DASC-PM v1.0 Ein Vorgehensmodell für Data-Science-Projekte

Schulz, Michael; Neuhaus, Uwe; Kaufmann, Jens; Badura, Daniel; Kerzel, Ulrich; Welter, Felix; Prothmann, Maik; Kühnel, Stephan; Passlick, Jens; Rissler, Raphael; Badewitz, Wolfgang; Dann, David; Gröschel, Alexander; Kloker, Simon; Alekozai, Emal M.; Felderer, Michael; Lanquillon, Carsten; Brauner, Dorothee; Gölzer, Philipp; Binder, Harald; Rohde, Heiko; Gehrke, Nick

Michael Schulz ▪ Uwe Neuhaus

DASC-PM v1.0

Ein Vorgehensmodell für Data-Science-Projekte

Jens Kaufmann ▪ Daniel Badura ▪ Ulrich Kerzel ▪ Felix Welter ▪ Maik Prothmann ▪ Stephan Kühnel ▪ Jens Passlick ▪ Raphael Rissler ▪ Wolfgang Badewitz ▪ David Dann ▪ Alexander Gröschel ▪ Simon Kloker ▪ Emal M. Alekozai ▪ Michael Felderer ▪ Carsten Lanquillon ▪ Dorothee Brauner ▪ Philipp Gölzer ▪ Harald Binder ▪ Heiko Rohde ▪ Nick Gehrke

gefördert durch

valantic

NORDAKADEMIE
HOCHSCHULE DER WIRTSCHAFT

Hamburg, **Elmshorn** 2020

Michael Schulz

michael.schulz@nordakademie.de

Uwe Neuhaus

uwe.neuhaus@nordakademie.de

ISBN e-Book: 978-3-00-064898-4

Herausgeber:

valantic Business Analytics GmbH

Beim Strohhause 17

20097 Hamburg

NORDAKADEMIE gAG Hochschule der Wirtschaft

Köllner Chaussee 11

25337 Elmshorn

VORWORT

Das Thema Data Science hat in den letzten Jahren in vielen Organisationen stark an Aufmerksamkeit gewonnen. Häufig herrscht jedoch weiterhin große Unklarheit darüber, wie diese Disziplin von anderen abzugrenzen ist, welche Besonderheiten der Ablauf eines Data-Science-Projekts besitzt und welche Kompetenzen vorhanden sein müssen, um ein solches Projekt durchzuführen.

In der Hoffnung, einen kleinen Beitrag zur Beseitigung dieser Unklarheiten leisten zu können, haben wir von April 2019 bis Februar 2020 in einer offenen und virtuellen Arbeitsgruppe mit Vertretern aus Theorie und Praxis das vorliegende Dokument erarbeitet, in dem ein Vorgehensmodell für Data-Science-Projekte beschrieben wird – das Data Science Process Model (DASC-PM). Ziel war es dabei nicht, neue Herangehensweisen zu entwickeln, sondern vielmehr, vorhandenes Wissen zusammenzutragen und in geeigneter Form zu strukturieren. Die Ausarbeitung ist als Zusammenführung der Erfahrung sämtlicher Teilnehmerinnen und Teilnehmer dieser Arbeitsgruppe zu verstehen.

Als Zielgruppe des Dokumentes sind all diejenigen zu sehen, die direkt oder aber auch indirekt an Data-Science-Projekten beteiligt sind. Grundlegende Kenntnisse über den Komplex der analytischen Informationssysteme werden dabei vorausgesetzt. Das Vorgehensmodell soll dazu dienen, allen Interessengruppen von Data-Science-Projekten ein Verständnis der notwendigen Aufgaben und Zusammenhänge zu vermitteln. Zudem kann es von Studierenden genutzt werden, um sich dem Themenfeld zu nähern.

Die Data Science befindet sich noch am Anfang ihrer Entwicklung. Deshalb soll dieses Dokument nicht als abgeschlossenes Werk betrachtet werden. Wir wünschen uns sehr, dass es zukünftig in der Durchführung von Data-Science-Projekten Berücksichtigung findet. Dadurch gewonnene Erkenntnisse sollen sowohl genutzt werden, um die bestehenden Ausarbeitungen in Frage zu stellen, als auch, um sie zu vervollständigen und zu detaillieren.

Falls Sie Verbesserungsvorschläge zum Vorgehensmodell haben oder sich aktiv an seiner Weiterentwicklung beteiligen möchten, freuen wir uns über eine Kontaktaufnahme. Das nächste Treffen der virtuellen Arbeitsgruppe ist für September 2020 geplant.

Unser Dank gilt allen Teilnehmerinnen und Teilnehmern der Arbeitsgruppe. In produktiver und konstruktiver Atmosphäre haben wir ein unserer Meinung nach nutzbringendes und verständnisförderndes Ergebnis erzielt – und dabei auch selbst viel Neues über Data Science gelernt.

Hamburg, im Februar 2020

Uwe Neuhaus und Michael Schulz

Kontakt: michael.schulz@nordakademie.de

1. EINLEITUNG

Trotz gestiegener Aufmerksamkeit fehlt derzeit eine allgemein akzeptierte und einheitliche Definition der Data Science. Während der Begriff in wissenschaftlichen Veröffentlichungen vieler Disziplinen noch selten verwendet wird, fallen Definitionen aus der Praxis vor allem durch ihre Heterogenität auf. Dieser Sachverhalt wiederum führt zu sehr unterschiedlichen Erwartungen und möglichen Missverständnissen bei den beteiligten Personengruppen. Von den Teilnehmern der Arbeitsgruppe wurden wiederholt zwei Definitionen genannt:

Definition 1:

> „*Data science is an interdisciplinary field aiming to turn data into real value. Data may be structured or unstructured, big or small, static or streaming. Value may be provided in the form of predictions, automated decisions, models learned from data, or any type of data visualization delivering insights. Data science includes data extraction, data preparation, data exploration, data transformation, storage and retrieval, computing infrastructures, various types of mining and learning, presentation of explanations and predictions, and the exploitation of results taking into account ethical, social, legal, and business aspects.*" (van der Aalst, 2016)

Definition 2:

> „ *At a high level, data science is a set of fundamental principles that support and guide the principled extraction of information and knowledge from data. Possibly the most closely related concept to data science is data mining - the actual extraction of knowledge from data via technologies that incorporate these principles. There are hundreds of different data mining algorithms, and a great deal of detail to the methods of the field. We argue that underlying all these many details is a much smaller and more concise set of fundamental principles.*
>
> *(…)*
>
> *Data science involves principles, processes, and techniques for understanding phenomena via the (automated) analysis of data. (…) the ultimate goal of data science is improving decision making, as this generally is of paramount interest to business. (…) Data-driven decision making refers to the practice of basing decisions on the analysis of data rather than purely on intuition."* (Provost & Fawcett, 2013)

Auch wenn es sich bei dem Zitat von Provost und Fawcett nicht um eine Definition im klassischen Sinn handelt, umreißt es doch zentrale Aspekte der Data Science.

Die beiden aufgeführten Definitionen besitzen zwei Nachteile: Zum einen umfassen sie nicht alle Aspekte, die von den Teilnehmerinnen und Teilnehmern unserer Arbeitsgruppe als wichtig erachtet werden, zum anderen sind sie recht wortreich, vermischen Aspekte von sehr unterschiedlichem Detailgrad und verstellen dadurch den Blick auf die zentralen Konzepte der Data Science. Auf Basis der Teilnehmerbeiträge unserer Arbeitsgruppe empfehlen wir folgende, prägnantere Definition, die sich auf die übergeordneten Aspekte der Data Science konzentriert:

> *Data Science ist ein interdisziplinäres Fachgebiet, in welchem mit Hilfe eines wissenschaftlichen Vorgehens, semiautomatisch und unter Anwendung bestehender oder zu entwickelnder Analyseverfahren Erkenntnisse aus teils komplexen Daten extrahiert und unter Berücksichtigung gesellschaftlicher Auswirkungen nutzbar gemacht werden.*

Im Folgenden sollen die einzelnen Bestandteile der Definition einzeln betrachtet werden.

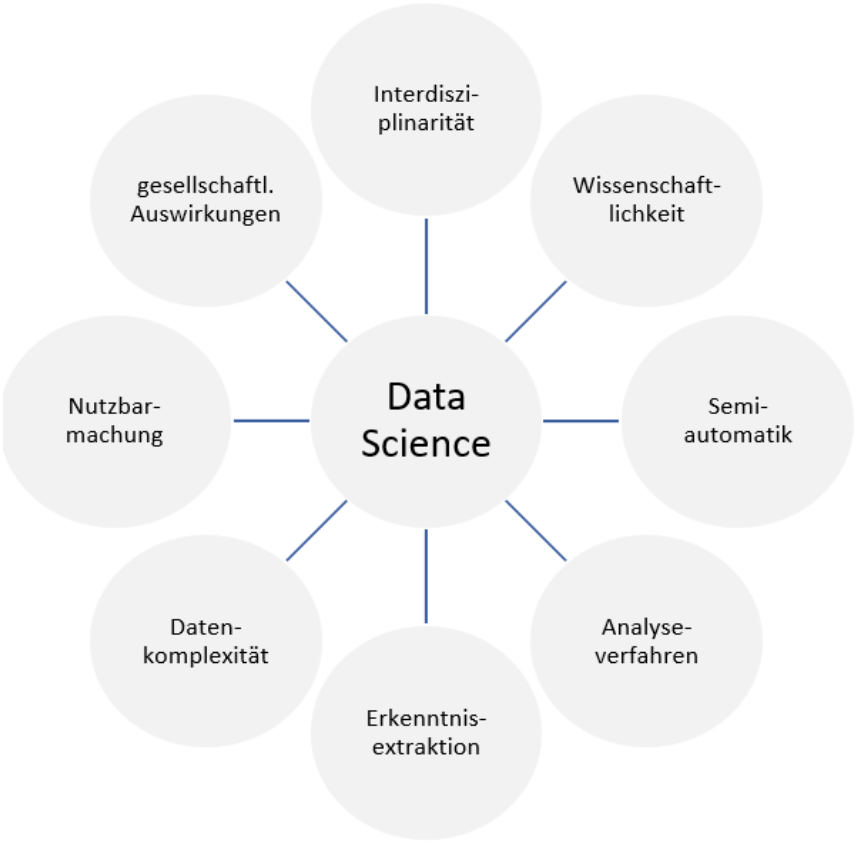


Abbildung 1: Merkmale der Data Science

Wie in der Definition von van der Aalst (2016) bereits benannt, wird die Tatsache, dass es sich bei der Data Science um ein **interdisziplinäres Fachgebiet** handelt, für sehr relevant angesehen. Dies ist in der häufig notwendigen starken Kooperation verschiedener Forschungsdisziplinen wie etwa von Mathematik (insbesondere Statistik und Numerik), Informatik, Künstlicher Intelligenz und Linguistik begründet. In all diesen Disziplinen existieren bereits seit langem Ansätze zur wissenschaftlichen Auseinandersetzung mit Daten (Provost & Fawcett, 2013). Das unter dem Namen *Data Science* entstandene Fachgebiet scheint dort begründet worden zu sein, wo die Mittel der traditionellen Disziplinen nicht mehr ausgereicht haben, um aktuellen Herausforderungen (z. B. größeren, häufig unstrukturierten oder sich dynamisch ändernden Datenmengen) zu begegnen. Das rasante Wachstum an öffentlich verfügbaren Informationen durch das Internet, der Preisverfall für Computerspeicher und das Wachstum an Rechenkapazität ermöglichen es, zunehmend komplexere Analyseverfahren anzuwenden (McAfee & Brynjolfsson, 2012), was immer mehr zu der Herausbildung einer eigenen Fachdisziplin führt.

Weiterhin sind im Kontext der Interdisziplinarität auch die verschiedenen Domänen zu nennen, in denen die Anwendung der Data Science als unterstützender Wissenschaft von Interesse ist. Obwohl in vielen Texten zu diesem Thema das Domänenumfeld traditionell auf die Betriebswirtschaft beschränkt wird, ist dies nicht mehr zu rechtfertigen. Andere unterstützende Wissenschaften, wie die Mathematik oder die Informatik, sind ebenfalls nicht auf die Anwendung innerhalb einer bestimmten Domäne beschränkt, sodass eine weite Auslegung einer Disziplin weder neu noch problematisch erscheint. In Domänen wie der Biologie, der Medizin, der Physik, der Astronomie und vielen mehr ist die Anwendung der Data Science nicht nur als hilfreich zu bewerten, sondern auch bereits vorzufinden.

Der Data-Science-Begriff lässt bereits erkennen, dass auf ein **wissenschaftliches Vorgehen** abgezielt wird. Diese Wissenschaftlichkeit spiegelt sich unter anderem häufig in dem Ziel eines allgemeinen Erkenntnisgewinns wider. Nicht immer steht die direkte Nutzung der Analyseergebnisse im Fokus, sondern die Untersuchung allgemeinerer Aspekte wie beispielsweise die Eignung von Verfahren für bestimmte Fragestellungen, die Aussagekraft einzelner Verfahren bezogen auf die zu Grunde liegende Datenbasis oder die Bewertung der Komplexität verschiedener Verfahren. Weiterhin zeichnet sich die Wissenschaftlichkeit dadurch aus, dass die untersuchte Problemstellung nicht trivial ist, die Verfahren vollständig verstanden sein sollten und objektiv nachvollziehbar, reproduzierbar, dokumentiert und systematisch angewandt werden. Aus Unternehmenssicht ist ein weiterer Aspekt der Wissenschaftlichkeit darin zu sehen, dass Analyseverfahren aus dem wissenschaftlichen Umfeld übernommen werden, was im betrieblichen Umfeld häufig eine Neuerung darstellt. Die tatsächliche Tiefe der wissenschaftlichen Auseinandersetzung, auch hier vor allem bezogen auf die Anwendung im betriebswirtschaftlichen Kontext, variiert, ist abhängig von der Domäne und kann sich auf ein „ingenieurmäßiges“ Vorgehen beschränken.

Bei der Betrachtung von Datenanalysen unter Verwendung eines wissenschaftlichen Vorgehens wird häufig auch der Begriff *Data Mining* genannt, der teilweise synonym zum Data-Science-Begriff verwendet wird. Dies liegt unter anderem in den beiden bekannten Data-Mining-Vorgehensmodellen *Knowledge Discovery in Databases* (KDD) (Fayyad et al., 1996) – siehe Abbildung 2 – und *Cross Industry Standard Process for Data Mining* (CRISP-DM) (Wirth & Hipp, 2000) – siehe Abbildung 3 – begründet, deren Anwendung zumindest in betriebswirtschaftlichen Domänen stark verbreitet ist und auch dem Data Mining ein in Grenzen strukturiertes „wissenschaftliches“ Vorgehen auferlegt. Dies hatte wiederum Einfluss auf das Begriffsverständnis. Beim KDD-Prozess wird nur ein einzelner Teilschritt, nämlich die eigentliche Datenanalyse, als *Data Mining* bezeichnet. Es existieren aber auch Data-Mining-Definitionen, welche die datenorientierten Prozessschritte und auch die Aufgabe der Untersuchung von Analyseergebnissen enthalten. Der CRISP-DM fügte mit dem Business Understanding zusätzlich noch explizit einen nicht-technischen, anwendungsspezifischen Prozessschritt hinzu.

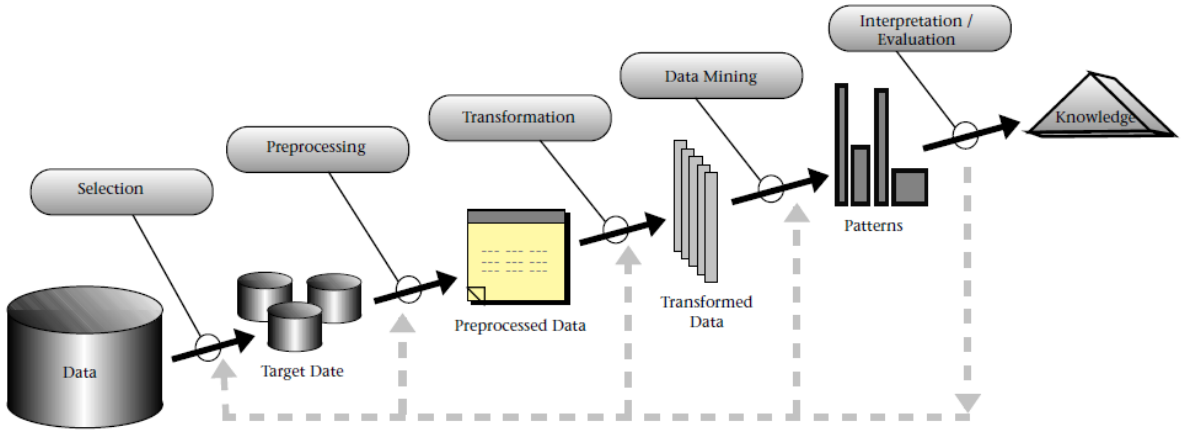


Abbildung 2: Knowledge Discovery in Databases, aus: Fayyad et al., 1996

Die Notwendigkeit, die zunächst eng gefasste Datenanalyse um ein geeignetes Vorgehensmodell zu ergänzen, ist nachvollziehbar. Diese Aufweitung des ursprünglichen Verständnisses des Data-Mining-Begriffs erschwert jedoch dessen einheitliche Verwendung. Die Grenzen zur Data Science verschwimmen.

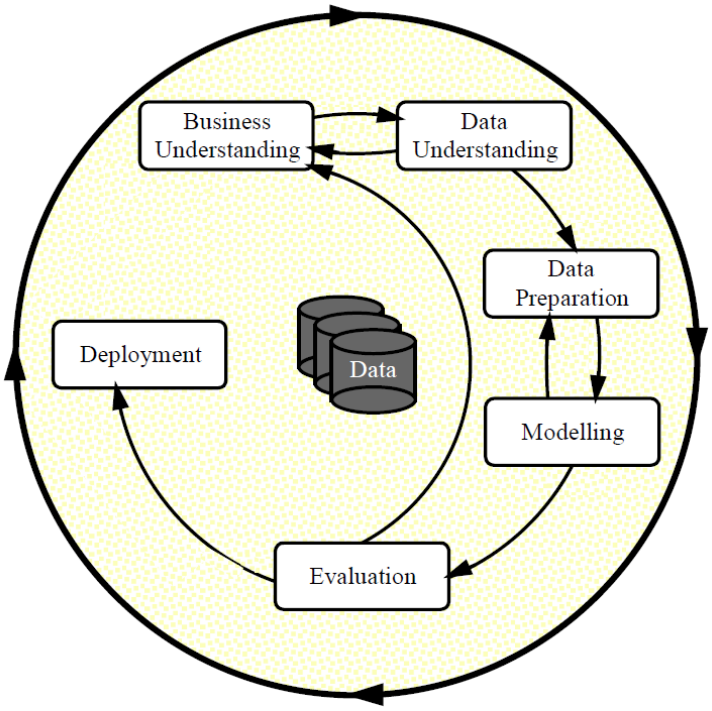


Abbildung 3: Cross Industry Standard Process for Data Mining, aus: Wirth & Hipp, 2000

Auch der Bereich des *Operations Research* besitzt eine große Ähnlichkeit zur Data Science, durch die eine Abgrenzung häufig erschwert wird: Beide Bereiche basieren auf konkreten, wissenschaftlichen Verfahren, die auf Daten angewandt werden, nutzen also ggf. auch Data-Mining-Methoden, um ein Datenverständnis zu erlangen. Spezifisch für das Operations Research ist jedoch die Fokussierung auf die Optimierung des untersuchten Systems. Operations Research besitzt ferner in den meisten Definitionen eine Domäneneinschränkung, die bei der Data Science, wie bereits diskutiert, nicht gesehen wird.

Durch die Nennung von Algorithmen(-gruppen) innerhalb der Data-Science-Definition, wie beispielsweise bei van der Aalst (2016), wird implizit eine Einschränkung vorgenommen, auf die an dieser Stelle verzichtet werden soll. Auf Grund der immer noch hohen Entwicklungsgeschwindigkeit des Fachgebiets besteht bei einer Aufzählung von Algorithmen die Gefahr, dass diese schnell veraltet sind. Selbst der Begriff *Algorithmus* an sich ist im gegebenen Zusammenhang nicht ideal, da nicht alle Algorithmen als Teil von Data Science zu sehen sind. Aus diesem Grund wird in der Definition der Begriff **Analyseverfahren** verwendet, welcher in der Kombination mit einer Anwendung auf Daten den Kern der Data Science adressiert. Es können unter anderem hypothesenprüfende und hypothesenfreie Analysen mit deskriptivem, prädiktivem und präskriptivem Ziel durchgeführt werden. Möglich ist sowohl unüberwachtes als auch überwachtes Lernen. Hierbei können der Zweck von Data Science das Aufdecken von Mustern, Trends und Zusammenhängen sowie die Optimierung sein. Abhängig von der gegebenen Problemstellung können bestehende Analyseverfahren verwendet werden. Es kann jedoch auch nötig sein, Analyseverfahren weiterzuentwickeln oder vollständig neu zu entwickeln, da keine geeigneten Ansätze existieren.

