bei 10,45 Minuten. Sämtliche Ergebnisse wurden anonymisiert gespeichert. Keiner der Datensätze musste auf Grund üblicher Filterkriterien (zu kurze Bearbeitungszeit, auffälliges Bearbeitungsmuster, etc.) aus dem Datenset entfernt werden.

Abbildung 2 fasst einige Informationen zu den Teilnehmer:innen zusammen. 15 Personen arbeiten bei kleinen Unternehmen mit maximal 250 Mitarbeiter:innen, elf bei Unternehmen mit 251 bis 2000 Mitarbeiter:innen und 14 bei Unternehmen mit mehr als 2000 Mitarbeitenden. Mit einer Anzahl von 21 besaß etwas mehr als die Hälfte der Teilnehmer:innen keine Führungsposition, bei 19 Teilnehmer:innen war eine Führungsposition gegeben. Ebenfalls 19 Personen haben angegeben, große Erfahrungen im Data-Science-Bereich zu besitzen (Einschätzung 4 oder 5 auf einer Skala von "1 – n. v."bis "5 - Experte"), sieben eher wenig (Einschätzung 1 oder 2). 37 der 40 Teilnehmer:innen haben bereits mindestens ein Data-Science-Projekt mit durchgeführt, 23 bereits mehrere.

Von den Teilnehmer:innen schätzten sich acht am ehesten als Berater, elf als Wissenschaftler, 14 als Praktiker und sieben als Lernende ein. Letztere gehören jedoch vermutlich nicht zu einer Gruppe „Studierende“, sondern sind einer der zuvor genannten Gruppen zuzuordnen, betrachten sich selbst aber offenbar nicht als erfahren genug, um sich „Praktiker“ zu nennen. Diese Annahme wird auch dadurch gestützt, dass alle Teinehmer:innen dieser Gruppe angeben, bereits Projekte durchgeführt zu haben, drei davon in Unternehmen.

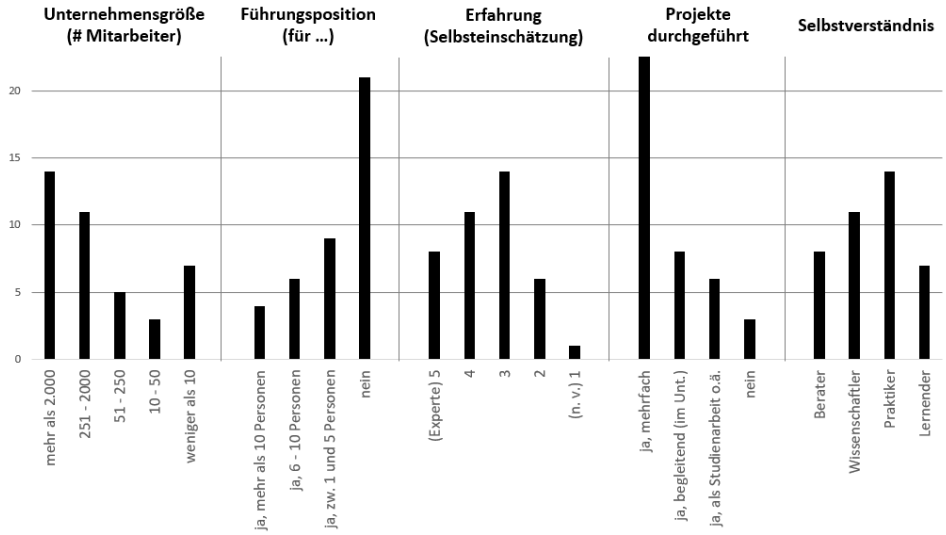

Abb. 2: Informationen zu den Teilnehmer:innen an der Umfrage

3.2 Status quo: Data-Science-Projekte und Kompetenzen

Der erste inhaltliche Themenblock der Umfrage beschäftigte sich damit, wer Data-Science Projekte anstößt (Management, Fachabteilung, Dats-Science-Team, IT-Abteilung, Externe) und auf welche Geschäftsfelder (Hauptgeschäftsfeld, weitere bestehende Geschäftsfelder, neue Geschäftsfelder) die Use-Cases der Data-Science-Projekte vorwiegend einzahlen. Dabei wurde zwischen den aktuell vorhandenen Gegebenheiten in der Organisation der Umfrageteilnehmer:innen (Ist-Zustand) und den angestrebten Verhältnissen (Soll-Zustand) unterschieden.