Die Anwendung von Analyseverfahren erfolgt **semiautomatisch**, umfasst also sowohl menschliche als auch maschinelle Arbeitsschritte. Neben der Tatsache, dass Verfahren in der Regel nicht vollständig automatisiert werden können, sind an dieser Stelle auch hybride Lernverfahren zu nennen, die speziell dafür entwickelt werden, um Problemen im Zusammenspiel von Expertenwissen und Analyseverfahren zu begegnen (Olivotti et al., 2018). Häufig, jedoch nicht ausschließlich sind hierfür hochleistungsfähige Hard- und Software-Plattformen nötig,

welche in Kombination eine komplexe Infrastruktur bilden. Abhängig vom konkreten Szenario kann eine vollständige Automatisierung angestrebt werden. Um diese vollständige Automatisierung zu erreichen, sind aber vorbereitende manuelle Arbeitsschritte notwendig. Auch ist ein Erkenntnisgewinn letztendlich nur durch menschliche Beteiligung zu erreichen.

Die **Extraktion von Erkenntnissen aus häufig komplexen Daten** ist als Ziel der Data Science anzusehen. Daten unterscheiden sich in ihrer Struktur, ihrer Qualität, ihrer Vollständigkeit, ihrer Größe und ihrer Dimensionalität. Zudem kann es sich um statische Daten oder Datenströme handeln. Außerdem können Daten in komplexen Beziehungen zueinander stehen. Auch wenn die Entwicklung der Data Science stark durch den Anstieg der Datenmengen getrieben wird (Dhar, 2013) und die Analyse sehr großer Datenbestände die Schaffung neuer Verfahren erfordert, ist die Data Science nicht auf Big-Data-Anwendungen beschränkt. Bevor Analyseverfahren auf Daten angewendet werden können, müssen diese aus den Quellsystemen extrahiert, aufbereitet und bereitgestellt werden. Auch hierfür werden häufig komplexe Infrastrukturen benötigt.

Data Science beinhaltet nicht nur die Extraktion von Erkenntnissen, sondern zusätzlich auch die **Nutzbarmachung**, also den Einsatz der Erkenntnisse. Diese kann sowohl aus einer Bereitstellung der Erkenntnisse für Domänenexperten oder andere Abnehmer bestehen als auch in der Integration in bestehende Systeme und/oder der Automatisierung der Anwendung auf neue Daten. Verschiedene Autoren stellen bei Data-Science-Projekten explizit die Schaffung eines ökonomischen Wertes in den Vordergrund. Wir sprechen in der Definition jedoch allgemeiner von Nutzbarmachung, um neben wirtschaftlichen Zielsetzungen etwa auch rein wissenschaftliche abzudecken.

Sowohl die semiautomatische Extraktion der Erkenntnisse, die Komplexität der Datenbereitstellung und -aufbereitung als auch eine spätere Nutzbarmachung als Software-System erfordern bei Data-Science-Projekten häufig die Bereitstellung oder Entwicklung einer spezifischen IT-Infrastruktur. Diese umfasst Hard- und Software-Komponenten, die an die konkreten Rahmenbedingungen des Projekts angepasst werden müssen. Stichwörter sind hierbei etwa skalierbare Architekturen, Arbeit mit verteilten Daten oder Cloud-Anbindung. Die hierfür benötigten spezifischen IT-Kompetenzen werden oftmals von Projektmitarbeitern eingebracht, die Data Engineers genannt werden. Diese Form der Arbeitsteilung erlaubt es Analyse- und IT-Experten, sich auf ihre speziellen Aufgabenbereiche zu konzentrieren.

Die Auseinandersetzung mit den **gesellschaftlichen Auswirkungen** der Data Science mithilfe einer aktiven Teilnahme am Diskurs zu sich ergebenden ethischen und rechtlichen Fragestellungen in Bezug sowohl auf die Analyseergebnisse als auch auf Daten als Rohmaterial der Analysen soll ebenfalls berücksichtigt werden.

2. FESTLEGUNG GRUNDLEGENDER ANFORDERUNGEN AN DAS DATA-SCIENCE-VORGEHENSMODELL

Bei der Entwicklung eines Vorgehensmodells für Data-Science-Projekte sind Anforderungen ganz unterschiedlicher Natur zu berücksichtigen. Generell soll durch die Verwendung eines Vorgehensmodells die Qualität von Data-Science-Projekten erhöht werden. Das Durchlaufen sämtlicher Schritte – von der Projektkonzeption bis zur Nutzbarmachung der gewonnenen Erkenntnisse – ist dabei zu dokumentieren. Insbesondere muss erkennbar sein, an welcher Stelle Erkenntnisse durch die Anwendung von Analyseverfahren gewonnen und Interpretationen durch Domänenwissen ergänzt werden. Dadurch kann eine Reproduzierbarkeit, Wiederverwendbarkeit und Generalisierbarkeit der Ergebnisse sichergestellt werden. Zudem muss das Modell skalierbar sein, um Projekte unterschiedlicher Größe zu unterstützen. Hierzu wird eine Unterscheidung zwischen auszuführenden Projektaktivitäten und qualitativen Anforderungen an die Projektkoordination und -organisation vorgenommen. Es gilt, Besonderheiten jeder einzelnen Phase abzubilden und klare Handlungsempfehlungen für den weiteren Projektverlauf zu geben. Hierfür ist es nötig, klare Richtlinien festzulegen, wie das Modell in einem konkreten Anwendungsfall einzusetzen ist.

Bei der Entwicklung eines Vorgehensmodells ist die Wahl der Abstraktionsebene der enthaltenen Aufgaben von hoher Relevanz. Ist die gewählte Abstraktionsebene zu hoch, resultiert nur ein geringer Nutzen, der sich auf die konzeptionelle Ebene beschränkt. Ist die gewählte Abstraktionsebene zu niedrig, erschwert dies sowohl die Verallgemeinerbarkeit des Modells, welche gerade durch die verschiedenen Einsatzgebiete der Data Science so wichtig ist, als auch die Verständlichkeit, was wiederum die Akzeptanz des Modells gefährdet. Eine hohe Komplexität führt zu einem bewussten oder unbewussten Auslassen von Aufgaben und stellt damit die generelle Nutzung eines standardisierten Vorgehens in Frage. Durch eine Aufteilung in Ebenen unterschiedlicher Abstraktionsgrade bleibt die Übersichtlichkeit des Modells gewahrt und es kann zugleich eine Hilfestellung in Detailfragen geboten werden. Auf niedrigerer Abstraktionsebene kann auch eine Modularisierung sinnvoll sein. In der konkreten Anwendung können irrelevante Modellbausteine somit übersprungen werden, ohne dass dies einen wesentlichen Einfluss auf den Projektablauf hat. Ein solches Überspringen benötigt allerdings eine geeignete und dokumentierte Begründung, um die Nachvollziehbarkeit der Ergebnisgenerierung nicht zu gefährden. Alternativ zu einer Modularisierung können spezialisierte Varianten des Vorgehensmodells entstehen – abhängig von der betrachteten Domäne und/oder den eingesetzten Analyseverfahren. Wichtig ist hierbei, dass auf jeder Abstraktionsebene und in jeder Form der Instanziierung geeignete Schnittstellen zwischen den einzelnen Modellbausteinen definiert werden.

Zunächst soll in diesem Dokument ein Modell auf hoher Abstraktionsebene erarbeitet werden, das für den Einsatz in sämtlichen Data-Science-Projekten geeignet ist. Eine Detaillierung des Vorgehensmodells bis hin zu expliziten Templates, Prozeduren oder sogar Skripten ist in späteren Arbeiten denkbar, würde aber in einem ersten Schritt zu weit führen. Der Fokus dieser Arbeitsgruppe liegt nicht auf der abschließenden Festlegung und Beschreibung aller Details eines Vorgehensmodells. Vielmehr sollen durch den Einsatz des Modells in realen Projekten neue Erkenntnisse gewonnen und mögliche Probleme identifiziert werden, die im Rahmen eines kontinuierlichen Weiterentwicklungsprozesses Berücksichtigung finden.

Große Projekte erfordern ein Team von Experten aus unterschiedlichen Bereichen, die sich gegenseitig ergänzen und deren Zusammenarbeit durch das Vorgehensmodell unterstützt wird. Dabei sollen alle Projektbeteiligten Berücksichtigung finden. Die Projektbeteiligten sollen in der Lage sein, unter Zuhilfenahme des Modells ihre eigenen Aufgaben zu identifizieren und

die Aufgaben anderer nachzuvollziehen. Hierzu ist es zweckmäßig, Personengruppen zu definieren, wobei jede Gruppe eine geeignete Bezeichnung erhält und ein definiertes Aufgabenspektrum übernimmt. Das Vorgehensmodell bietet einen Rahmen für ein einheitliches Begriffsverständnis, sodass die Kommunikation zwischen den verschiedenen Personengruppen vereinfacht wird. Auch den aktiven Austausch zwischen den einzelnen Teammitgliedern eines Data-Science-Projekts gilt es durch geeignete Handlungsempfehlungen zu fördern. So kann sichergestellt werden, dass Analyseverfahren aus Sicht sämtlicher beteiligten Personengruppen korrekt und zielführend angewendet werden.

Da in Data-Science-Projekten häufig eine Vielzahl von Analyseverfahren eingesetzt werden kann, sind auch die Einarbeitungszeit in neue Themenfelder sowie das Testen und Verwerfen verschiedener Analyseverfahren zu berücksichtigen. Diese Aufgaben tragen zwar unter Umständen nicht unmittelbar zum Projekterfolg bei, sind aber ein notwendiger Bestandteil des Projektablaufs. Da Data Science Auswirkungen auf ökonomische, gesellschaftliche und ökologische Dimensionen hat, sind diese Aspekte im Vorgehensmodell ebenfalls zu berücksichtigen. Das kann jedoch nur im Kontext der spezifischen Anwendungsdomäne geschehen.

3. STRUKTURIERUNG DER AUFGABEN EINES DATA-SCIENCE-PROJEKTS NACH SCHLÜSSELBEREICHEN

Aufbauend auf den in Kapitel 1 gewonnenen Erkenntnissen sowie den ebenfalls dort angesprochenen Data-Mining-Vorgehensmodellen, soll nun die Identifikation und Charakterisierung der Schlüsselbereiche eines Data-Science-Projekts im Fokus stehen.

3.1 Vorgehensmodelle aus dem Bereich der Data Science

Zusätzlich zu den bereits aufgeführten Vorgehensmodellen KDD und CRISP-DM, deren zugrundeliegende Logik eine konzeptionelle Verbundenheit zur Data Science aufweist, gilt es, in einem zweiten Schritt nach verwandten Modellen zu suchen, die speziell für den Data-Science-Bereich entwickelt wurden. Im Rahmen des Suchprozesses wurde der *Team Data Science Process* (TDSP) identifiziert, der von der Firma Microsoft entwickelt und propagiert wird. Microsoft beschreibt den TDSP als „an agile, iterative data science methodology to deliver predictive analytics solutions and intelligent applications efficiently“ (Microsoft, 2017). Der TDSP fokussiert sich auf Data-Science-Projekte im Unternehmenskontext und beinhaltet viele Aspekte, die sich auch in den anderen, älteren Prozessmodellen wiederfinden. Im Gegensatz zu Data-Mining-Modellen wie dem CRISP-DM ist der TDSP in seiner Gesamtheit jedoch eng mit Artefakten, Rollen im Team/Unternehmen sowie Arbeitsweisen und -werkzeugen verknüpft. Microsoft bietet spezielle Produkte zur Unterstützung des Prozesses an, der jedoch grundsätzlich auch mit anderen Werkzeugen realisierbar ist. Eine nähere Erläuterung des TDSP kann direkt bei Microsoft eingesehen werden.

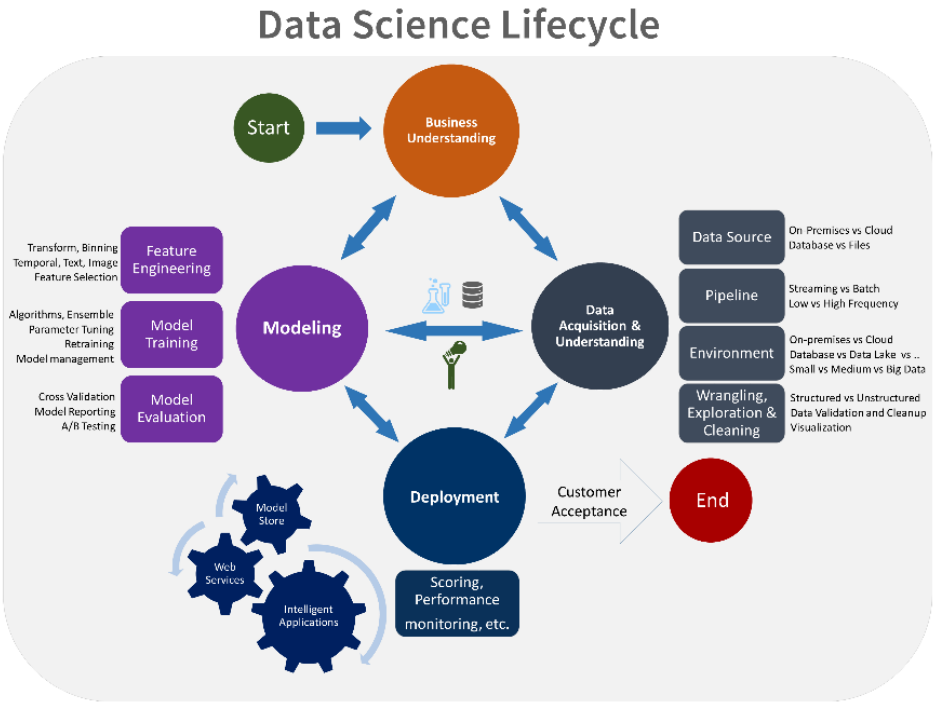


Abbildung 4: Team Data Science Process, aus: Microsoft, 2017

3.2 Schlüsselbereiche der Data Science

Unter Berücksichtigung der Stärken und Schwächen der Vorgehensmodelle KDD, CRISP-DM und TDSP sowie des in Kapitel 1 erarbeiteten Verständnisses der Data-Science-Disziplin lassen sich Schlüsselbereiche identifizieren, auf deren Basis strukturiert ein Vorgehensmodell abgeleitet werden kann.

Die Kernkompetenzen des KDD-Prozesses sind seine Einfachheit und klare Verständlichkeit. Die einzelnen Prozessschritte sind klar definiert und voneinander abgegrenzt. Der KDD-Prozess ist jedoch stark zentriert auf die Anwendung von Analyseverfahren. Vor- und nachgelagerte Aufgaben, wie der Aufbau eines Domänenverständnisses, die Nutzbarmachung von Analyseergebnissen oder die Überführung eines Modells in den Produktivbetrieb, werden nicht ausreichend berücksichtigt. Die Tatsache, dass es sich bei der Durchführung eines Data-Science- bzw. Data-Mining-Projekts um eine iterativ durchzuführende Aufgabe handelt, wird beim KDD-Prozess lediglich rudimentär berücksichtigt: Es ist weder klar, an welcher Stelle mit einer weiteren Iteration begonnen werden müsste, noch geht aus ihm hervor, wann ein Abbruch der Arbeiten empfehlenswert ist. Da die einzelnen Prozessschritte eines Analyseprojekts häufig in verschiedenen Teams, teilweise sogar von verschiedenen Abteilungen bearbeitet werden, ist ein unilateraler Informationsfluss, wie im Modell suggeriert, ebenfalls nicht gegeben.

Im Gegensatz zum KDD wird beim CRISP-DM die Iterativität der einzelnen Prozessschritte innerhalb eines Analyseprojekts sehr viel stärker deutlich. Der CRISP-DM bleibt dennoch einfach und klar verständlich. Die schwächer ausgeprägte Differenzierung der einzelnen Prozessschritte erfordert einerseits eine engere Zusammenarbeit der beteiligten Personengruppen und erlaubt andererseits keine exakte Abgrenzung der Aufgaben, wie sie im Rahmen des KDD möglich ist. Der CRISP-DM enthält neben daten-, analyse- und auswertungsbezogenen Prozessschritten mit dem *Business Understanding* einen weiteren Prozessschritt mit sehr starkem Fokus auf der Domäne. Da der CRISP-DM aus der Industrie heraus entwickelt wurde, ist ihm die Einschränkung auf den Unternehmenskontext inhärent. Auch der Prozessschritt der Nutzbarmachung stellt eine Erweiterung gegenüber dem KDD dar, obgleich dieser Schritt nur wenig ausgeprägt ist, was im Hinblick auf die heute gängigen datengetriebenen Produkte und Dienstleistungen problematisch sein kann. Hieran und auch an anderen Abschnitten des CRISP-DM-Leitfadens ist zu erkennen, dass das Vorgehensmodell in Teilen nicht mehr den aktuellen Anforderungen entspricht und neue, u. a. auch technische Herausforderungen nicht adäquat berücksichtigt.

Im Gegensatz zu KDD und CRISP-DM ist der TDSP explizit für die Verwendung in Data-Science-Projekten entwickelt worden. Einerseits werden neben den bereits in den anderen Vorgehensmodellen berücksichtigten Prozessschritten explizit Rollen und zugeordnete Aufgaben benannt, andererseits werden die Bereiche *IT-Infrastruktur* und *Customer Acceptance* ergänzt. Durch Letztere wird die Domänenrelevanz noch stärker hervorgehoben als im CRISP-DM. Nichtsdestotrotz bleibt die Ausgestaltung dieses Bereichs unklar und es findet auch hier eine Fokussierung auf den Unternehmenskontext statt. Obwohl das Modell bezogen auf die berücksichtigten Bereiche das vollständigste ist, wird der iterative Charakter von Data-Science-Projekten nicht in geeigneter Form berücksichtigt. Zudem ist der Ablauf der einzelnen Prozessschritte nicht eindeutig definiert und in der Dokumentation wird häufig – dem Ursprung des Modells mit der damit verbundenen Besetzung eines Marktsegments geschuldet – auf Microsoft-Technologien verwiesen.

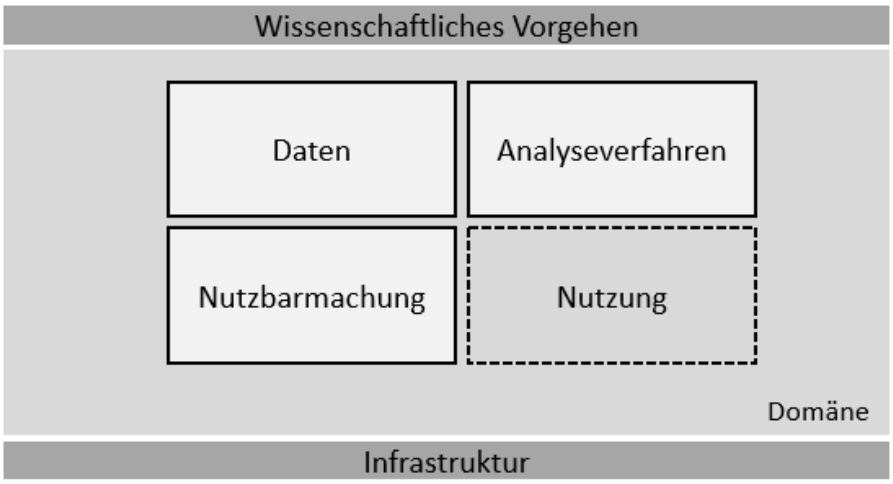


Abbildung 5: Schlüsselbereiche der Data Science

Basierend auf der kritischen Diskussion von KDD, CRISP-DM und TDSP können sieben Schlüsselbereiche der Data Science abgeleitet werden, deren grober Zusammenhang in Abbildung 5 dargestellt ist. Im Zentrum stehen die Bereiche *Daten* und *Analyseverfahren*, die im Rahmen unterschiedlicher Prozessschritte von den drei betrachteten Vorgehensmodellen adressiert werden. Die ebenfalls im Fokus stehende „Nutzbarmachung" wird im KDD angerissen, im CRISP-DM explizit adressiert und im TDSP durch den Bereich der *Customer Acceptance* besonders hervorgehoben. Die *Nutzung* des durch die Nutzbarmachung entstehenden Analyseartefaktes wird jedoch in keinem der zuvor genannten Vorgehensmodelle betrachtet. Große Teile dieses Schlüsselbereichs sind zwar auch nicht als Kernelement eines Data-Science-Projekts anzusehen, doch entstehen häufig Artefakte, die entweder im Laufe der Nutzung durch Data Scientists angepasst werden müssen, oder solche, die in zukünftige (Weiter-)Entwicklungsprojekte einfließen.

Die vier zuvor genannten Schlüsselbereiche sind eingebettet in den Schlüsselbereich *Domäne*. Im Gegensatz zur Spezifikation im CRISP-DM und TDSP sollte dieser Bereich, unter Berücksichtigung der Definition des Data-Science-Begriffes, nicht explizit auf den Unternehmenskontext beschränkt sein, sondern in jedem der zuvor genannten Schlüsselbereiche berücksichtigt werden. Zudem werden die vier oben genannten Schlüsselbereiche durch den übergreifenden Schlüsselbereich *Wissenschaftliches Vorgehen* flankiert, der bisher in keinem der bestehenden Vorgehensmodelle explizite Berücksichtigung findet, jedoch einen Kernbestandteil der Data-Science-Disziplin bildet. Der begleitende Schlüsselbereich *IT-Infrastruktur*, der bis dato nur im TDSP adressiert wird, ist ebenfalls zu berücksichtigen, da er in vielen Data-Science-Projekten eine zunehmend wichtige Rolle spielt (siehe auch Kapitel 1). Die folgenden Abschnitte enthalten eine kurze Charakterisierung der sieben Schlüsselbereiche. Eine detaillierte Beschreibung erfolgt in späteren Kapiteln.

Schlüsselbereich „Daten"

Daten werden als der „Rohstoff" der Data Science betrachtet (Palmer, 2006). Mit Daten sind unmittelbar zahlreiche Arbeitsschritte verbunden, die zusammengenommen häufig den Aufwandsschwerpunkt eines Data-Science-Projekts bilden. Zu diesen Arbeitsschritten gehören die Datenbeschaffung, -integration, -bereinigung, -transformation und -speicherung. Es muss geklärt werden, ob die zur Erfüllung des Projektziels notwendigen Daten in ausreichender Menge und Qualität zur Verfügung stehen, ob sie genutzt werden dürfen (Datenschutz) und welche Struktur sie besitzen. Mit den Daten verbunden ist aber noch ein weiterer wichtiger

Bereich: der Aufbau eines gemeinsamen Datenverständnisses im Kontext des Anwendungsproblems. Ein wichtiger Arbeitsschritt ist dabei die explorative Datenanalyse, gegebenenfalls inklusive einer ersten Datenvisualisierung. Die Vorbereitung der Daten im Hinblick auf das später genutzte Analyseverfahren ist eine weitere wichtige Aufgabe, da je nach verwendetem Verfahren spezielle Anforderungen an die Form der Daten gestellt werden.

Schlüsselbereich „Analyseverfahren“

Während die Daten den Rohstoff der Data Science bilden, wird die Anwendung eines geeigneten Datenanalyseverfahrens als zentraler Schritt im Data-Science-Prozess angesehen, da sie (im Erfolgsfall) den Grundstein für den angestrebten Erkenntnisgewinn liefert. Dieser Schritt wird häufig auch Modellierung oder Modellbildung genannt, da durch Anwendung von Analyseverfahren auf Daten ein Modell des untersuchten Wirklichkeitsbereichs entsteht, das anschließend, z. B. zur Klassifikation neuer Daten oder zur Prognose zukünftiger Werte, verwendet werden kann. Wichtig bei der Modellbildung ist die Auswahl eines für den Anwendungsfall und die gegebenen Daten passenden Analyseverfahrens und dessen geeignete Parametrisierung. Die Bandbreite der zur Verfügung stehenden Verfahren ist sehr hoch und reicht von statistischen Methoden über klassisches Data Mining bis zu neuronalen Netzen, Deep Learning und Künstlicher Intelligenz. Stehen keine passenden Analyseverfahren zur Verfügung, so müssen bestehende Verfahren angepasst oder sogar neue Verfahren entwickelt werden. Vorbereitende Schritte der Modellierung sind die verfahrensspezifische Datenaufbereitung und die Merkmalskonstruktion, Folgeschritt ist die Modellevaluierung.

Schlüsselbereich „Nutzbarmachung“

Die Nutzbarmachung rechtfertigt den entstehenden zeitlichen und finanziellen Aufwand. Eine unzureichende spätere Nutzung der Ergebnisse kann daher sogar ein theoretisch erfolgreiches Projekt praktisch scheitern lassen. Die einfachste, aber auch unbestimmteste Form der Nutzbarmachung ist die Aufbereitung der Ergebnisse in Form eines Abschlussberichts oder einer Veröffentlichung. Wenn sachdienlich, sollte es das Ziel sein, das entwickelte Analysemodell in eine dauerhaft nutzbare Form zu überführen. Dies kann je nach Anwendungsfall und Nutzergruppe z. B. durch ein Software-System erreicht werden. Dabei sind zahlreiche sowohl fachliche als auch technische Aspekte zu berücksichtigen, etwa die Form der Aufbereitung der Analyseergebnisse oder die Bereitstellung einer angemessenen Benutzerschnittstelle.

Schlüsselbereich „Nutzung“

Unstrittig scheint zu sein, dass der rein operative Betrieb, also die Anwendung der entwickelten Analyseartefakte, nicht als Teil von Data-Science-Projekten zu sehen ist. Dies gilt aber nicht für das Monitoring der Analysequalität während der Nutzung mit dem Ziel, entweder die Verwendung der Modelle zu beenden oder deren Anpassungsbedarf zu identifizieren (sogenanntes „Model-Lifecycle-Management“).

Schlüsselbereich „Domäne“

Die vier zuvor charakterisierten Schlüsselbereiche können nur im Kontext der Domäne, also des analysierten Anwendungsfeldes, konkretisiert werden. Ein breites Hintergrundwissen auf diesem Anwendungsfeld ist an vielen Stellen des Data-Science-Prozesses relevant, so bei der Identifikation eines lohnenden Analyseziels, dem korrekten Verständnis von Daten, ihrer Herkunft, Qualität und Zusammenhänge, der Bewertung und Einordnung der erzielten Analyseergebnisse im Kontext der Anwendung sowie der späteren praktischen Nutzung der Ergebnisse.

Auch die Beurteilung von Stärken und Schwächen bestehender Lösungen, die fachliche Anforderungsanalyse, die Unterstützung bei der Modellparametrisierung und die abschließende Evaluation des Projekterfolgs werden diesem Bereich zugeordnet. Schließlich lassen sich auch die rechtlichen, gesellschaftlichen und ethischen Aspekte des Data-Science-Projekts an dieser Stelle aufnehmen.

Schlüsselbereich „Wissenschaftliches Vorgehen“

Ein Data-Science-Projekt sollte einem bewährten Vorgehensmodell folgen und auf Grundlage des aktuellen wissenschaftlichen Erkenntnisstandes durchgeführt werden. Wichtige Aspekte sind dabei zum einen die Standardisierung und Strukturierung des Vorgehens, ein geeignetes Projektmanagement sowie die Kommunikation der beteiligten Anspruchsgruppen. Hier sind weniger spezifische technische oder domänenspezifische Kenntnisse wichtig, sondern organisatorische, kommunikative und vermittelnde Kompetenzen. Zum anderen gilt es, eine angemessene wissenschaftliche Arbeitsweise sicherzustellen, etwa das Aufstellen einer Forschungshypothese, die Evaluation angewandter Methoden im Hinblick auf Angemessenheit und Effizienz, die Überprüfung der Validität der Ergebnisse, die Sicherstellung ihrer Reproduzierbarkeit und das fundierte, fakten- und datenbasierte Treffen von Entscheidungen. Das Festhalten von neuem, verallgemeinerbarem (also nicht projektspezifischem) Wissen über Datensätze und Methoden wird ebenso dazugezählt wie die Veröffentlichung der generalisierbaren Erkenntnisse.

Schlüsselbereich „IT-Infrastruktur“

Praktisch alle Aufgaben eines Data-Science-Projekts, sei es das Datenmanagement, die eigentliche Datenanalyse, die Evaluation und Nutzbarmachung der Analyseergebnisse oder auch das Projektmanagement, werden mit Hilfe von spezialisierten Software-Produkten umgesetzt. Die Bandbreite und Komplexität dieser Produkte sind insbesondere bei größeren Projekten, die mit umfangreichen Datenmengen arbeiten und aufwendige Analyseverfahren nutzen, hoch. Die Bereitstellung und der Betrieb der erforderlichen IT-Infrastruktur ist eine anspruchsvolle Aufgabe, die entsprechend spezielle IT-Kenntnisse (z. B. in Bezug auf Arbeit mit verteilten Daten, Cloud-Anbindung, Sandboxing, skalierbare Architekturen, verteiltes Berechnen von Modellen, Automatisierung) erfordert.

4. DER DATA SCIENTIST

Artikel wie der von Davenport und Patil (2012), in dem der Beruf des Data Scientists mit *The Sexiest Job of the 21 Century* betitelt wird, können den Anschein erwecken, dass alle auf diesem Gebiet nötigen Kompetenzen in einer einzelnen Person vereinigt sein können beziehungsweise müssen. Diese Sichtweise wurde in vielen Publikationen übernommen[1], ist aber problematisch.

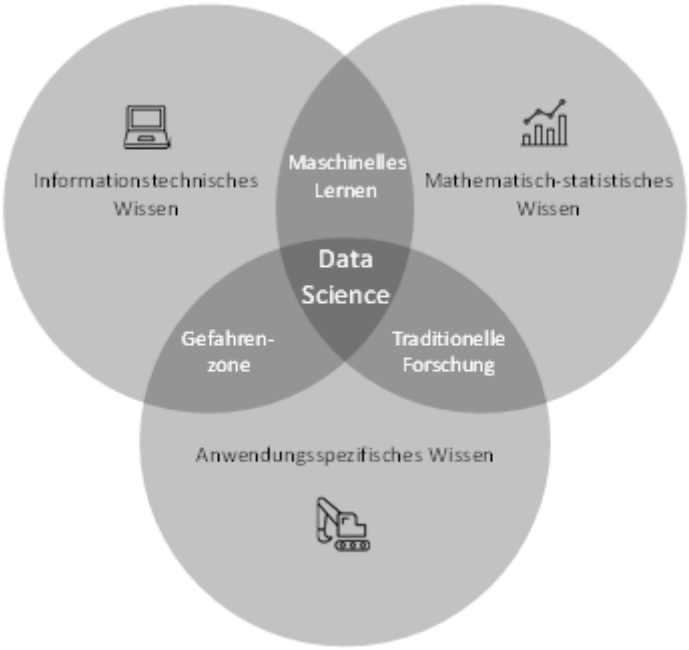


Abbildung 6: Nötige Kompetenzen eines Data Scientists, in Anlehnung an: Conway, 2010

In verschiedenen Quellen werden aufbauend auf dem Beitrag von Conway (2010) von einem Data Scientist Kompetenzen in **mathematisch-statistischen**, **informationstechnischen** und **anwendungsspezifischen** Bereichen gefordert. Sind nur Kompetenzen in einzelnen Bereichen vorhanden, handelt es sich demnach nicht um einen ausgebildeten Data Scientist. Laut Dhar (2013) genügen grundlegende Fähigkeiten in den o. g. Bereichen, auch das häufig zitierte Diagramm von Conway (2010) – siehe Abbildung 6 – legt dies durch die geringen Überschneidungen der einzelnen Fachgebiete nahe. Eine oberflächliche Kenntnis ist jedoch nicht immer ausreichend: Abhängig vom konkreten Anwendungsfall können vertiefte Kompetenzen in einem oder mehreren der genannten drei Bereiche nötig sein. Weiterhin müssen Data Scientists in der Lage sein, mit allen Anspruchsgruppen in einer geeigneten Sprache zu **kommunizieren** (Davenport & Patil, 2012), das **Management** eines Data-Science-Projekts zu übernehmen und die **strategische Einordnung** von Aktivitäten vorzunehmen. Abbildung 7 fasst sämtliche Kompetenzen zusammen, die für die Durchführung von Data-Science-Projekten benötigt werden. Für eine einzelne Person ist es in der Regel nicht möglich, tiefreichende Fähigkeiten in allen genannten Bereichen aufzubauen (Zschech et al., 2018). Data Scientists können sich daher entweder in einer Disziplin oder in wenigen Disziplinen spezialisieren oder übergeordnete bzw. weniger datenorientierte Rollen übernehmen.

[1] Eine Übersicht über existierende Data-Scientist-Definitionen ist in Chatfield et al. (2014) zu finden.

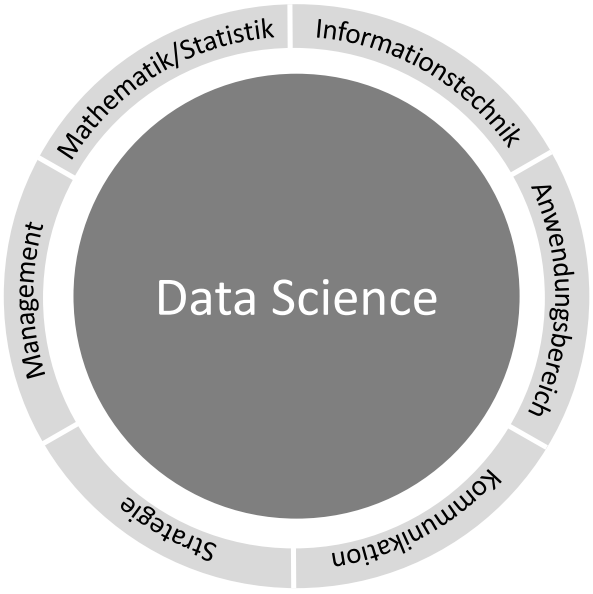


Abbildung 7: Nötige Kompetenzen in einem Data-Science-Projekt

Um der sich etablierenden Spezialisierung innerhalb von Data-Science-Projekten Rechnung zu tragen, werden vermehrt verschiedene Rollen unterschieden. Im Folgenden werden die Rollen dargestellt, die von den Teilnehmerinnen und Teilnehmern der Arbeitsgruppe als relevant identifiziert wurden, um sämtliche notwendigen Aktivitäten eines Data-Science-Projekts abzudecken. Um die Übersicht zu wahren, wurde allerdings auf eine sehr feingliedrige Rollenunterteilung verzichtet. Bei großen Projekten werden die hier beschriebenen Rollen häufig noch in spezifischere Unterrollen aufgeteilt. Entsprechende Hinweise finden sich bei den nachfolgenden Beschreibungen. Eine Rolle entspricht nicht zwangsläufig einer Person. Sie kann entweder auf mehrere Personen innerhalb eines Projekts aufgeteilt werden oder es können, gerade bei kleinen Projekten, mehrere Rollen von einer Person übernommen werden. In Abbildung 8 sind die Rollen innerhalb eines Data-Science-Projekts zusammengefasst.

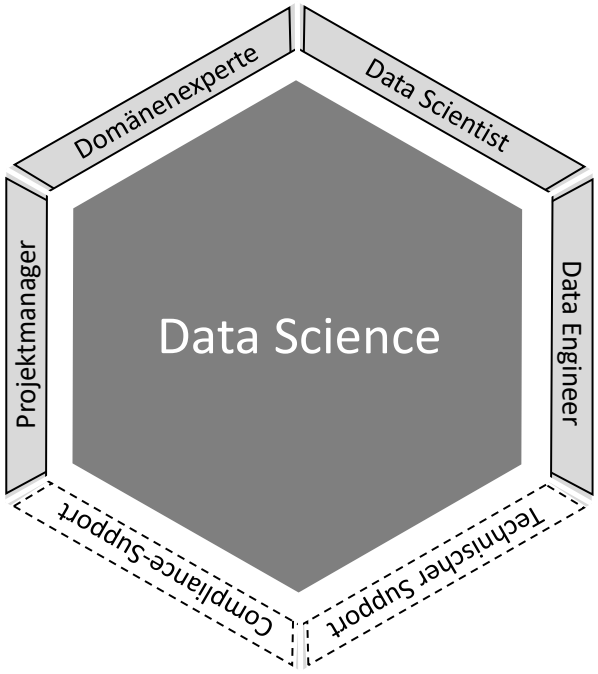


Abbildung 8: Rollen in einem Data-Science-Projekt

Die Kernrolle „Data Scientist“

Der Begriff *Data Scientist* wird in der Praxis auf zwei unterschiedliche Arten genutzt: zum einen als Oberbegriff für alle in einem Data-Science-Projekt tätigen Personen, zum anderen in einem spezifischeren Sinne für diejenigen Personen, die sich auf die tatsächliche Datenanalyse spezialisieren. Im Sinne des Oberbegriffs ist ein Data Scientist zuständig für die Durchführung aller Aspekte eines Data-Science-Projekts. Er arbeitet mit Domänenexperten zusammen, ist aber für alle methodischen, technischen und organisatorischen Fragen verantwortlich. Obwohl dieses Verständnis von Data Scientist in der betrieblichen Praxis noch häufig angetroffen wird, kann eine einzelne Person – wie oben bereits erläutert – eine so umfassend definierte Rolle höchstens bei kleinen Data-Science-Projekten übernehmen. Im spezifischeren Sinne versteht man unter einem Data Scientist einen Spezialisten für den Analysebereich eines Data-Science-Projekts, der insbesondere für die Auswahl der Analysemethoden und -werkzeuge, die Durchführung von Analysen und die Interpretation der Ergebnisse zuständig ist, in den übrigen Bereichen eines Data-Science-Projekts jedoch nur beratend tätig ist. Weitere Aufgaben müssen dementsprechend von anderen Rollen (siehe unten) übernommen werden. In diesem Dokument wird im Weiteren unter *Data Scientist* immer die spezifischer definierte Rolle verstanden.

Die Rolle eines Data Scientists kann bei größeren und komplexeren Data-Science-Projekten noch in weitere Unterrollen aufgespalten werden:

- *Data Analyst*: Ein *Data Analyst* ist eine in Stellenanzeigen weiterhin häufig anzutreffende Bezeichnung für Personen, die sich mit unterschiedlichen Aspekten der Datenaufbereitung, -analyse und -auswertung befassen. Sie nutzen datengetriebene Analyseverfahren, statistische Modelle und Methoden der Datenvisualisierung. Inhaltlich gibt es also sehr große Überschneidungen mit den Aufgaben eines Data Scientists, sodass *Data Analyst* häufig als älteres Synonym zu *Data Scientist* betrachtet wird. Wenn *Data Analyst* tatsächlich als Unterrolle eines Data Scientists verstanden werden soll, so erfolgt die Abgrenzung meist durch seine Schwerpunktsetzung: Der Data Analyst ist verstärkt in der explorativen Datenanalyse tätig und bedient sich stärker traditioneller Analysemethoden, der Data Scientist fokussiert eher auf das Analysemodell und verwendet auch komplexe, neue Methoden.
- *Methodenspezialist*: Methodenspezialisten beschäftigten sich mit der Erforschung und Weiterentwicklung von Data-Science-Methoden (Datenanalyse, Datentransformation etc.). Sie entwerfen beispielsweise neue Analysealgorithmen und führen Untersuchungen zur Wirkung von Analyseparametern durch. Außerdem sind sie über den aktuellen Forschungsstand im Data-Science-Bereich informiert. Obwohl Methodenspezialisten auch im Kontext anwendungsbezogener Data-Science-Projekte einen Beitrag leisten können, ist ihr Fokus stärker theorie- bzw. forschungsbezogen.
- *Data Scientist Consultant*: Data Scientist Consultants besitzen genügend methodisches, technisches und gegebenenfalls auch Domänenwissen, um bei der Definition geeigneter Analysefragestellungen bzw. -anwendungsfälle beraten zu können. Es handelt sich idealtypisch um erfahrene Data Scientists, die ihr Know-how Unternehmen, Organisationen oder Organisationseinheiten zur Verfügung stellen, in denen Data-Science-Projekte durchgeführt werden sollen, aber nicht genügend Expertise vorhanden ist.

Die Rolle eines Data Scientists kann bei Bedarf ferner auf Grund des Erfahrungshintergrunds (Junior Data Scientist, Senior Data Scientist, Advanced Data Scientist etc.) oder der Ausbildung (Absolventen spezieller Studiengänge, Absolventen von anerkannten Weiterbildungen/Zertifizierungen, Quereinsteiger/Praktiker) unterteilt werden.