Aktuell ist häufig ein dediziertes Data-Science-Team Initiator eines Data-Science-Projekts (34 %), gefolgt von den Fachabteilungen (27 %) und dem Management (21 %). Von geringerer Bedeutung sind Externe (14 %) und die IT-Abteilung (4 %). Für den Soll-Zustand wünschen sich die Umfrageteilnemer:innen, dass Data-Science-Projekte noch häufiger von den Fachabteilungen (35 %) und dem Management (25 %) angestoßen werden, dem Data-Science-Team kommt in dieser Frage aber weiterhin eine wichtige Rolle zu (26 %). Der Einfluss externer Projekt-Initiatoren sinkt entsprechend (7 %), die IT-Abteilung ist weiterhin nur gelegentlich als Ideengeberin für Projekte vorgesehen (7 %).

Die Mehrheit der Use-Cases von Data-Science-Projekten zahlt bislang auf das Hauptgeschäftsfeld des Unternehmens oder der Organisation ein (51 %), auf weitere bestehende Geschäftsfelder entfallen immerhin 29 %. Nur 20 % der Use-Cases zielen bislang auf neue oder neu zu entwickelnde Geschäftsfelder. Beim Soll-Zustand ändert sich dies, dort sollen Haupt- und neue Geschäftsfelder in etwa gleich stark durch Data-Science-Projekte unterstützt werden (36 % bzw. 34 %).

Als nächstes wurden die Umfrageteilnehmer:innen zu den Projektdimensionen *Teamgröße*, *Laufzeit* und *Budget* befragt. Die überwiegende Mehrheit aller Data-Science-Projektteams (77 %) umfasst zwischen 3 und 5 Personen, 17 % sind kleiner und 7 % größer. Die typische Projektdauer beträgt zwischen drei und sechs Monaten (52 %), 28 % der Projekte sind bereits nach ein bis zwei Monaten abgeschlossen, 17 % dauern länger als sechs Monate. Auch sehr kurze Projekte (Laufzeit von weniger als einem Monat) gibt es gelegentlich (3 %). Dabei könnte es sich um Proof-of-Concepts handeln. Das Budget von Data-Science-Projekten kann sehr unterschiedlich ausfallen. In 36 % der in der Umfrage genannten Fälle betrug das Budget weniger als 20.000 Euro, in 32 % zwischen 20.000 und 50.000 Euro und in 21 % zwischen 50.000 und 100.000 Euro. Nur wenige Projekte verfügten über größeres Budget (100.000 - 200.000 Euro: 4 %; mehr als 200.000 Euro: 7 %).

Stark ausgeprägt ist gemäß der Umfrageergebnisse der Zusammenhang von Projekt- und Unternehmensgröße. Kleinere Unternehmen ($\leq$ 250 Mitarbeiter:innen) haben tendenziell kleinere Projektteams, kürzere Projektlaufzeiten und geringere Budgets. In größeren Unternehmen ($>$ 250 Mitarbeiter:innen) findet sich hingegen die ganze Projektbandbreite, angefangen von kurzen und preiswerten Proof-of-Concepts bis zu langlaufenden Projekten mit großem Data-Science-Teams und entsprechendem Projektbudget.

Ein weiterer Block der Umfrage befasste sich mit den für ein Data-Science-Projekt benötigten Kompetenzen. Zuerst wurden die Teilnehmer:innen befragt, ob die im weiteren Projektverlauf benötigten Kompetenzen bereits zu Projektstart bekannt sind. Nur 16 % von ihnen haben diese Frage vollständig oder überwiegend bejaht. 62 % brachten hingegen zum Ausdruck, dass die benötigten Kompetenzen zu Projektstart noch nicht oder größtenteils nicht vorhanden sind. Dies spiegelt den häufig stark ausgeprägten Forschungscharakter von Data-Science-Projekten wieder, bei denen das konkrete Vorgehen und die eingesetzten Methoden vielfach erst während der Laufzeit festgelegt werden können.