Die Rolle „Data Engineer“

Data Engineers kümmern sich um die Beschaffung, Speicherung, Aufbereitung, Strukturierung und Weitergabe von Daten. Sie sind insbesondere in den Vorstufen der eigentlichen Analyse tätig. Sie haben einen technischeren Fokus als Data Scientists und befassen sich auch mit der für das Data-Science-Projekt benötigten IT-Infrastruktur. Gelegentlich wird für diese Rolle auch der Begriff *Data Architect* verwendet.

Eine Unterrolle des Data Engineers, die insbesondere bei größeren Data-Science-Projekten häufig separat besetzt wird, ist die des *Data Stewards* (auch *Data Manager* oder *Data Quality Engineer*). Dieser kümmert sich fortwährend um den Zugang zu den Daten und ihren Schutz sowie um die dauerhafte Gewährleistung einer hohen Datenqualität. Ein Data Steward hat somit starke Berührungspunkte zum fachlichen Anwendungsbereich.

Die Rolle „Domänenexperte“

Domänenexperten sind Fachanwender oder Vertreter der Fachanwender. Sie verfügen über spezifisches Wissen in Bezug auf die Anwendungsdomäne und besitzen ein inhaltliches Verständnis der Problemstellung / des Anwendungsfalls. Domänenexperten können Prioritäten für zu modellierende/analysierende Aspekte setzen und sind Bindeglieder zu den methodischen und technischen Experten.

Innerhalb der Domänenexperten kann es wieder Unterrollen geben. Häufig anzutreffen sind etwa die *Business Developer*, welche die zugrundeliegende domänenspezifische Fragestellung entwickeln und somit das Bindeglied zwischen Unternehmenszielen und Datenanalysen bilden, oder die *Business Analysts*, die später die entwickelten Analysemodelle im Rahmen ihrer fachlichen Aufgaben nutzen.

Die Rolle „Projektmanager“

Der Projektmanager plant, steuert und koordiniert den Gesamtablauf eines Data-Science-Projekts. Dazu benötigt er – neben den traditionellen Projektmanagementfertigkeiten – ein gutes Verständnis der methodischen und technischen Aspekte der Data Science, Kenntnisse geeigneter Vorgehensmodelle und einen Einblick in die Anwendungsdomäne. Insbesondere bei kleineren Projekten wird das Projektmanagement häufig von Personen übernommen, die auch die Rolle eines Data Scientists oder eines Data Engineers ausfüllen. Das Projektmanagement kann aber auch von Personen ohne spezifisches Data-Science-Knowhow übernommen werden, wenn ihnen entsprechende Experten zur Seite stehen. Solche – auch *Methodical Lead* oder *Technical Lead* genannten – Experten besitzen ein tiefergehendes Hintergrundwissen, um das Projekt methodisch und technisch zu begleiten. Zusammen mit Domänenexperten bestimmen sie den Scope der Analyse und Umsetzung.

Neben den vier oben beschriebenen Kernrollen eines Data-Science-Projekts, die einen starken inhaltlichen Bezug zu den Projektzielen besitzen, sind noch zwei Unterstützungsrollen relevant. Personen in Unterstützungsrollen sind für die erfolgreiche Durchführung des Data-Science-Projekts zwar erforderlich, die Projektergebnisse haben für ihre Arbeit jedoch nur mittelbar Bedeutung. Diese Personen tragen somit im Rahmen ihrer normalen Tätigkeit zum Gelingen des Projekts bei, ohne direkt von seinen Data-Science-spezifischen Aspekten betroffen zu sein.

Die Rolle „Technischer Support“

Der technische Support umfasst alle Aufgaben, die erledigt werden müssen, um die technischen Voraussetzungen für die Durchführung des Data-Science-Projekts zu schaffen. Typische Unterrollen des technischen Supports sind etwa der *IT Infrastructure Architect*, der eine geeignete IT-Infrastruktur für das Projekt entwirft, und der *IT-Techniker/IT-Administrator*, welcher die benötigte Hard- und Software bereitstellt sowie die zugrundeliegenden Systeme konfiguriert. Aber auch Anwendungsentwickler, die sich mit der Implementierung von Anwendungssoftware/-werkzeugen zur produktiven Nutzung der Analyseergebnisse befassen, werden hier dem technischen Support zugeordnet.

Die Rolle „Compliance-Support“

Der Compliance-Support ist für die Einhaltung gesetzlicher Vorgaben, die Kompatibilität des Data-Science-Projekts mit den organisationsinternen Regelwerken und das korrekte Verhalten der Projektmitarbeiter verantwortlich. Er ist außerdem für das übergreifende Sicherheitsmanagement zuständig und gewährleistet den Datenschutz, insbesondere den Schutz personenbezogener Daten.

5. DATA-SCIENCE-VORGEHENSMODELL DASC-PM

Aufbauend auf den zuvor, vor allem in Kapitel 3.2, beschriebenen Ausarbeitungen soll in diesem Kapitel das Data-Science-Process-Model (DASC-PM) als Vorgehensmodell für Data-Science-Projekte eingeführt werden – siehe Abbildung 9.

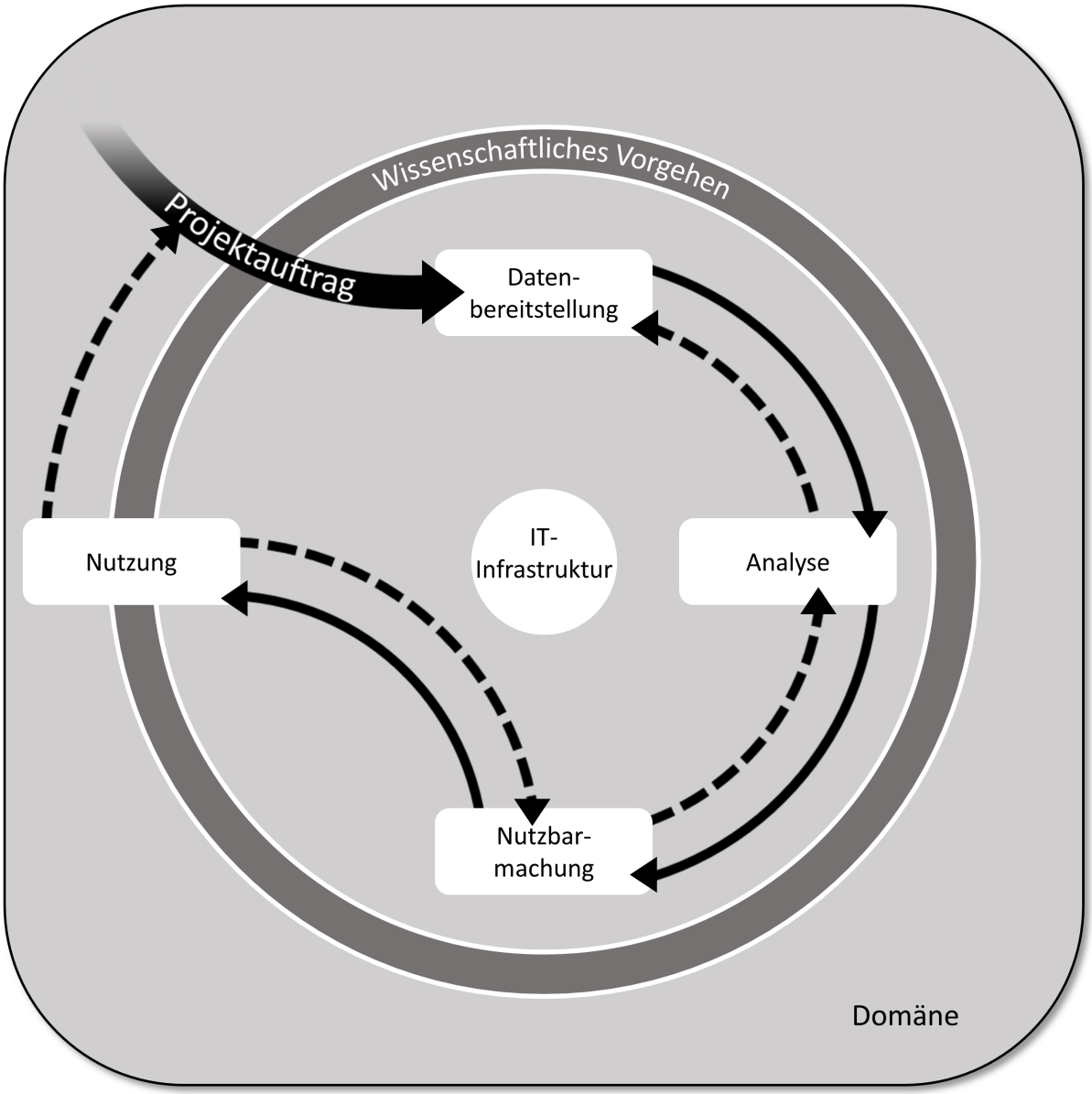


Abbildung 9: Data-Science-Vorgehensmodell DASC-PM

Die Visualisierung des DASC-PM leitet sich aus den zuvor identifizierten Schlüsselbereichen eines Data-Science-Projekts und auch aus den bereits angedeuteten Abhängigkeiten zwischen ihnen ab. Dabei soll die Abbildung das Vorgehen innerhalb eines Projekts vereinfacht und prägnant darstellen. Details sind den Kapiteln zu entnehmen, auf die in den folgenden Beschreibungen verwiesen wird. Data-Science-Projekte können von sehr unterschiedlicher Größe, Komplexität und Schwerpunktsetzung sein. Nicht alle adressierten Aufgaben sind im individuellen Projekt daher in gleichem Maße zu berücksichtigen. Unabhängig von der Ausgestaltung im konkreten Anwendungsfall sollte jedoch für alle im Folgenden dargestellten Phasen geklärt werden, ob die jeweils enthaltenen Aufgaben berücksichtigt werden müssen.

Die in Abbildung 9 dargestellten durchgezogenen Pfeile bilden den primären Pfad bei der Verwendung des DASC-PM. Die gestrichelten Pfeile zeigen die Möglichkeit von Rückkopplungen zu vorherigen Phasen, die durch die Gewinnung neuer Erkenntnisse im Projektverlauf immer wieder nötig sein können.

Eingebettet ist das Analysevorhaben in eine **Domäne**. Innerhalb dieser werden Anwendungsfälle identifiziert, die eine datenwissenschaftliche Betrachtung rechtfertigen. Aus einem oder mehreren Anwendungsfällen wird ein Projektauftrag formuliert, der in Form eines Data-Science-Projekts zu bearbeiten ist (für eine detaillierte Beschreibung vgl. Kapitel 10). Können in dieser Phase aus der Domäne heraus explizite Aufgaben formuliert werden, stellen diese in anderen Phasen häufig domänenspezifische Rahmenbedingungen dar, die Aufgaben beeinflussen. Die Domäne muss daher durchgängig berücksichtigt werden.

Der definierte Projektauftrag ist in jeder einzelnen Projektphase dem **wissenschaftlichen Vorgehen** entsprechend zu bearbeiten. Hervorzuheben sind hier vor allem das Projektmanagement und eine strukturierte Bearbeitung, die bereits durch die Verwendung eines Vorgehensmodells in den Vordergrund gestellt wird. Details zum nötigen Grad an Wissenschaftlichkeit sind unter Berücksichtigung der Projektgegebenheiten und der Domänenspezifika festzulegen. Bereiche, in denen das wissenschaftliche Vorgehen häufig als besonders relevant angesehen wird, finden sich in Kapitel 12.

Sämtliche Aktivitäten, die dem Schlüsselbereich Daten zuzuordnen sind, werden innerhalb der Visualisierung des DASC-PM in der Phase der **Datenbereitstellung** zusammengefasst, weshalb der verwendete Begriff hier weit zu fassen ist (vgl. Kapitel 6). Er enthält sowohl die Datenaufbereitung (bestehend aus Datenbeschaffung, -integration, -transformation und -speicherung) und das Datenmanagement als auch die explorative Analyse zur Erkundung der Daten. Als Ergebnis der Bearbeitung dieser Phase entsteht eine Datenquelle, die aus methodischer und fachlicher Sicht für die Analyse geeignet ist.

Die Phase der **Analyse** (Kapitel 7) beinhaltet alle Aufgaben, die dem Schlüsselbereich Analyseverfahren zuzuordnen sind. Entweder können in einem Data-Science-Projekt bestehende Verfahren angewendet werden oder es müssen zunächst neue Verfahren entwickelt werden. Die Identifikation geeigneter bestehender Verfahren kann dabei eine große Herausforderung darstellen. Artefakt dieser Phase ist ein Analyseergebnis, das eine methodische und fachliche Evaluation durchlaufen hat.

In der Phase der **Nutzbarmachung**, beschrieben in Kapitel 8, müssen die Analyseergebnisse so aufbereitet werden, dass sie für die geplante Nutzung geeignet sind. Die Ergebnisse könnenabhängig von dem spezifischen Projekt sehr unterschiedlich sein: Analyseartefakte können aus Ergebnissen bestehen, die den Adressaten verbal oder in Form von Berichten zur Verfügung gestellt werden. Weiterhin können Modelle oder auch Analyseverfahren selbst das Ergebnis eines Data-Science-Projekts sein.

Die sich an die Projektdurchführung anschließende **Nutzung** von Analyseartefakten (vgl. Kapitel 9) ist nicht als primärer Teil eines Data-Science-Projekts anzusehen. Ein Monitoring der Verwendung ist aber abhängig von der konkreten Form der Nutzbarmachung nötig, um die fortbestehende Eignung des Modells in der Anwendung zu prüfen und ggf. Erkenntnisse aus der Nutzung für die Weiter- und Neuentwicklung von Analyseartefakten zu erlangen.

Sämtliche Schritte, die ein Data-Science-Projekt durchlaufen muss, sind von der zu Grunde liegenden **IT-Infrastruktur** abhängig, das tatsächliche Ausmaß ist allerdings projektindividuell zu bewerten (vgl. Kapitel 11). Auch wenn die Nutzung spezifischer Hard- und Software häufig bereits organisationsintern festgelegt ist, sollte man, wenn auch nicht die Auswahl, so doch

zumindest die limitierenden und befähigenden Merkmale der IT-Infrastruktur in sämtlichen Projektphasen berücksichtigen.

Obwohl aus Gründen der Übersichtlichkeit auf eine Visualisierung dieser Tatsache in Abbildung 9 verzichtet wurde, ist der Abbruch des Data-Science-Projekts in jeder einzelnen Projektphase als Option zu berücksichtigen. Auch wenn dadurch das im Projektauftrag definierte Ziel i. d. R. nicht erreicht werden kann, bedeutet dies nicht zwangsläufig, dass das Projekt vollständig fehlgeschlagen ist. Erkenntnisse, die bis zum Zeitpunkt des Abbruchs gesammelt wurden, können innerhalb des bearbeiteten Anwendungsfalls bzw. der bearbeiteten Anwendungsfälle Verwendung finden.

In den folgenden Kapiteln werden die einzelnen Schlüsselbereiche des DASC-PM detailliert betrachtet. Die Darstellungen basieren dabei auf der Expertise der Teilnehmerinnen und Teilnehmer der Arbeitsgruppe. Ein Anspruch auf Vollständigkeit besteht nicht, vielmehr soll die Ausarbeitung dazu dienen, dem Leser ein Gefühl für relevante Aspekte innerhalb der verschiedenen Projektschritte zu vermitteln.

Die das Vorgehensmodell detaillierenden Kapitel sind alle nach derselben Struktur aufgebaut. Zunächst werden die Teilbereiche dargestellt, die im jeweiligen Schlüsselbereich eine Rolle spielen. Unterschieden werden dabei merkmalstragende und aufgabentragende Bereiche. Letztere werden weiterhin unterschieden in Kernaufgaben des Teilprozesses und begleitende Aufgaben. Die verbindenden Kanten beinhalten ebenfalls Aufgaben, die durch die Verzahnung der einzelnen Bereiche entstehen. Das Schema einer solchen Darstellung findet sich in Abbildung 10.

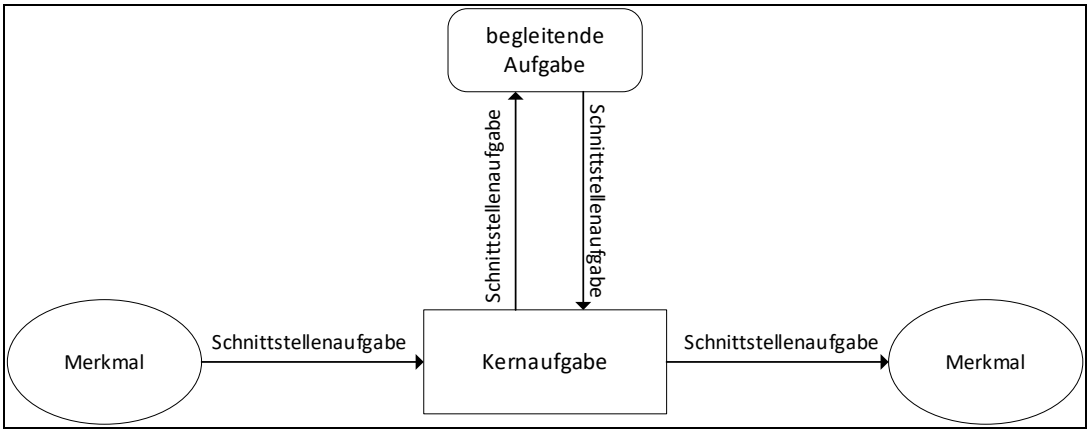


Abbildung 10: Verwendete Notation zur Detaillierung von Schlüsselbereichen

Merkmale, Kernaufgaben und begleitende Aufgaben werden anschließend in jeweils einem eigenen Unterkapitel detailliert. In den die Aufgaben beschreibenden Kapiteln sind jeweils zwei Darstellungen zu finden, die auf den in Kapitel 4 hergeleiteten Abbildungen 7 und 8 basieren. In der ersten wird in Form eines Netzdiagramms das Kompetenzprofil von Personen dargestellt, die sich im jeweiligen Aufgabenbereich spezialisieren. Die Darstellungen sind dabei durch eine Aggregation der Rückmeldungen von Teilnehmerinnen und Teilnehmern entstanden. Die Diagramme sollen dem Leser eine ungefähre Einschätzung des Kompetenzprofils für ein typisches Data-Science-Projekt ermöglichen, die Ausprägungen sind im Detail sicherlich diskutabel und können sich im individuellen Projekt stark unterscheiden. In der jeweils zweiten Abbildung werden nach demselben Prinzip die am spezifischen Aufgabenbereich beteiligten Rollen dargestellt. Sämtliche Netzdiagramme eines Schlüsselbereichs werden, weiterhin unterteilt nach Kompetenzprofil- und Rollendarstellung, in der Einführung zu jedem Schlüsselbereich sich überlagernd abgebildet, sodass sowohl ein Kernbereich als auch ein Maximalbereich ersichtlich ist.

6. SCHLÜSSELBEREICH DATEN

Der Schlüsselbereich *Daten* wird in die in Abbildung 11 dargestellten und in Verbindung stehenden Teilbereiche untergliedert, die in den folgenden Kapiteln einzeln betrachtet werden.

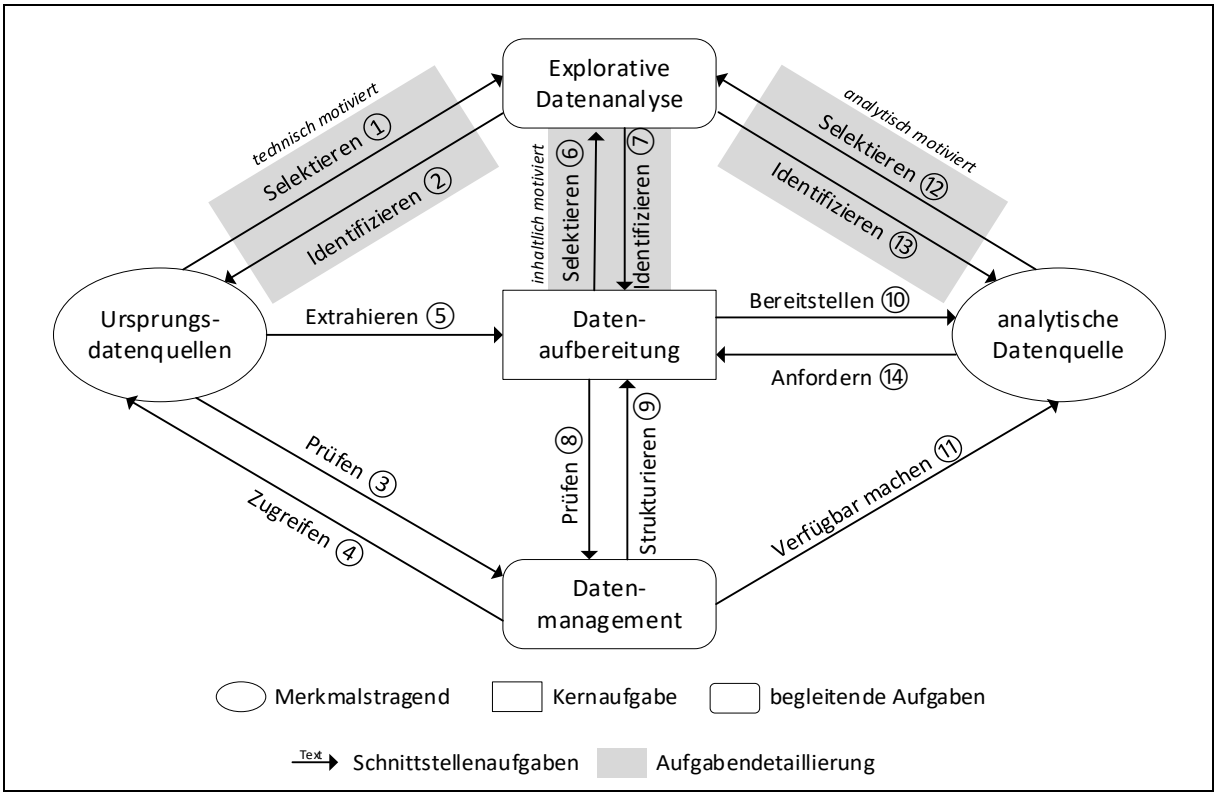


Abbildung 11: Schlüsselbereich Daten (Beschreibung siehe Folgeseite)

①	Die Daten werden direkt aus den Quellen selektiert, um ihre Eigenschaften zu untersuchen. Dabei ist häufig bereits grob bekannt, welche Daten für die Durchführung des Data-Science-Projekts benötigt werden. Es ist aber auch denkbar, dass eine Prüfung der existierenden Datenquellen zur Generierung neuer Ideen führt.
②	Für die spätere Analyse der Daten und deren grundlegender Eigenschaften werden die relevanten Daten identifiziert.
③	Auf Basis von Eigenschaften und Funktionen der Ursprungsdatenquellen werden Strategien für das Datenmanagement (z. B. bzgl. Detailgrad, Vorberechnungen etc.) geprüft und festgelegt.
④	Ein korrekter und für die Fragestellung des Data-Science-Projekts geeigneter Zugriff auf die Datenquellen wird sichergestellt.
⑤	Auf Basis der in ② gewonnenen Erkenntnisse und des in ④ festgelegten Datenzugriffs werden die potenziell relevanten Daten aus den Quellen extrahiert.
⑥	Daten werden zur Analyse selektiert, um Eigenschaften zu untersuchen, Aufbereitungsstrategien festzulegen oder Aufbereitungsergebnisse zu validieren.
⑦	Als Ergebnis der explorativen Datenanalyse werden sinnvolle Aufbereitungsstrategien identifiziert.
⑧	Anforderungen an das Datenmanagement aus dem aktuellen Data-Science-Projekt werden geprüft.
⑨	Anforderungen an die Datenstrukturierung bzgl. der Datenspeicherungsarchitektur aus dem aktuellen Data-Science-Projekt und der bestehenden IT-Infrastruktur werden berücksichtigt.
⑩	Die aufbereiteten Daten werden für die Anwendung von Analyseverfahren zur Verfügung gestellt.
⑪	Die Daten werden in einem geeigneten Datenmodell gespeichert, über eine Virtualisierungsschicht oder auch als Stream verfügbar gemacht, zudem ggf. geschützt und archiviert.
⑫	Die Daten werden selektiert, um sie bezogen auf ein explizites Analysevorhaben zu bewerten.
⑬	Für ein explizites Analysevorhaben relevante Eigenschaften werden identifiziert.
⑭	Auf Basis der in ⑫ identifizierten Erkenntnisse wird die Datenaufbereitung angepasst.

In Abbildung 12 ist das Kompetenzprofil von Personen dargestellt, die sich im Schlüsselbereich *Daten* spezialisieren. Abbildung 13 stellt diejenigen Rollen dar, die bei der Betrachtung dieses Schlüsselbereichs eine Relevanz besitzen.

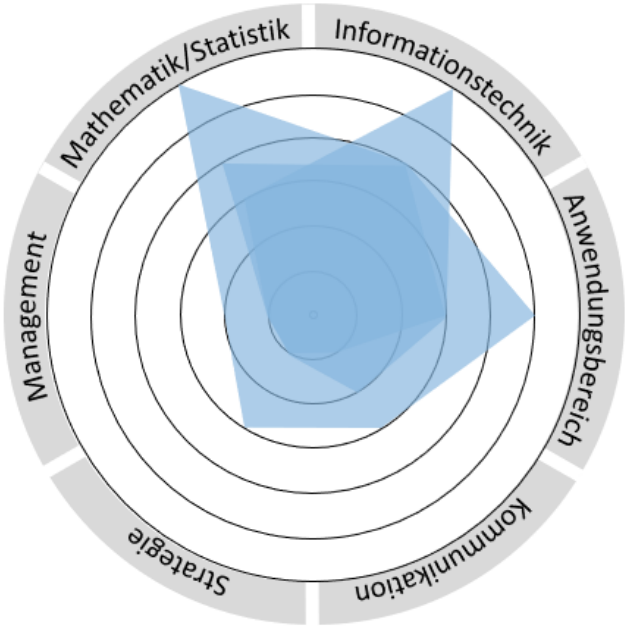


Abbildung 12: Kompetenzprofil des Schlüsselbereichs „Daten“

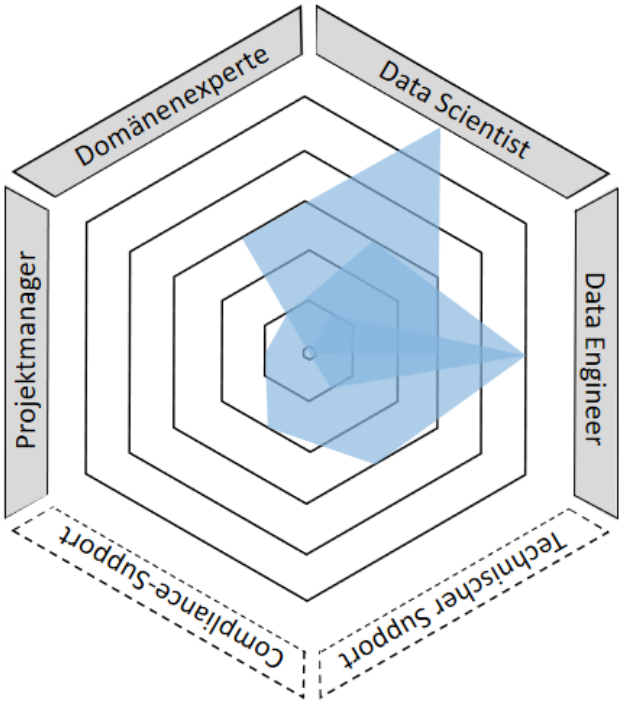


Abbildung 13: Rollen im Schlüsselbereich „Daten“

6.1 Merkmale von „Ursprungsdatenquellen“

Daten werden meist nicht für analytische Aufgaben erhoben. Wird auf bestehende Datenquellen zurückgegriffen, muss zunächst ein Verständnis des Ablaufs, der Erfassung und der Rahmenbedingungen aufgebaut werden, unter denen diese Quellen entstanden sind. Diese Metadaten gilt es in geeigneter Form zu dokumentieren und dem Datensatz zuzuordnen. Metadaten mehrerer Datenquellen werden idealerweise in einem Metadaten-Repository verwaltet, um diese Datenquellen nachhaltig auch in anderen Projekten nutzen zu können.

Merkmale von *Daten(-quellen)* können, wie in Tabelle 1 dargestellt, in vier Kategorien aufgeteilt werden, die für ein Data-Science-Projekt Relevanz haben können. Ziel dieser Darstellung

ist es nicht, eine ausführliche Checkliste aller erdenklichen Merkmale zu bieten, sondern eine strukturierte Herangehensweise an eine elementare Bestandsaufnahme zu erleichtern. Eine ausführlichere Betrachtung der möglichen Merkmale von Datenquellen ist z. B. bei Helfert et al. (2001) zu finden. Die Aufzählung der in Tabelle 2 dargestellten und der letztgenannten Quelle entnommenen Datenqualitätskriterien erhebt dabei, trotz ihres Umfangs, keinen Anspruch auf Vollständigkeit. Die Relevanz der einzelnen Merkmale ist projektindividuell zu bewerten.

Tabelle 1: Beschreibung der Merkmalskategorien im Bereich Datenquellen

Merkmalskategorie	Beschreibung
Beschaffungsaufwand	Die Verfügbarkeit von Daten kann große Auswirkungen darauf haben, welche Analysen letztendlich durchgeführt werden. Sind Daten beispielsweise organisationsintern schon vorhanden und können sie automatisch geladen werden, stellt dies einen wesentlich geringeren Aufwand dar als die Verwendung externer Daten, die zunächst erhoben, gekauft oder ausfindig gemacht werden müssen.
Verwaltungsaufwand	Je nach Menge, Veränderungsgeschwindigkeit und Vertraulichkeit können unterschiedliche Formen der Datenspeicherung gefordert sein. Ein weiteres relevantes Merkmal ist, ob auf die Daten nur einmal oder immer wieder zugegriffen werden soll.
Verarbeitungsaufwand	Wie die Daten transformiert werden müssen, um für Analysen nutzbar zu werden, wird unter anderem von der Granularität, Redundanz und Strukturierung sowie von der bereits in den Quellsystemen durchgeführten Vorverarbeitung beeinflusst.
Datenqualität	Welche Qualität die Daten besitzen, hängt unter anderem von ihrer Aktualität, dem Anteil an fehlenden oder fehlerhaften Werten und ihrer Relevanz in Bezug auf das Data-Science-Projekt ab. Um sich ein Bild von der Qualität machen zu können, ist neben Wissen über ihre Herkunft und den Erhebungsprozess eine explorative Datenanalyse im Vorfeld der Anwendung komplexer Analyseverfahren nötig.

Tabelle 2: Häufig genannte Datenqualitätskriterien, aus: Helfert et al. (2001)

Aktualität	Allgemeingültigkeit	Alter	Änderungshäufigkeit	Aufbereitungsgrad
Bedeutung	Benutzbarkeit	Bestätigungsgrad	Bestimmtheit	Detailliertheit
Effizienz	Eindeutigkeit	Fehlerfreiheit	Flexibilität	Ganzheit
Geltungsdauer	Genauigkeit	Glaubwürdigkeit	Gültigkeit	Handhabbarkeit
Integrität	Informationsgrad	Klarheit	Kompaktheit	Kompression
Konsistenz	Konstanz	Korrektheit	Neutralität	Objektivität
Operationalität	Performance	Portabilität	Präzision	Problemadäquatheit
Prognosegehalt	Prüfbarkeit	Quantifizierbarkeit	Rechtzeitigkeit	Relevanz
Reliabilität	Richtigkeit	Robustheit	Seltenheit	Sicherheit
Signifikanz	Speicherbedarf	Standardisierungsgrad	Subjektadäquatheit	Testbarkeit
Umfang	Unabhängigkeit	Überprüfbarkeit	Übertragbarkeit	Validität
Verdichtungsgrad	Verfügbarkeit	Verfügungsmacht	Verknüpfbarkeit	Verlässlichkeit
Verschlüsselungsgrad	Verständlichkeit	Vertrauenswürdigkeit	Verwendungsbereitschaft	Vollständigkeit
Wahrheitsgehalt	Wahrscheinlichkeit	Wartungsfreundlichkeit	Wiederverwendbarkeit	Wirkungsdauer
Zeitadäquanz	Zeitbezug	Zeitoptimal	Zugänglichkeit	Zuverlässigkeit

6.2 Aufgabe „*Datenaufbereitung*“

Grundsätzlich geht es im Bereich der *Datenaufbereitung* darum, die aus einem oder mehreren Quellsystemen extrahierten Daten in ein geeignetes Format für die anzuwendenden Analyseverfahren zu überführen. Ein weiteres Hauptziel liegt in der Erhöhung der Datenqualität.

In Tabelle 3 werden von Teilnehmerinnen und Teilnehmern häufig genannte Aufgaben aufgeführt und beschrieben, die in Data-Science-Projekten durchgeführt werden müssen.

Tabelle 3: Häufig genannte Aufgaben des Bereichs „Datenaufbereitung“

Aufgabe	Beschreibung
Merkmalserzeugung	Aus den bestehenden Daten können zusätzliche bzw. alternative Merkmale abgeleitet werden.
Datenanonymisierung	Werden innerhalb von Data-Science-Projekten vertrauliche Daten (z. B. personenbezogene Daten) benötigt, müssen diese ggf. zunächst anonymisiert oder pseudonymisiert werden.
Datenaggregation	Wenn Daten einen zu hohen Detaillierungsgrad besitzen, sind sie zu aggregieren.
Datenannotation	Das Annotieren von Merkmalen ist unter anderem nötig, um überwachte Lernverfahren anwenden zu können.
Datenbereinigung	Identifizierte Fehler oder auch fehlende Werte können ggf. manuell oder auch automatisiert bereinigt werden. Wenn dies nicht möglich ist, ist eine Datenfilterung oder Dimensionsreduzierung zu prüfen.
Datenfilterung	Nicht benötigte oder auch fehlerhafte Daten sollten aus der Datenbasis entfernt werden.
Datenintegration	Daten aus verschiedenen Quellen müssen zusammengeführt und vereinheitlicht werden.
Datenstrukturierung	Abhängig von den anzuwendenden Analyseverfahren müssen unstrukturierte Daten zuvor strukturiert werden. Dafür können bspw. Methoden des Natural Language Processing oder der Bilderkennung genutzt werden.
Datentransformation	Transformationen sind durchzuführen, um Daten für die Analyse vorzubereiten. Dies beinhaltet sowohl den bei der explorativen Datenanalyse identifizierten Transformationsbedarf als auch durch das Datenmanagement getriebene Transformationen aus eher technischer Sicht.
Dimensionsreduzierung	Irrelevante oder redundante Merkmale sollten aus der Datenbasis entfernt werden.
Erstellung von Datenaufbereitungsplänen	Vor der Datenaufbereitung sind basierend auf dem Datenbedarf Aufbereitungspläne zu erstellen.
Formatanpassung	Quellformate sind i. d. R. nicht primär für die Anwendung von Analyseverfahren definiert worden. Deshalb ist hier häufig eine Überführung in ein geeignetes Format nötig.
Protokollierung der Datenaufbereitung	Sämtliche Schritte der Datenaufbereitung sind zu protokollieren. Dies ist unter anderem wichtig für die Reproduzierbarkeit und Repräsentativität der Projektergebnisse.
Prozessautomatisierung	Wenn Daten wiederholt bezogen oder auf Grund der Anwendung verschiedener Analyseverfahren aufbereitet werden müssen, kann der Prozess der Aufbereitung ganz oder teilweise automatisiert werden.
Schemaintegration	Schemata aus verschiedenen Quellen müssen zusammengeführt und vereinheitlicht werden.

Die Verarbeitung großer Datenmengen erfordert den Einsatz leistungsfähiger Hard- und Software sowie teilweise innovativer Verfahren. Dies wird im Bereich *IT-Infrastruktur* adressiert.

Als Artefakte der Datenaufbereitung entstehen Skripte, die möglicherweise auch automatisierbar sind, den Prozess auf jeden Fall aber dokumentieren und wiederholbar machen. Das Resultat einer Ausführung dieser Skripte ist eine für das Data-Science-Projekt geeignete, die oben genannten Aufgaben berücksichtigende, aufbereitete Datenbasis. Eine Dokumentation der Aufbereitungsschritte ist genauso erforderlich wie eine Dokumentation der Merkmale in einem Datenkatalog.

In Abbildung 14 ist das Kompetenzprofil von Personen dargestellt, die sich im Bereich *Datenaufbereitung* und den mit ihm direkt verbundenen Aufgaben spezialisieren, in Abbildung 15 sind beteiligte Rollen zu erkennen.

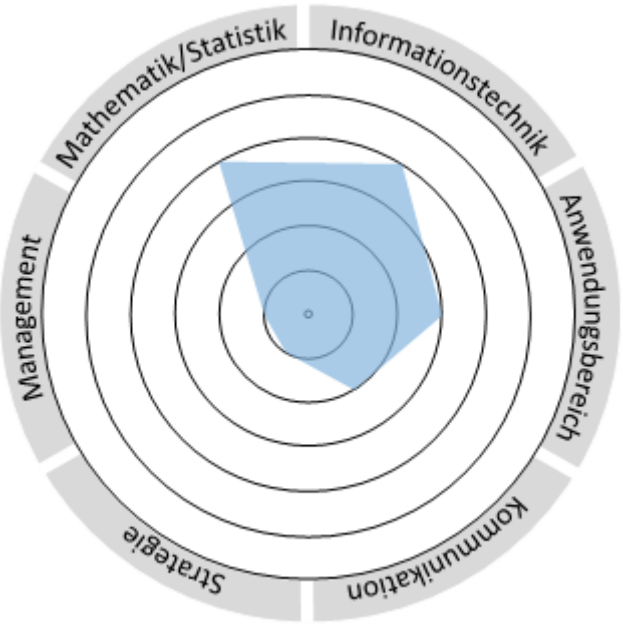


Abbildung 14: Kompetenzprofil des Bereichs „Datenaufbereitung“

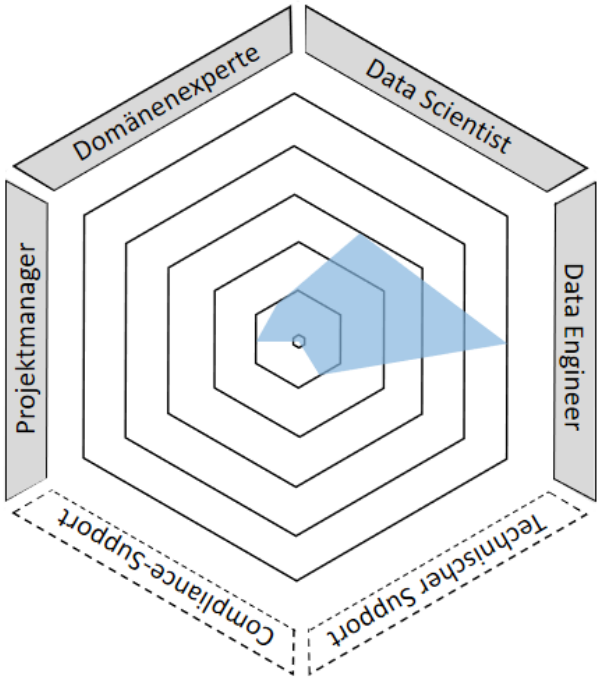


Abbildung 15: Rollen im Bereich „Datenaufbereitung“

6.3 Begleitende Aufgabe „Datenmanagement“

Im Bereich *Datenmanagement* soll der Fokus auf die Verfügbarmachung der benötigten Daten gelegt werden, ohne dabei bereits Anforderungen an eine IT-Infrastruktur zu formulieren. Diese Thematik wird in Kapitel 11 adressiert.

In Tabelle 4 werden von Teilnehmerinnen und Teilnehmern häufig genannte Aufgaben des *Datenmanagements* aufgeführt und beschrieben.

Tabelle 4: Beschreibung der Aufgaben des Bereichs „Datenmanagement“

Aufgabe	Beschreibung
Datenarchivierung	Wenn Analyseverfahren auf Basis identischer Daten reproduzierbar sein sollen und diese Möglichkeit nicht durch die Quellsysteme sichergestellt ist, müssen die verwendeten Daten archiviert werden. Dabei sind neben technischen Herausforderungen bspw. auch Themen wie das Urheberrecht zu berücksichtigen, die eine dauerhafte Speicherung unmöglich machen können.
Datenschutz	Abhängig von den verwendeten Daten kann die Notwendigkeit entstehen, Daten vor unbefugtem Zugriff zu schützen oder sie ggf. ausschließlich anonymisiert oder pseudonymisiert zu speichern. Dabei sind auch verschiedene Zugriffsrollen bzw. Zugriffsrechte zu berücksichtigen.
Datensicherung von aufbereiteten Daten	Es ist zu prüfen, ob die aufbereiteten Daten während der Durchführung des Data-Science-Projekts gesichert werden müssen oder ob sie über die im Bereich *Datenaufbereitung* erarbeiteten Skripte automatisiert wiederhergestellt werden können.
Datenspeicherung von Ursprungsdaten	Es muss geprüft werden, ob die Ursprungsdaten für das Projekt separat gesichert werden. Falls Daten im Laufe des Projekts anwachsen bzw. laufend hinzukommen, müssen geeignete Prozesse und Infrastrukturen vorgesehen werden.
Datenzugriff	Daten können entweder einmalig, in definierten Abständen über eine Batchverarbeitung oder in (Nahe-)Echtzeit als Stream geladen und auch verarbeitet werden. Im Kontext von Open Science kann ggf. auch Dritten Zugriff auf die Daten gewährt werden.
Metadatenmanagement	Aus den Quellen extrahierte oder über die durchgeführten Aufgaben ergänzte bzw. ermittelte Metadaten sind sinnvoll zu verwalten.