Unabhängig von der Frage, ob die benötigten Kompetenzen bereits feststehen, ergibt sich bei Data-Science-Projekten häufig ein weiteres Problem: Sind die Mitarbeiter:innen mit den entsprechenden Kompetenzen im Unternehmen bzw. in der Organisation verfügbar? Nur 32 % der Umfrageteilnehmer:innen haben diese Frage vollständig oder überwiegend bejaht. 55 % sagten aus, dass entsprechend qualifiziertes Personal ganz oder überwiegend fehlt.

Um ein möglichst detailliertes Bild davon zu erhalten, an welchen konkreten Kompetenzen ein Mangel herrscht, wurden die Teilnehmer:innen zu einem vorher definierten Data-Science-Kompetenzspektrum befragt, das an die Ergebnisse von [Sc21] angelehnt ist und die folgenden elf Kompetenzen umfasst: *Mathematik*, *Methoden*, *Daten*, *Programmierung*, *Systemadministration*, *Anwendbarkeit*, *Domaine*, *Kommunikation*, *Strategie*, *Projektmanagement*, *Riskmanagement*. Die Teilnehmer:innen sollten dazu Auskunft geben, ob sie die jeweilige Kompetenz generell für relevant in Data-Science-Projekten halten, in welchem Maße die Kompetenz im eigenen Unternehmen / in der eigenen Organisation vorhanden ist und in welchem Maße die Kompetenz tatsächlich benötigt wird. Für die letzten beiden Fragen wurde ein Skala von 1 (Kompetenz überhaupt nicht vorhanden) bis 5 (Kompetenz vollständig vorhanden) verwendet.

Alle elf aufgeführten Kompetenzen wurden von den Teilnehmer:innen als relevant für Data-Science-Projekte eingestuft. Wie Abbildung 3 zeigt, gibt es jedoch erhebliche Diskrepanzen zwischen den benötigten und den tatsächlich vorhandenen Kompetenzen. In allen elf Bereichen liegt der Wert für die vorhandenen Kompetenzen unter dem der benötigten. Am stärksten ausgeprägt ist diese Differenz bei den Data-Science-Methoden, der Mathematik und der Programmierung. Die kleinsten Unterschiede ergaben sich bei der Befragung bei den Punkten Projektmanagement, Domäne und Systemadministration.

3.3 Vorgehensmodelle in der Data Science

Bei der Durchführung von Data-Science-Projekten werden häufig Vorgehensmodelle eingesetzt (vgl. Kapitel 2), die einen strukturierten Projektablauf mit vorgegebenen Phasen und definierten Aufgaben unterstützen. In der Umfrage wurde daher erfragt, welche dieser Vorgehensmodelle bekannt sind, ob sie in der Praxis tatsächlich eingesetzt werden und wie die Umfrageteilnehmer:innen die Nützlichkeit der Modelle generell einschätzen.

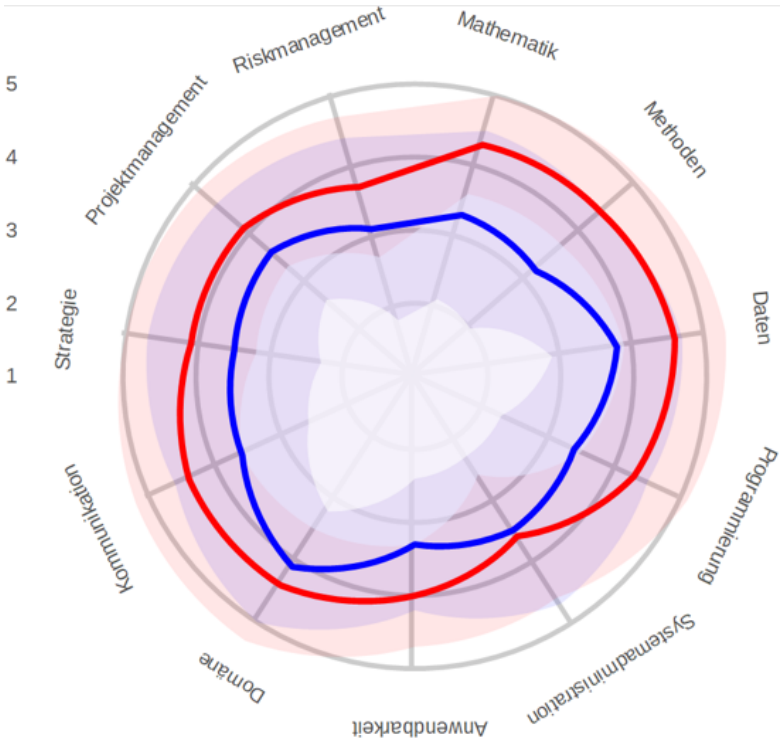

Abb. 3: Differenzen zwischen vorhandenen (blau) und benötigten (rot) Kompetenzen.
(Skala: 1 = überhaupt nicht vorhanden (Kreismittelpunkt) bis 5 = vollständig vorhanden (Kreisrand))