Als Artefakt dieses Bereichs ist eine Erweiterung des Datenkatalogs zu nennen, die die Nachvollziehbarkeit des Datenmanagements sicherstellt.

In Abbildung 16 ist das Kompetenzprofil von Personen dargestellt, die sich in dem Bereich *Datenmanagement* und den mit diesem direkt verbundenen Aufgaben spezialisieren, in Abbildung 17 sind beteiligte Rollen zu erkennen.

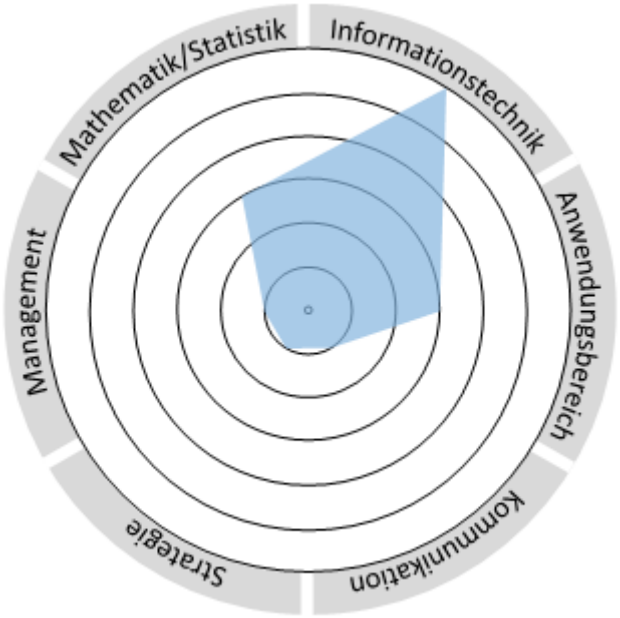


Abbildung 16: Kompetenzprofil des Bereichs „Datenmanagement“

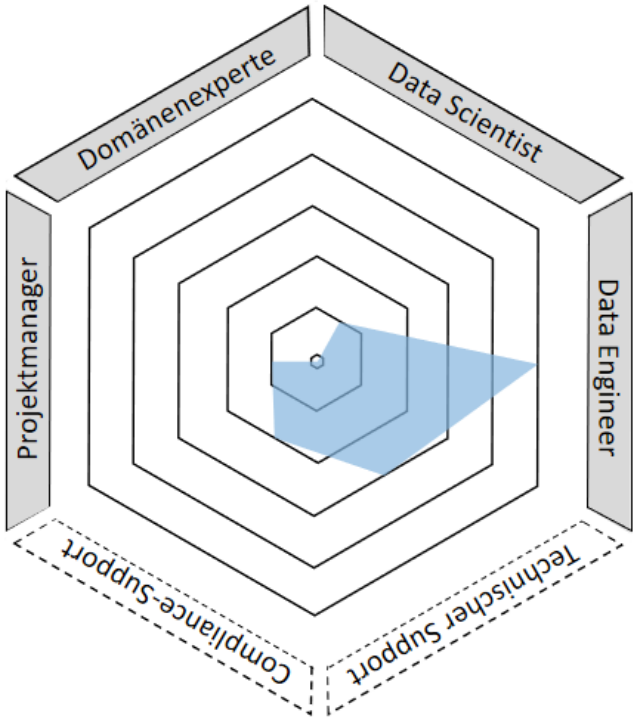


Abbildung 17: Rollen im Bereich „Datenmanagement“

6.4 Begleitende Aufgabe „Explorative Datenanalyse“

Im Bereich *Explorative Datenanalyse* geht es darum, ein besseres inhaltliches Verständnis der vorliegenden Daten zu erlangen und mögliche Ansatzpunkte für spätere, tiefergehende Analysen zu bestimmen. Auch soll geklärt werden, ob die Menge und Qualität der vorliegenden Daten für die gewählte Fragestellung ausreichend ist und ob die geplante Analyse noch weitere Datenaufbereitungsschritte erfordert.

In Tabelle 5 werden von Teilnehmerinnen und Teilnehmern häufig genannte Aufgaben der explorativen Datenanalyse aufgeführt und beschrieben.

Tabelle 5: Beschreibung der Aufgaben innerhalb des Bereichs der „Explorativen Datenanalyse“

Aufgabe	Beschreibung
Ausreißeridentifikation	Ausreißer können das spätere Analyseergebnis stark beeinflussen. Es muss entschieden werden, ob die identifizierten Ausreißer realen Datenpunkten entsprechen oder durch andere Effekte entstanden sind. Entsprechend sind diese Werte gegebenenfalls herauszufiltern oder zu ersetzen.
Datenvalidierung	Unter Nutzung von Domänenwissen können in Datensätzen Werte identifiziert werden, die zwar formal einwandfrei, inhaltlich aber nicht korrekt oder sinnvoll sind.
Datenvisualisierung	Durch einfache Diagramme (z. B. Histogramme, Linien- oder Punktdiagramme) wird die Verteilung der vorliegenden Daten deutlich und können einfache Zusammenhänge zwischen Attributen aufgedeckt werden.
Identifikation zentraler Attribute	Die spätere Datenanalyse kann effizienter durchgeführt werden, wenn die Datensätze weniger Attribute besitzen. Ziel ist es daher, möglichst zentrale, aussagekräftige Attribute zu identifizieren bzw. unerhebliche Attribute auszuschließen. Dabei wird häufig auf Domänen- und Statistikwissen zurückgegriffen.
Inhaltliches Verständnis	Die Daten sind bzgl. ihrer Eignung in der spezifischen Domäne und unter Berücksichtigung der Ziele des aktuellen Data-Science-Projekts zu bewerten.
Statistische Analysen	Einfache statistische Maße wie Median, Mittelwert, Standardabweichung oder Korrelation helfen dabei, schnell ein besseres Verständnis der vorliegenden Daten zu erlangen und unerwartete Abweichungen aufzuspüren.
Untersuchung der Notwendigkeit von Datentransformationen	Um die Vergleichbarkeit von Attributen zu gewährleisten, ist häufig eine Normierung der Daten notwendig. Ein weiterer Grund für Transformationen sind die später anzuwendenden Analyseverfahren, die häufig eine bestimmte Datenbeschaffenheit voraussetzen. Die Identifikation der Transformationsaufgaben ist Teil der *explorativen Datenanalyse*, die Umsetzung ist im Bereich der *Datenaufbereitung* anzusiedeln.
Untersuchung fehlender Werte	Fehlen in Datensätzen Attributwerte, muss entschieden werden, ob diese Datensätze oder die betroffenen Attribute gelöscht werden können. Da dies die Menge, die Repräsentativität und die Aussagekraft der zugrundeliegenden Daten beeinflussen kann, ist auch ein Ersatz der fehlenden Werte denkbar. Die Identifikation geeigneter Verfahren zur Behandlung fehlender Werte ist Teil der *explorativen Datenanalyse,* die Umsetzung entsprechender Maßnahmen ist im Bereich der *Datenaufbereitung* anzusiedeln.

Da bei der explorativen Datenanalyse gerade noch nicht bekannte Aspekte aufgespürt werden sollen, gibt es keine feste Abfolge der anzuwendenden Verfahren. Neben der Datenvisualisierung kommen jedoch meist verschiedene statistische Methoden zum Einsatz, etwa Korrelations-, Faktor- und Clusteranalysen sowie statistische und ggf. auch kausale Modellierungen. Die entstandenen Visualisierungen und Modelle stellen dementsprechend die Artefakte dar. Zu dokumentieren sind identifizierte Probleme bei der Datenqualität sowie alle nötigen Änderungen des Datenmaterials. Da bei der explorativen Datenanalyse in schneller Abfolge viele

Hypothesen untersucht und eventuell auch wieder verworfen werden, verzichtet man bei der Dokumentation jedoch meist auf einen hohen Detailgrad.

In Abbildung 18 ist das Kompetenzprofil von Personen dargestellt, die sich in dem Bereich *Explorative Datenanalyse* und den mit diesem direkt verbundenen Aufgaben spezialisieren, in Abbildung 19 sind beteiligte Rollen zu erkennen.

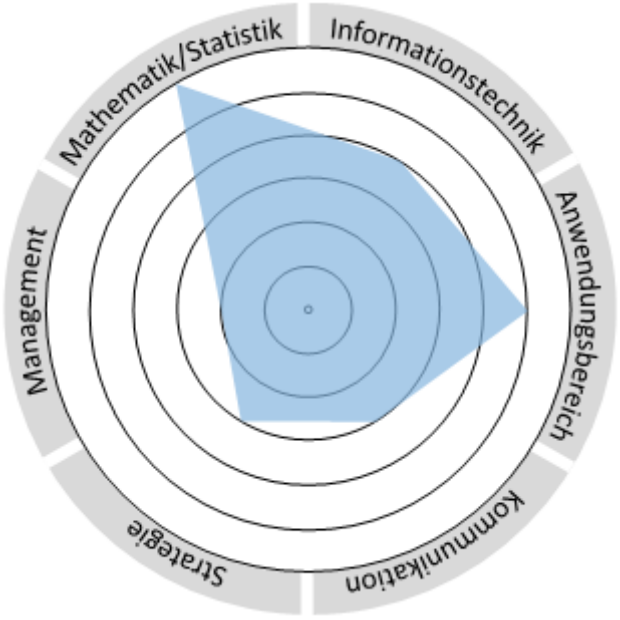


Abbildung 18: Kompetenzprofil des Bereichs „Explorative Datenanalyse“

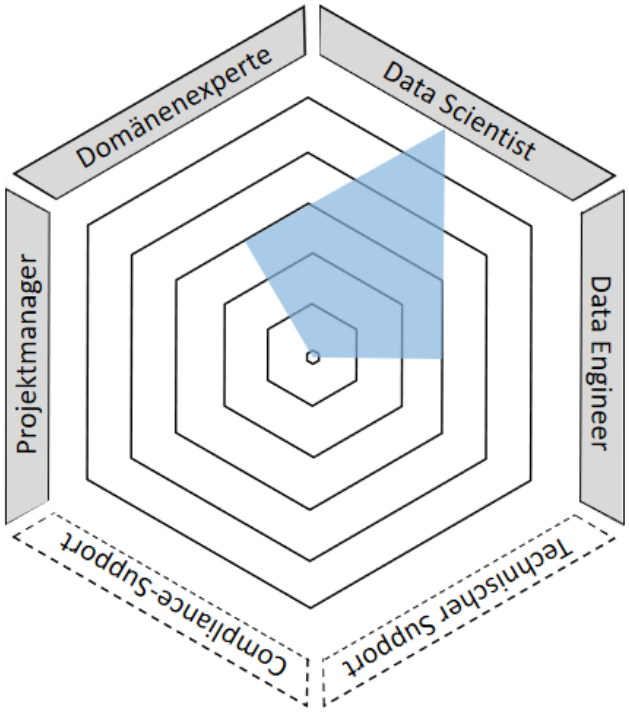


Abbildung 19: Rollen im Bereich „Explorative Datenanalyse“

6.5 Merkmale „analytischer Datenquellen“

Ursprungsdatenquellen und analytische Datenquellen stimmen zwar in wesentlichen Merkmalen überein, aber durch die Aufbereitung der Daten für Data-Science-Anwendungen ergeben sich im Detail Unterschiede in Inhalt, Umfang, Struktur und Format. So wird etwa hinsichtlich

des Analyseziels angestrebt, dass die Attribute bereinigt und möglichst redundanzfrei sind. Weiterhin sollen die Attribute für das Analyseziel eine besondere Relevanz haben. Jedoch ist gerade die Bewertung der Relevanz zu einem frühen Stadium eines Data-Science-Projekts nicht immer eindeutig möglich, eine Einschätzung durch Domänenexperten ist daher empfehlenswert. In Abhängigkeit von den anzuwendenden Analyseverfahren müssen die Datenformate und Skalenniveaus angepasst werden. Viele Lernverfahren verarbeiten beispielsweise ausschließlich numerische Attribute. Analytische Datenquellen können meist durch projektbeteiligte Personengruppen eigenständig bearbeitet werden. Der Datenzugriff kann dabei in Echtzeit, kontinuierlich oder einmalig erfolgen. Ergänzend können Metadaten der Datenquellen zur Verfügung gestellt werden.

7. SCHLÜSSELBEREICH ANALYSEVERFAHREN

Der Schlüsselbereich *Analyseverfahren* wird in die in Abbildung 20 dargestellten und in Verbindung stehenden Teilbereiche untergliedert, die in den folgenden Unterkapiteln einzeln betrachtet werden.

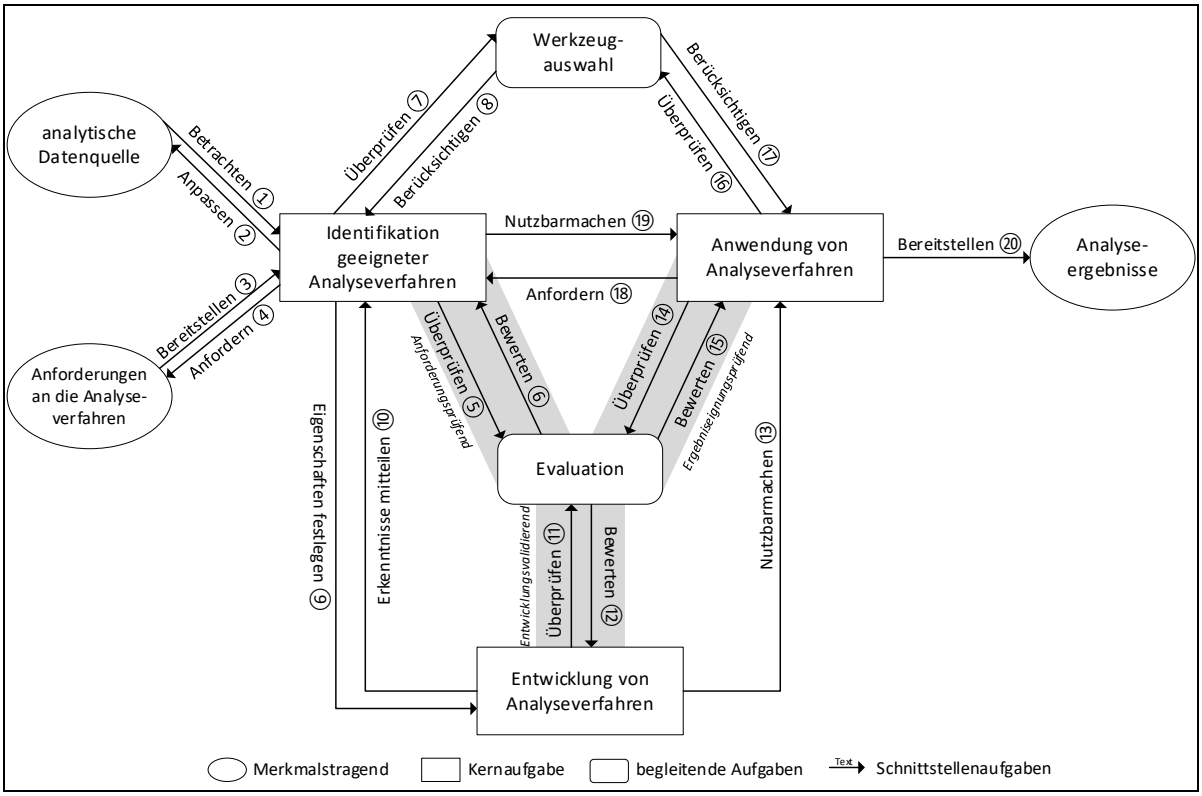


Abbildung 20: Schlüsselbereich „Analyseverfahren“ (Beschreibung siehe Folgeseite)

① Die analytische Datenquelle wird durch Bearbeitung der im Schlüsselbereich *Daten* beschriebenen Aufgaben erstellt. Die Identifikation geeigneter Analyseverfahren ist nur unter Berücksichtigung von Merkmalen der zur Verfügung stehenden Daten möglich.

② Nach der Identifikation möglicherweise geeigneter Analyseverfahren kann es zur Sicherstellung der Anwendbarkeit nötig sein, die analytische Datenquelle anzupassen.

③ Bei der Identifikation geeigneter Analyseverfahren sind die definierten nicht-funktionalen Anforderungen zu berücksichtigen.

④ Sollte kein geeignetes Analyseverfahren identifiziert werden können, kann es ggf. sinnvoll oder notwendig sein, die festgelegten Anforderungen anzupassen.

⑤ Ausgewählte Verfahren sind dahingehend einer Evaluation zu unterziehen, ob gegebene Analyseanforderungen erfüllt werden können.

⑥ Die Ergebnisse der Evaluation sind bei der Identifikation geeigneter Analyseverfahren zu berücksichtigen.

⑦ Die Auswahl geeigneter Werkzeuge ist unter Berücksichtigung identifizierter Analyseverfahren zu prüfen.

⑧ Die Ergebnisse der Werkzeugauswahl sind bei der Identifikation geeigneter Analyseverfahren zu berücksichtigen.

⑨ In Sonderfällen ist eine Entwicklung von Analyseverfahren nötig. Die bei der Identifikation geeigneter Analyseverfahren im Detail betrachteten Anforderungen sind dabei zu berücksichtigen.

⑩ Sollte der Schritt der Entwicklung von Analyseverfahren nicht erfolgreich sein und diese Tatsache nicht zu einem Projektabbruch führen, sind die gewonnenen Erkenntnisse bei der erneuten Identifikation geeigneter Analyseverfahren zu berücksichtigen.

⑪ Während der Entwicklung muss die Eignung des Analyseverfahrens immer wieder evaluiert werden.

⑫ Die Erkenntnisse aus der Evaluation sind bei der (Weiter-)Entwicklung des Analyseverfahrens zu berücksichtigen.

⑬ Das entwickelte Analyseverfahren ist / die entwickelten Analyseverfahren sind für die Anwendung zur Verfügung zu stellen.

⑭ Bei der Anwendung von Analyseverfahren sind verschiedene Parametrisierungen einer Evaluation zu unterziehen.

⑮ Die Ergebnisse der Evaluation sind bei der Anwendung von Analyseverfahren zu berücksichtigen.

⑯ Die Auswahl geeigneter Werkzeuge für die Anwendung von Analyseverfahren ist zu prüfen.

⑰ Die Ergebnisse der Werkzeugauswahl sind bei der Anwendung von Analyseverfahren zu berücksichtigen.

⑱ Sollte die Anwendung von Analyseverfahren keine akzeptablen Ergebnisse liefern, muss der Prozess abgebrochen oder zum Schritt der Identifikation geeigneter Analyseverfahren zurückgekehrt werden.

⑲ Geeignete Analyseverfahren können angewendet werden.

⑳ Führt die Anwendung von Analyseverfahren zu akzeptablen Ergebnissen, können diese für die Nutzbarmachung bereitgestellt werden.

In Abbildung 21 ist das Kompetenzprofil von Personen dargestellt, die sich im Schlüsselbereich *Analyseverfahren* spezialisieren. Abbildung 22 stellt diejenigen Rollen dar, die bei der Betrachtung dieses Schlüsselbereichs eine Relevanz besitzen.