Fast alle Umfrageteilnehmer:innen (85 %) kennen mindestens ein Vorgehensmodell für Data-Science-Projekte, viele sogar zwei (58 %), drei (38 %), vier (25 %) oder noch mehr (10 %). Die beiden am häufigsten angegebenen Modelle sind dabei der *Cross-Industry Standard Process for Data Mining (CRISP-DM)* und der *Knowledge Discovery in Data Bases Process (KDD)*. Sie sind 75 % bzw. 40 % der Teilnehmer:innen bekannt. Der Bekanntheitsgrad des *Team Data Science Process (TDSP)*, der *Analytics Solutions Unified Method for Data Mining/Predictive Analytics (ASUM-DM)* und von *Sample, Explore, Modify, Model, and Assess (SEMMA)* ist mit 18 %, 13 % und erneut 13 % hingegen geringer.

Die große Mehrheit (83 %) der Umfrageteilnehmer:innen sind vollständig oder überwiegend der Auffassung, dass der Einsatz eines Vorgehensmodells die Erfolgswahrscheinlichkeit von Data-Science-Projekten erhöht. Nur 8 % stimmten eher nicht zu. Der Aussage, dass der Einsatz von Vorgehensmodellen allgemein der Standard bei Data-Science-Projekten ist, haben allerdings nur 30 % der Umfrageteilnehmer:innen ganz oder überwiegend zugestimmt, 53 % gaben an, dass dies bislang nicht bzw. überwiegend nicht der Fall ist. Eine mögliche Erklärung dafür könnte sein, dass für Data-Science-Projekte andere,

allgemeinere Vorgehensmodelle genutzt werden. Dies wäre aufgrund der Erfahrungen der Teilnehmer:innen jedoch weniger empfehlenswert, denn 60 % sind der Auffassung, dass sich ein Vorgehensmodell für Data-Science-Projekte von einem allgemeinen Vorgehen für Projekte unterscheidet. Nur 15 % waren eher nicht dieser Meinung.

4 Konzept des WISDAP-Workshops und weitere Schritte

Wissenschaftlichkeit („Science“) und rigoroses wissenschaftliches Vorgehen nehmen in der Data Science bereits auf Basis der Gebietsbezeichnung eine entscheidende Rolle ein. Die Ergebnisse der Studie zeigen, dass die Teilnehmer:innen sowohl vor dem Hintergrund ihrer eigenen Tätigkeit als auch vor der geplanten Anwendung von Ergebnissen durchgeführter Data-Science-Aktivitäten einen Bedarf sehen, die Verlässlichkeit von Algorithmen und Analysen durch Nutzung stringenter Methoden und Vorgehensmodelle zu sichern.

4.1 Vorgehensmodelle in der Wissenschaft

Dass Vorgehensmodelle in der (Wirtschafts-)Informatik eine hohe Relevanz besitzen, zeigt sich bereits an der Existenz einer entsprechenden Fachgruppe *Vorgehensmodelle für die betriebliche Anwendungsentwicklung (WI-VM)* innerhalb der Gesellschaft für Informatik e.V. [Ge21]. Dabei werden Vorgehensmodelle üblicherweise als eine Form von Referenzmodellen beschrieben ([Br12; VF16]), die ihrerseits eine Standard-Methode zur Erstellung von IT-Artefakten im Rahmen der Forschung in der Wirtschaftsinformatik darstellt [Fr17]. Tatsächlich existieren anerkannte Vorgehensmodelle zur Entwicklung von Referenzmodellen [Fe14]. Die Verwendung von Vorgehensmodellen kann daher als etabliertes Vorgehen nicht nur im praktischen, sondern auch im wissenschaftlichen Kontext angesehen werden.