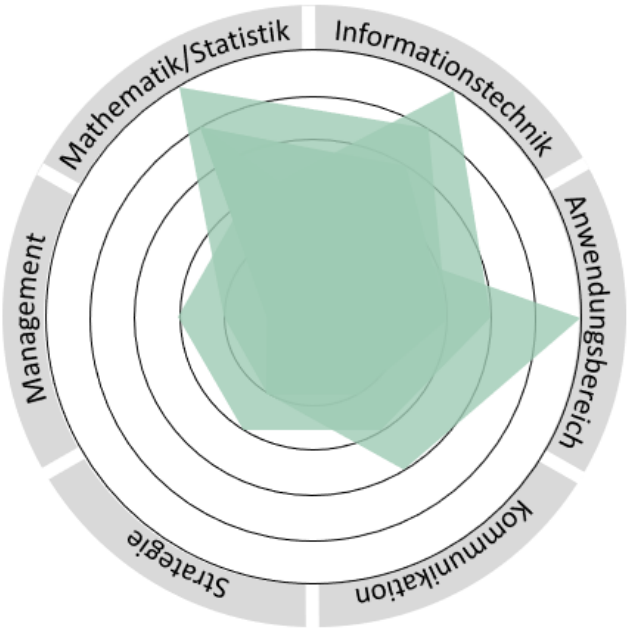


Abbildung 21: Kompetenzprofil des Schlüsselbereichs „Analyseverfahren“

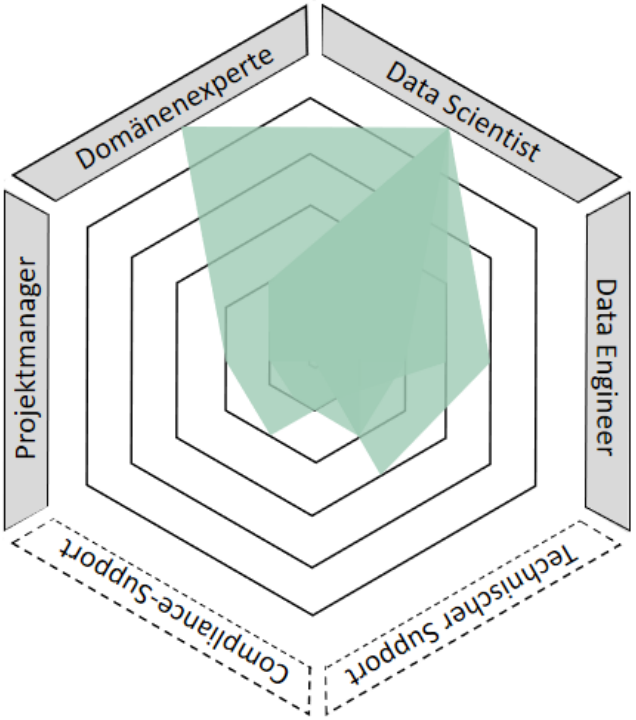


Abbildung 22: Rollen im Schlüsselbereich „Analyseverfahren“

7.1 Merkmale „*analytischer Datenquellen*“

Die Identifikation geeigneter Analyseverfahren fußt auf den Merkmalen der vorliegenden analytischen Datenquelle (vgl. Kapitel 4.1). Durch die betrachtete analytische Fragestellung bzw. das geforderte Analyseergebnis entstehen häufig spezielle Anforderungen an die Datenquelle.

Andersherum können unabänderliche Merkmale der analytischen Datenquelle auch die Menge der beantwortbaren Fragestellungen einschränken.

7.2 Merkmale der „*Anforderungen an das Analyseverfahren*“

Die in diesem Abschnitt betrachteten Merkmale stellen die nicht-funktionalen Anforderungen an Analyseverfahren dar. Im individuellen Projekt können sie auch bereits mit expliziten Grenzwerten versehen sein und als Spezifikationsanforderungen verwendet werden.

In Tabelle 6 werden von Teilnehmerinnen und Teilnehmern häufig genannte Merkmale der Anforderungen an das Analyseverfahren aufgeführt und beschrieben.

Tabelle 6: Häufig genannte Merkmale des Bereichs „Anforderungen an das Analyseverfahren“

Merkmal	Beschreibung
Anforderungsabdeckung	Nicht immer können die gewählten Analyseverfahren alle Anwendungsanforderungen vollständig erfüllen. Wünschenswert ist dennoch ein möglichst hoher Abdeckungsgrad.
Effizienz	Das Verfahren muss auf die IT-Infrastruktur in geeigneter Zeit angewendet werden können. Je weniger Daten und Rechenzeit benötigt werden, desto einfacher lässt sich das Verfahren in den laufenden Betrieb der Organisation integrieren und desto wirtschaftlicher lässt es sich anwenden.
Innovative Problemlösung	Das Verfahren muss ein Problem lösen, das durch bestehende Verfahren noch nicht im selben Umfang oder in derselben Qualität gelöst wird.
Reproduzierbarkeit	Damit das Ergebnis (von anderen) reproduziert und das verwendete Verfahren im Idealfall in unterschiedlichen Szenarien eingesetzt werden kann, müssen Technologien und Algorithmen eingesetzt werden, die ausführlich dokumentiert und allgemein verfügbar sind.
Robustheit	Die eingesetzten Verfahren sollten möglichst fehlerunanfällig sein. Beispielsweise ist es hilfreich, wenn fehlerhafte Daten oder Ausreißer automatisch erkannt werden oder das Ergebnis nur geringfügig beeinflussen.
Skalierbarkeit	In der Praxis nehmen die Menge und/oder die Dimension der neu zu analysierenden Daten im Zeitverlauf häufig erheblich zu. Daher ist es von Vorteil, wenn das gewählte Verfahren auch eine wachsende Datenmenge mit vertretbarem Zusatzaufwand verarbeiten kann.
Umsetzbarkeit	Das Verfahren muss mit zur Verfügung stehenden Ressourcen (z. B. technischer Infrastruktur und Fachpersonal) umsetzbar sein. Zudem sollte es möglichst wenig Aufwand in der Umsetzung erfordern.
Validität	Die Vorhersagen oder abgeleiteten Strukturen sollten zuverlässig die Realität der Fragestellung möglichst zutreffend widerspiegeln. Die akzeptable Fehlertoleranz ist dabei von der Problemstellung abhängig.
Verständlichkeit	Die Ergebnisse der Verfahren sollten nach Möglichkeit nachvollziehbar sein und sich leicht kommunizieren und/oder visualisieren lassen.

7.3 Aufgabe „*Identifikation geeigneter Analyseverfahren*“

An dieser Stelle sollte bereits klar sein, dass sich die gegebene Aufgabenstellung tatsächlich mit Hilfe von Data Science lösen lässt, d. h., dass sie auf der einen Seite ein potenziell lösbares Problem darstellt, auf der anderen Seite aber auch nicht so trivial ist, dass sie beispielsweise mit Hilfe eines Standardberichtes gelöst werden kann. Die Identifikation von geeigneten Analyseverfahren stellt häufig eine große Herausforderung dar. Obwohl es eine sehr große Anzahl von Analyseverfahren gibt, kann es auch sein, dass keines für die Problemstellung geeignet

ist. In diesem Fall ist zu prüfen, ob bestimmte Projektrahmenbedingungen geändert werden können, ob die Entwicklung eines neuen Analyseverfahrens denkbar ist oder ob das Projekt nötigenfalls abgebrochen werden muss.

In dieser Phase stehen die Gewinnung eines Überblicks über existierende Verfahren und die Identifikation der besten Verfahren für die Anwendung im Fokus. Da ohne eine weitere Evaluation noch keine abschließende Auswahl getroffen werden kann, können zunächst mehrere Verfahren für die weitere Bewertung berücksichtigt werden. Die Entscheidung für eine Neuentwicklung von Verfahren sollte unter Berücksichtigung des Aufwandes und der bestehenden Unsicherheit getroffen werden.

In Tabelle 7 werden von Teilnehmerinnen und Teilnehmern häufig genannte Aufgaben bei der Identifikation geeigneter Analyseverfahren aufgeführt und beschrieben.

Tabelle 7: Häufig genannte Aufgaben des Bereichs „Identifikation geeigneter Analyseverfahren“

Aufgabe	**Beschreibung**
Identifikation von Anforderungen	Bevor verschiedene Verfahren geprüft werden, ist Klarheit darüber zu schaffen, welche Probleme durch sie gelöst werden sollen.
Bestimmung der Problemklasse	Anhand der identifizierten Anforderungen kann die Problemstellung meist einer konkreten Problemklasse zugeordnet werden, die dann die Suche nach einem konkreten Analyseverfahren leiten kann.
Recherche zu vergleichbaren Problemstellungen	Bei der Suche nach geeigneten Analyseverfahren ist es hilfreich zu recherchieren, ob es Publikationen zu ähnlichen Anwendungsfällen gibt.
Bestimmung potenziell geeigneter Verfahren	Vor dem Hintergrund der Problemklasse und auf Basis der Recherche zu vergleichbaren Problemstellungen können nun grundsätzlich erfolgversprechende Analyseverfahren/Analyseverfahrensvarianten benannt werden.
Auswahl	Nach Aufstellung der in Frage kommenden Verfahren sollten diejenigen ausgewählt werden, die den projektspezifischen Kriterien und Ressourcen am besten entsprechen.

Als Artefakt dieser Phase entsteht eine Liste von Analyseverfahren, in der auch Begründungen enthalten sind, weshalb diese Verfahren für die Fragestellung geeignet sind. Sollten keine passenden Analyseverfahren identifiziert werden, können Verfahren ausgewählt werden, die weiterzuentwickeln sind, ggf. kann sogar bereits ein Prototyp erstellt werden, der die Eignung der Auswahl sicherstellt.

Die Erkenntnisse dieser Phase sollten so dokumentiert werden, dass nicht nur die Auswahl für das aktuelle Projekt begründet wird, sondern auch die Entscheidungen in einer Form festgehalten werden, die für zukünftige Fragestellungen angewendet werden können (etwa in einem internen Wiki).

In Abbildung 23 ist das Kompetenzprofil von Personen dargestellt, die sich in dem Bereich *Identifikation geeigneter Analyseverfahren* und den mit diesem direkt verbundenen Aufgaben spezialisieren, in Abbildung 24 sind beteiligte Rollen zu erkennen.

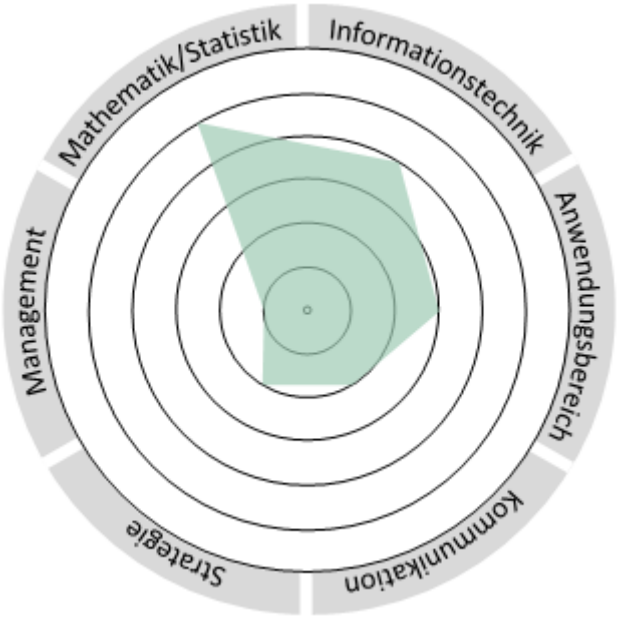


Abbildung 23: Kompetenzprofil des Bereichs „Identifikation geeigneter Analyseverfahren“

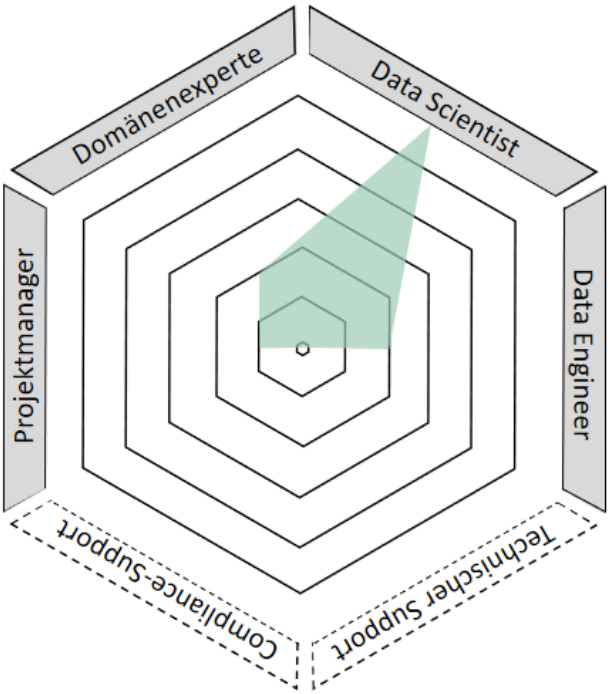


Abbildung 24: Rollen im Bereich „Identifikation geeigneter Analyseverfahren“

7.4 Aufgabe „Anwendung von Analyseverfahren“

Für die korrekte Anwendung von Analyseverfahren sind detaillierte Kenntnisse über bestehende Verfahren vonnöten. Werden Verfahren falsch angewendet, führt dies zu willkürlichen Ergebnissen, was zur Folge hat, dass fehlerhafte oder falsche Aussagen entstehen.

Es ist zu gewährleisten, dass anzuwendende Verfahren die jeweiligen Aufgaben in geeigneter Form erfüllen. Dies muss bereits bei der Identifikation (vgl. Kapitel 7.3) eine hervorgehobene Rolle spielen. Sichergestellt werden kann das jedoch erst in der tatsächlichen Anwendung auf die zu analysierenden Daten.

Ziel ist es, das beste Analyseergebnis zu finden. Im Detail hängt dies von dem angewandten Verfahren und den individuellen Domänenanforderungen ab. Bei einigen Verfahren ist zu entscheiden, ob ein möglichst genaues Ergebnis das Ziel sein soll oder ein Modell, das auf möglichst viele Szenarien anwendbar ist.

In Tabelle 8 werden von Teilnehmerinnen und Teilnehmern häufig genannte Aufgaben bei der Anwendung bestehender Analyseverfahren aufgeführt und beschrieben.

Tabelle 8: Häufig genannte Aufgaben des Bereichs „Anwendung von Analyseverfahren“

Aufgabe	Beschreibung
Aufsetzen einer Entwicklungsumgebung	Besonders wenn mehrere Anwender beteiligt sind, sollte es eine leistungsstarke und gut zugängliche Entwicklungsumgebung mit Versionsverwaltung geben, um einen langfristig reibungslosen Ablauf des Data-Science-Projekts zu gewährleisten.
Konstruktion der Prozesse	Die einzelnen Bestandteile der Prozesse müssen angelegt und in die richtige Reihenfolge gebracht werden.
Dimensionsreduktion	Da viele Algorithmen auf hochdimensionalen Daten keine guten Ergebnisse liefern, sollte geprüft werden, ob Datendimensionen entfernt oder zusammengefasst werden können.
Sicherstellung der Validität	Schon während der Konstruktion der Modelle kann z. B. durch eine Aufteilung in Trainings- und Testpartitionen sowie durch Kreuzvalidierung die Wahrscheinlichkeit einer Überanpassung verringert werden.
Berücksichtigung mehrerer Analyseverfahren	Gegebenenfalls sind mehrere Analyseverfahren zu erproben oder auch durch die Bildung von Ensembles zu kombinieren.
Auswahl der besten Parameterkonfiguration	Ein systematisches Testen verschiedener Kombinationen zur Auswahl geeigneter oder gewünschter Einstellungen ist nötig.
Abwägen zwischen Zeit und Nutzen	Die Qualität des Ergebnisses muss für die Problemstellung geeignet sein. Die gesamten Rechenkosten für die Analyse dürfen dabei aber den Nutzen des Modells nicht übersteigen.
Sicherstellung von Reproduzierbarkeit und Transparenz	Unter anderem durch Speichern der transformierten Daten und aller Konfigurationen des Trainingsprozesses (z. B. verwendeter Seeds) sind Reproduzierbarkeit und Transparenz sicherzustellen.

Ein großer Teil der entstehenden Artefakte und benötigten Dokumentationen hängt von dem individuellen Projekt ab, ist also untrennbar mit der Problemstellung, den verwendeten Daten und den angewandten Analyseverfahren verbunden. Grundsätzlich entstehen als Artefakte eine Dokumentation der Analysedurchführung und der Evaluationsergebnisse (auch von Zwischenergebnissen und Grafiken), eine Begründung der Auswahl für das finale Modell, eine Sicherung der Entwicklungsumgebung, die trainierten Modelle, eine Schnittstellendokumenta-

tion und die Parameterkonfigurationen. Fachliche Informationen sollten für die Domänenexperten gut verständlich aufbereitet werden, zudem mit Hinweisen, welche Fehler und Auffälligkeiten es gegeben hat und welche weiteren Problemstellungen mit Hilfe der Analyseverfahren untersucht werden könnten. Abhängig von den verwendeten Werkzeugen wird bereits beim Analysevorgang selbst eine grundlegende Dokumentation erstellt.

In Abbildung 25 ist das Kompetenzprofil von Personen dargestellt, die sich in dem Bereich *Anwendung von Analyseverfahren* und den mit diesem direkt verbundenen Aufgaben spezialisieren, in Abbildung 26 sind beteiligte Rollen zu erkennen.

Abbildung 25: Kompetenzprofil des Bereichs „Anwendung von Analyseverfahren“

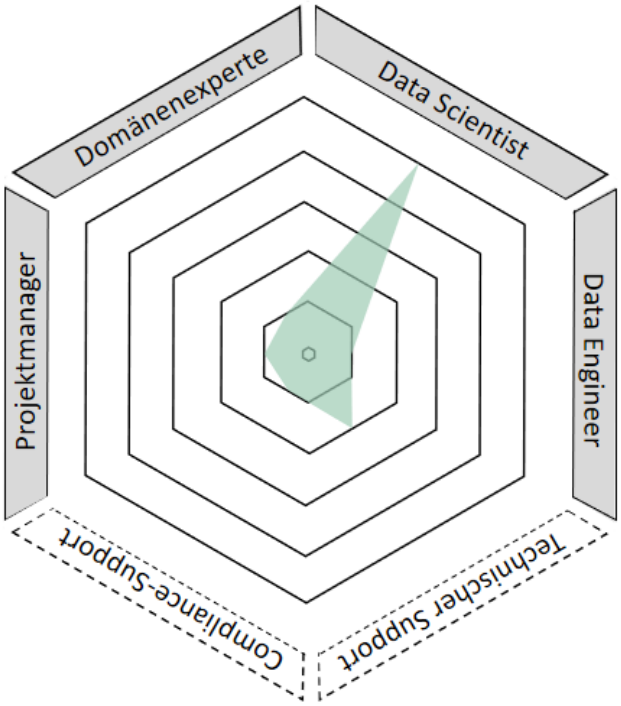


Abbildung 26: Rollen im Bereich „Anwendung von Analyseverfahren“

7.5 Aufgabe „Werkzeugauswahl“

Ziel der Werkzeugauswahl ist es, für die ausgewählten Verfahren eine passende Implementierungsinfrastruktur zu identifizieren. Dies bezieht sich sowohl auf Hardware als auch auf Software. Somit überschneidet sich dieser Bereich auch teilweise mit dem Schlüsselbereich *IT-Infrastruktur (*vgl. Kapitel 11*)*, der jedoch normalerweise nicht zur Kernaufgabe von Data Scientists gehört und sehr viel weitläufiger gefasst werden muss. Unter dem Begriff *Werkzeugauswahl* ist somit eher die Selektion einzelner Komponenten der IT-Landschaft zu verstehen, die im Kontext der Fragestellung zur direkten Lösung beitragen. Organisationsabhängig kann es möglich sein, dass die Hard- und Software bereits vorgegeben sind und ihre Auswahl somit nicht mehr in den Rahmen des Projekts fällt, ihre notwendige Verwendung allerdings als Anforderung zu berücksichtigen ist.

In Tabelle 9 werden von Teilnehmerinnen und Teilnehmern häufig genannte Aufgaben der *Werkzeugauswahl* aufgeführt und beschrieben.

Tabelle 9: Häufig genannte Aufgaben des Bereichs „Werkzeugauswahl“

Aufgabe	Beschreibung
Recherche zu geeigneter Software	Sobald abzuschätzen ist, welche Analyseverfahren in Frage kommen, sollte geklärt werden, mit welcher Software die Verfahren umzusetzen sind und wie die Software beschafft oder geschaffen werden kann, wenn sie noch nicht vorhanden ist.
Recherche zu geeigneter Hardware	Abhängig davon, wie viel Rechenleistung benötigt und ob die Anwendung lokal oder in einer Cloud durchgeführt wird, kann unterschiedliche Hardware benötigt werden.
Abgleich mit den vorhandenen Fähigkeiten im Projektteam	Kann ein Werkzeug nicht oder nur unzureichend bedient werden, dann muss entweder ein anderes Werkzeug ausgewählt oder eine Fortbildungsmaßnahme eingeleitet werden oder es müssen externe Ressourcen hinzugezogen werden.
Bewertung der Werkzeugeignung	Wenn ein Werkzeug nicht vollständig kompatibel mit dem übrigen Workflow des Projekts ist, muss ein Kompromiss zwischen der vollkommenen Umsetzung des angestrebten Verfahrens und der Integrierung in die restliche Infrastruktur gefunden werden.
Qualitätssicherung bei der Implementierung	Die Qualität der Implementierung ist z. B. durch Software-Validierung, Peer Review o. Ä. sicherzustellen.