Die Frage nach der Wissenschaftlichkeit der Data Science wird aus diesem Grund mit Hilfe dreier Kernfragen (siehe Abschnitt 1) und anhand des Vorgehensmodells DASC-PM untersucht, sodass die gewonnenen Erkenntnisse auch am Modell in handlungsunterstützende Artefakte überführt werden können. Dass diese konkrete Form eines Ergebnisses auch im wissenschaftlichen Kontext angemessen sein kann, zeigt sich beispielsweise bei vom Brocke et al. (2015), die ihre Richtlinien zur Literatursuche um eine Tabelle im Checklisten-Format ergänzen [Br15].

4.2 Konzept des Workshops

Der Workshop verfolgt den Ansatz des User-centered Design (UCD), der Design-Prozesse beschreibt, bei denen potentielle Anwender:innen frühzeitig aktiv in den Design-Prozess einbezogen werden [AMP04]. Der Ansatz des UCD wurde bereits bei der initialen Konzeption

des DASC-PM verfolgt, wobei die Erfahrungen und das Wissen von Wissenschaftler:innen und Praktiker:innen, die als potentielle Anwender:innen des DASC-PM angesehen werden, erhoben und berücksichtigt wurden. Im Einklang mit diesem Vorgehen werden auch im WISDAP-Workshop beide Nutzergruppen einbezogen. Da bei der Untersuchung ein besonderer Fokus auf den unterschiedlichen Sichtweisen von Wissenschaft und Praxis liegt, erscheint eine initiale Unterteilung der Workshop-Teilnehmer:innen in Wissenschaftler:innen und Praktiker:innen als sinnvoll, um gegenseitige Beeinflussungen zu vermeiden. Innerhalb der beiden Gruppen - Wissenschaft und Praxis - werden schließlich moderierte Fokusgruppendiskussionen durchgeführt.

Die moderierte Fokusgruppe wird als eine bedeutende Methode des UCD angesehen [AMP04]. Dabei handelt es sich um eine qualitative Forschungsmethode, bei der ein:e Moderator:in eine Gruppendiskussion leitet, die sich auf die Interaktion zwischen den Teilnehmern:innen stützt, um neue Erkenntnisse zu generieren [Mo97]. Die erforderliche Anzahl moderierter Fokusgruppen, die für verlässliche Ergebnisse notwendig ist, ist in der Literatur stark umstritten. Basierend auf den empirischen Ergebnissen von Guest et al. [GNM17] sind zwei Fokusgruppen jedoch als ausreichend anzusehen.

Das Vorgehen in beiden moderierten Fokusgruppen besteht aus insgesamt vier Schritten:

1) Einführung in das Forschungsthema,
2) Vorstellung des Problemkontextes,
3) Diskussion über die Fragen 1-3 (vgl. Kapitel 1) und
4) Überführung der Diskussionsergebnisse in Guidelines.

4.3 Weiteres Vorgehen

Im Nachgang des Workshops werden die gewonnenen Ergebnisse zusammengeführt und der offenen und virtuellen DASC-PM Arbeitsgruppe zu Evaluationszwecken vorgestellt. Hierbei soll geprüft werden, ob die Guidelines verständlich, vollständig, nachvollziehbar, praktisch anwendbar und nützlich sind oder ob es weiterer Überarbeitungen bedarf.

Die Erkenntnisse fließen in die regelmäßige Weiterentwicklung des Vorgehensmodells ein. Soweit sie allgemeingültig scheinen, sollen sie der wissenschaftlichen Gemeinschaft auch durch geeignete weitere Publikationen zur Diskussion und Weiterentwicklung zur Verfügung gestellt werden.

Literatur

[ABH17] van der Aalst, W. M.; Bichler, M.; Heinzl, A.: Responsible Data Science. Business and Information Systems Engineering 59/5, S. 311–313, 2017.

[AMP04] Abras, C.; Maloney-Krichmar, D.; Preece, J.: User-Centered Design. In (Bainbridge, W. S., Hrsg.): Berkshire Encyclopedia of Human-Computer Interaction. Berkshire Publishing Group, S. 763–768, 2004.

[BCS19] Bhalla, A.; Chakravorti, B.; Shankar Chaturvedi, R.: Which Countries Are Leading the Data Economy?, 2019, URL: https://hbr.org/2019/01/which-countries-are-leading-the-data-economy, Stand: 20. 07. 2021.

[Br12] Breitner, M. H.: Vorgehensmodell - Enzyklopädie der Wirtschaftsinformatik, 2012, URL: https://www.enzyklopaedie-der-wirtschaftsinformatik.de/wi-enzyklopaedie/lexikon/is-management/Systementwicklung/Vorgehensmodell, Stand: 20. 07. 2021.

[Br15] vom Brocke, J.; Simons, A.; Riemer, K.; Niehaves, B.; Plattfaut, R.; Cleven, A.: Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research. Communications of the Association for Information Systems 37/, S. 205–224, 2015.

[Co10] Conway, D.: The Data Science Venn Diagram, 2010, URL: http://www.dataists.com/2010/09/the-data-science-venn-diagram/, Stand: 22. 06. 2021.

[Di12] Dilger, A.: Rigor, wissenschaftliche und praktische Relevanz. In: Diskussionspapier des Instituts für Organisationsökonomik. Diskussionspapier des Instituts für Organisationsökonomik 03/2012, Institut für Organisationsökonomik, Westfälische Wilhelms-Universität Münster, 2012, URL: https://www.wiwi.uni-muenster.de/io/sites/io/files/forschen/downloads/dp-io_03_2012.pdf.

[Fe14] Fettke, P.: Eine Methode zur induktiven Entwicklung von Referenzmodellen. In: Tagungsband Multikonferenz Wirtschaftsinformatik 2014. S. 1034–1047, 2014.

[Fr17] Friedrich, T.; Schlauderer, S.; Weidinger, J.; Raab, M.: On the Research Paradigms and Research Methods Employed in the BISE Journal - A Ten-Year Update. In: Proceedings of Wirtschaftsinformatik 2017. S. 1111–1125, 2017.

[Ge21] Gesellschaft für Informatik e.V.: Startseite FG WI-VM, 2021, URL: https://fg-wi-vm.gi.de/, Stand: 20. 07. 2021.

[GNM17] Guest, G.; Namey, E.; McKenna, K.: How Many Focus Groups Are Enough? Building an Evidence Base for Nonprobability Sample Sizes. Field Methods 29/1, S. 3–22, 2017.

[Go18] Goebel, R.; Chander, A.; Holzinger, K.; Lecue, F.; Akata, Z.; Stumpf, S.; Kieseberg, P.; Holzinger, A.: Explainable AI: The new 42? 2018.

[He14] Hendler, J.: Data integration for heterogenous datasets. Big Data 2/4, S. 205–215, 2014.

[Mo97] Morgan, D. L.: Focus groups as qualitative research. Sage Publ, Thousand Oaks, Calif., 1997.

[Sc20a] Schulz, M.: Data-Science-Projekte und ihre Besonderheiten. Wirtschaftsinformatik & Management 12/5, S. 376–381, 2020.

[Sc20b] Schulz, M.; Neuhaus, U.; Kaufmann, J.; Badura, D.; Kuehnel, S.; Badewitz, W.; Dann, D.; Kloker, S.; Alekozai, E. M.; Lanquillon, C.: Introducing DASC-PM: A Data Science Process Model. In: ACIS 2020 Proceedings. 45. 2020.

[Sc21] Schulz, M.; Neuhaus, U.; Kaufmann, J.; Badura, D.; Kerzel, U.; Welter, F.; Prothmann, M.; Kühnel, S.; Passlick, J.; Rissler, R.; Badewitz, W.; Dann, D.; Gröschel, A.; Kloker, S.; Alekozai, E. M.; Felderer, M.; Lanquillon, C.; Brauner, D.; Gölzer, P.; Binder, H.; Rhode, H.; Gehrke, N.: DASC-PM v1.0 - Ein Vorgehensmodell für Data-Science-Projekte. 2021.

[VF16] Vom Brocke, J.; Fettke, P.: Referenzmodell - Enzyklopädie der Wirtschaftsinformatik, 2016, URL: https://www.enzyklopaedie-der-wirtschaftsinformatik.de/lexikon/is-management/Systementwicklung/Softwarearchitektur/Wiederverwendung-von-Softwarebausteinen/Referenzmodell/index.html, Stand: 20. 07. 2021.