Im Gegensatz zu den Anforderungen an die Implementierungsinfrastruktur ist eine ausführliche Dokumentation des Auswahlprozesses in der Regel nur bei umfangreichen Projekten nötig.

In Abbildung 27 ist das Kompetenzprofil von Personen dargestellt, die sich in dem Bereich der *Werkzeugauswahl* und den mit diesem direkt verbundenen Aufgaben spezialisieren, in Abbildung 28 sind beteiligte Rollen zu erkennen. Die Darstellungen sind durch eine Aggregation der Rückmeldungen von Teilnehmerinnen und Teilnehmern entstanden.

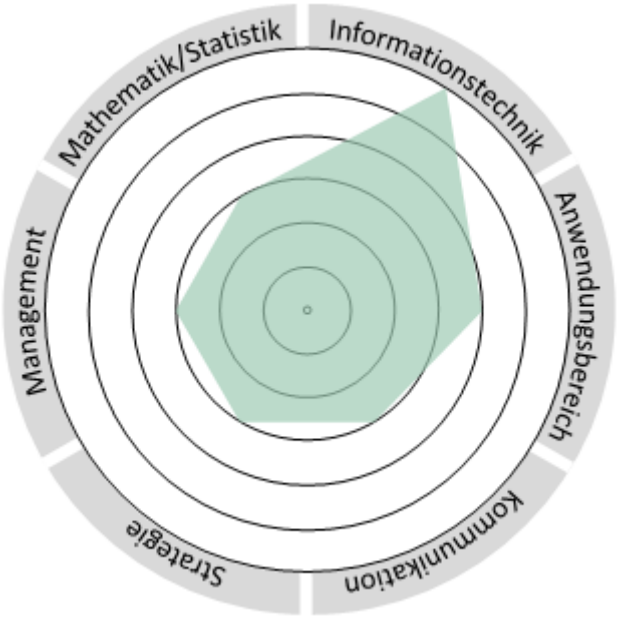


Abbildung 27: Kompetenzprofil im Bereich „Werkzeugauswahl“

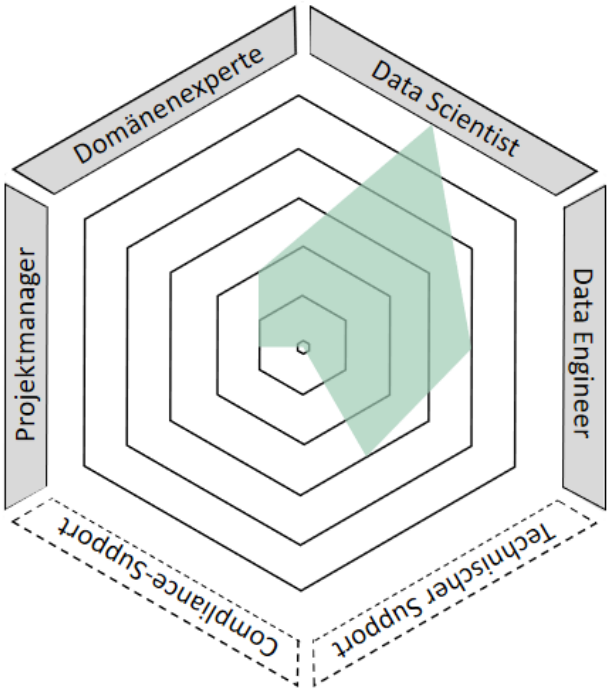


Abbildung 28: Rollen im Bereich „Werkzeugauswahl“

7.6 Aufgabe „Entwicklung von Analyseverfahren“

Wenn kein geeignetes Analyseverfahren existiert, müssen – falls möglich – bestehende Verfahren angepasst bzw. zusammengeführt werden oder es können vollständig neue Lösungen entwickelt werden. Dabei ist festzulegen, ob das Verfahren möglichst vielseitig anwendbar sein soll oder für den speziellen Anwendungsfall bzw. die vorliegenden Daten optimiert werden soll. Betrachtet werden muss außerdem die Effizienz der Eigenentwicklung, überflüssige Arbeiten, z. B. dadurch, dass bestehende (Hilfs-)Verfahren nicht genutzt werden, sind zu vermeiden. Das neuentwickelte Verfahren muss in die Implementierungsinfrastruktur eingefügt werden, Zeit- sowie Budgetbeschränkungen sind zu berücksichtigen.

In Tabelle 10 werden von Teilnehmerinnen und Teilnehmern häufig genannte Aufgaben der Entwicklung neuer Analyseverfahren aufgeführt und beschrieben.

Tabelle 10: Häufig genannte Aufgaben des Bereichs „Entwicklung von Analyseverfahren“

Aufgabe	Beschreibung
Festlegung von Kriterien	Es ist klar und genau zu definieren, was das Verfahren können soll und was nicht.
Bestimmung der Differenz zu relevanten bestehenden Verfahren	Eine Bestimmung der Unzulänglichkeiten relevanter bestehender Verfahren im Hinblick auf die Problemstellung (Gap-Analyse) ist durchzuführen.
Festlegung des Vorgehens	Es ist zu entscheiden, ob ein komplett neues Verfahren entwickelt werden soll oder ob auf einer bestehenden Idee aufgebaut werden kann.
Konzeption des Verfahrens	Eine technische Konzeption des neuen Analyseverfahrens ist durchzuführen.
Testen des Verfahrens	Eine empirische Modell-Validierung und Reliabilitätstests sind genauso durchzuführen wie ein Vergleich mit bestehenden Verfahren.
Implementierung	Das Analyseverfahren ist technisch umzusetzen.

Die Entwicklung eines neuen Analyseverfahrens muss sorgfältig und umfangreich dokumentiert werden. Dazu können beispielsweise gehören:

- Eine Begründung für die Neuentwicklung
- Die vollständige Herleitung des Verfahrens
- Eine Beschreibung des entwickelten Modells (inklusive aller getroffenen Annahmen und vorgenommenen Vereinfachungen)
- Die theoretische Basis / zugrundeliegende Mathematik
- Die ausführliche Darstellung des entwickelten Algorithmus
- Die Voraussetzungen für die Anwendung
- Eine Beschreibung der Ein- und Ausgaben
- Die Darstellung von Abhängigkeiten von bestehender Software
- Die Dokumentation des Verfahrens auf Code-Ebene
- Verschiedene Qualitätskriterien (Robustheit, Validität, Objektivität, Reliabilität)
- Ein Benutzerhandbuch
- Anwendungsbeispiele
- Ein Lessons-learned-Dokument
- Schwächen und Stärken des Verfahrens
- Potenzielle Weiterentwicklungsmöglichkeiten

In Abbildung 29 ist das Kompetenzprofil von Personen dargestellt, die sich in dem Bereich *Entwicklung von Analyseverfahren* und den mit diesem direkt verbundenen Aufgaben spezialisieren, in Abbildung 30 sind beteiligte Rollen zu erkennen.

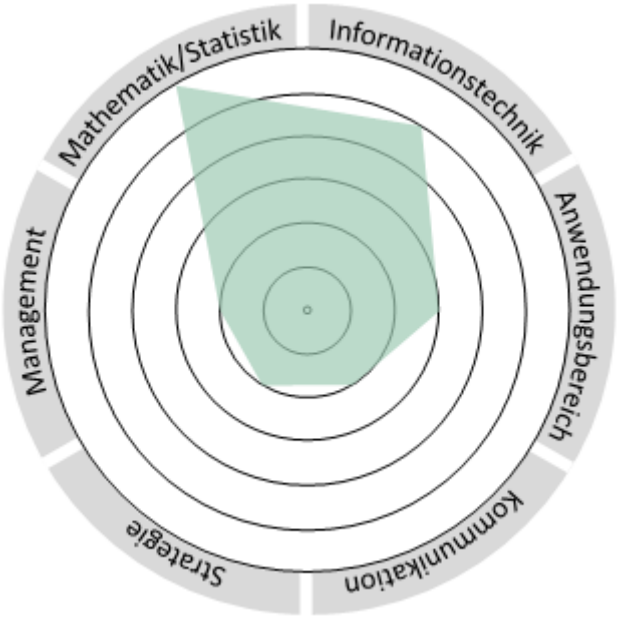


Abbildung 29: Kompetenzprofil des Bereichs „Entwicklung von Analyseverfahren“

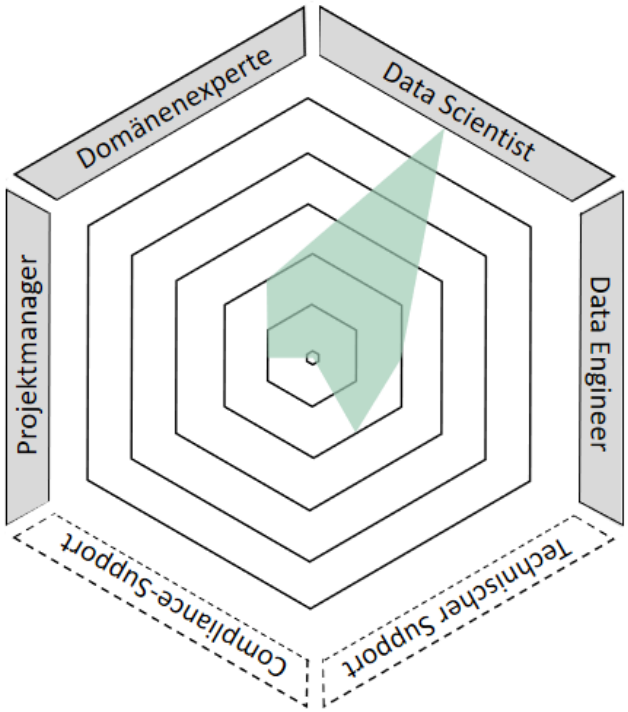


Abbildung 30: Rollen im Bereich „Entwicklung von Analyseverfahren“

7.7 Begleitende Aufgabe „Evaluation“

Die Evaluation ist im Schlüsselbereich Analyseverfahren eine vielfältige Aufgabe, da sie an drei Stellen ausgeführt wird: (1) bei der Auswahl potenziell für die Aufgabenstellung geeigneter Analyseverfahren, (2) bei der Entwicklung neuer Analyseverfahren und (3) bei der Anwendung des ausgewählten oder neuentwickelten Analyseverfahrens auf die konkrete Problemstellung. Ziel ist in allen drei Fällen eine nachvollziehbare Bewertung und Einordnung der Ergebnisse. Grundlage der Evaluation ist jeweils die Wahl einer geeigneten Metrik. Hierbei müssen neben technischen Metriken insbesondere auch die zentralen Kriterien der Anwendungsdomäne berücksichtigt werden, da nur diese Perspektive erlaubt, den tatsächlichen Wert der durchgeführten Analyse zu bestimmen.

In Tabelle 11 werden von Teilnehmerinnen und Teilnehmern häufig genannte Aufgaben der Evaluation aufgeführt und beschrieben.

Tabelle 11: Häufig genannte Aufgaben des Bereichs „Evaluation“

Aufgabe	Beschreibung
Bestimmung der Bewertungskriterien	Die Kriterien, nach denen die Evaluation vorgenommen wird, müssen domänenabhängig und im Hinblick auf das Projektziel gewählt werden.
Mehrwertschätzung	Der Nutzen, der durch die durchgeführte Analyse entstehen soll, muss im Vorfeld abgeschätzt werden. Dies kann nur im Kontext der domänenspezifischen Fragestellung geschehen. Die Mehrwertabschätzung setzt einen Rahmen für den vertretbaren Aufwand der Analysen.
Überprüfung der Umsetzbarkeit	Die Umsetzbarkeit der Analyse muss hinsichtlich der Erreichbarkeit des gesetzten Zieles, der Eignung der vorhandenen Daten und der Angemessenheit der verfügbaren Mittel beurteilt werden.
Benchmarking	Zur Beurteilung der späteren Ergebnisse muss ein geeigneter Vergleichsmaßstab (Benchmark) gewählt werden. Dies kann etwa ein bereits bestehendes Verfahren sein, das abgelöst werden soll, oder ein sehr einfaches Vergleichsverfahren, das mit wenig Aufwand nutzbar ist.
Aufwandsschätzung	Der Aufwand für die Durchführung der Analyseverfahren muss abgeschätzt werden. Der geschätzte Aufwand muss deutlich geringer sein als der Mehrwert, der von der Analyse erwartet wird.
Verfahrensvergleich	Die grundlegenden Merkmale der infrage kommenden Verfahren müssen herausgearbeitet und gegenübergestellt werden. Anschließend ist die Passung zwischen Verfahren und zu bearbeitender Problemstellung zu beurteilen.
Ergebnisevaluation	Die Ergebnisse der ausgeführten Analyse müssen beurteilt werden. Dies beinhaltet typischerweise eine Plausibilitätsprüfung, verschiedene statistische Auswertungen, die Validierung der Ergebnisse und eine Untersuchung der Robustheit des Verfahrens. Auch eine Überprüfung der Anwendbarkeit aus Domänensicht ist durchzuführen.
Performance-Tests	Soll das entwickelte Analyseverfahren später in den regulären Betrieb übernommen werden, ist die Performance des Verfahrens zu beurteilen (benötigte Hardware, Umfang der verarbeitbaren Datenmenge).

Die Ergebnisse der Evaluation müssen sorgfältig dokumentiert werden. Im Rahmen der Verfahrensauswahl gehören dazu vor allem die Gegenüberstellung von Vor- und Nachteilen der betrachteten Analyseverfahren sowie eine Beschreibung geeigneter Anwendungsfälle. Bei der Ergebnisevaluation zählen insbesondere die Darstellung der Bewertungskriterien und der Ausprägungen der Kriterien, die gewählte Vorgehensweise, das Test-Setup, Konfigurationstabellen, eine Aufstellung der untersuchten Parameterkombinationen und die konkreten Testergebnisse (inklusive der Angaben zur Ausführungsdauer) dazu. Auch sollten die während der Evaluation mit dem Verfahren gesammelten Erfahrungen und potenzielle Schwachstellen festgehalten werden. Schließlich sind die auf Basis der Evaluation getroffenen Entscheidungen nachvollziehbar und im Kontext der untersuchten Problemstellung zu begründen.

In Abbildung 31 ist das Kompetenzprofil von Personen dargestellt, die sich in dem Bereich *Evaluation* und den mit diesem direkt verbundenen Aufgaben spezialisieren. Da bei der Evaluation, wie oben beschrieben, drei unterschiedliche Aspekte untersucht werden, ist es denkbar, dass diese Untersuchungen auch von unterschiedlichen Personen durchgeführt werden. In diesem Fall müssen die beteiligten Personen nicht notwendigerweise in jeder Kompetenzdimension die unten gezeigten Maximalausprägungen besitzen. In Abbildung 32 sind in diesem Bereich beteiligte Rollen zu erkennen.

Abbildung 31: Kompetenzprofil des Bereichs „Evaluation“

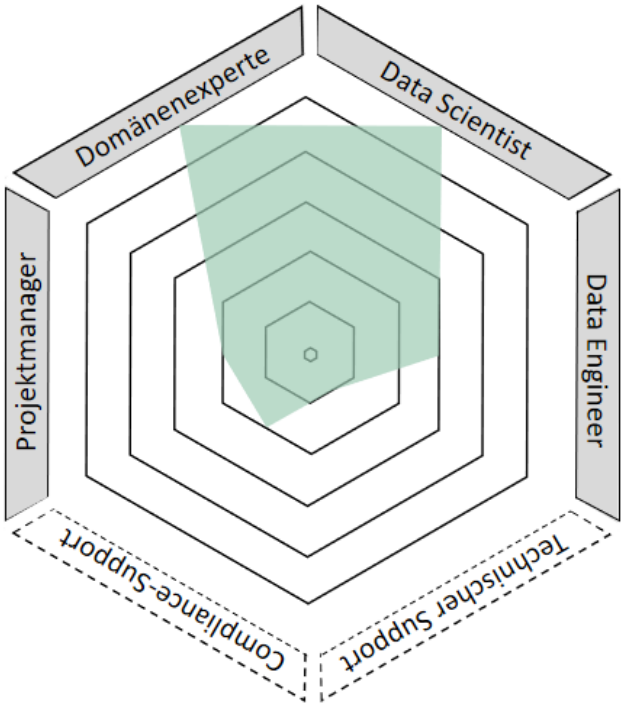


Abbildung 32: Rollen im Bereich „Evaluation“

7.8 Merkmale der „Analyseergebnisse“

Die Ergebnisse des Analyseprozesses können – je nach Fragestellung, Zielsetzung, verwendeten Methoden und vorhandener Datenbasis – sehr unterschiedliche Formen annehmen. Die Bandbreite reicht von deskriptiven und diagnostischen Analysen über prognostische und präskriptive Modelle bis zu sich selbst steuernden Systemen.

In Tabelle 12 werden von Teilnehmerinnen und Teilnehmern häufig genannte Merkmale der Analyseergebnisse aufgeführt und beschrieben.

Tabelle 12: Häufig genannte Merkmale des Bereichs „Analyseergebnisse“

Merkmal	Beschreibung
Aussagekraft	Welche Aussagen lassen sich aus dem Analyseergebnis ableiten? Handelt es sich eher um grobe Schätzungen oder um präzise Aussagen? Sind nur Aussagen über den Ist-Zustand möglich oder ist zu erwarten, dass die Ergebnisse auch in Zukunft gültig sein werden?
Darstellungsform	Wie werden die Analyseergebnisse vermittelt? Sind sie leicht verständlich beschrieben? Werden sie zur Erhöhung der Anschaulichkeit visualisiert? Werden die Ergebnisse detailliert dargestellt oder aggregiert?
Ergebnistyp	Welcher Art ist das Analyseergebnis (z. B. Beschreibung eines Zusammenhangs, Erklärung eines Zusammenhangs, Prognose zukünftigen Verhaltens, Ableitung einer Handlungsanweisung, Optimierung eines Systems)?
Generalisierbarkeit	Wie gut lassen sich die Ergebnisse auf unbekannte Daten übertragen?
Grenzen	Das entwickelte Analysemodell ist eine Abbildung der Realität und enthält notwendigerweise Vereinfachungen. Daher werden die Analyseergebnisse die Realität nur unvollkommen beschreiben. Welche Aussagegrenzen hat das entwickelte Modell? Welchen Grund haben diese Grenzen (z. B. geringe Datenmenge, fehlende Attribute, Beschränkungen des Analyseverfahrens)? Wie ließen sie sich gegebenenfalls überwinden?
Implementierbarkeit	Kann und soll das Analysemodell zu einer Software weiterentwickelt werden, welche die Analysefunktion dauerhaft und für neue Daten zur Verfügung stellt?
Komplexität	Wie einfach sind die Ergebnisse zu verstehen, und wie gut lassen sich Maßnahmen aus ihnen ableiten?
Neuartigkeit	Wurden Erkenntnisse gewonnen, die anders nicht zu Tage gekommen wären bzw. noch nicht vorhanden waren?
Quantitative Bewertung	Wie zuverlässig ist das erzielte Ergebnis? Welche quantitativen Bewertungsmaße (Signifikanzniveau, Fehlerrate usw.) liegen vor?
Relevanz	Tragen die Ergebnisse zur Lösung der ursprünglichen Problemstellung bei oder beantworten sie eine andere Frage? Auch statistisch signifikante Ergebnisse haben nicht notwendigerweise einen praktischen Nutzen. Sind die Ergebnisse trivial oder liefern sie neue Erkenntnisse? Lassen sich aus ihnen konkrete Handlungsvorschriften ableiten?
Transparenz	Ist der Entstehungsprozess der Analyseergebnisse transparent und nachvollziehbar?
Vergleichbarkeit	Lassen sich die Analyseergebnisse mit den Ergebnissen anderer, bereits bekannter Verfahren vergleichen?
Verständlichkeit	Sind die Ergebnisse aus sich selbst heraus verständlich? Werden Interpretationshilfen benötigt?
Vollständigkeit	Wie vollständig sind die vorliegenden Ergebnisse? Wurden nur Teilaspekte untersucht oder erfolgte eine umfangreiche Analyse? Ist die Notwendigkeit weiterer Analysen erkennbar?

8. SCHLÜSSELBEREICH NUTZBARMACHUNG

Der Schlüsselbereich *Nutzbarmachung* wird in die in Abbildung 33 dargestellten und in Verbindung stehenden Teilbereiche untergliedert, die in den folgenden Unterkapiteln einzeln betrachtet werden.

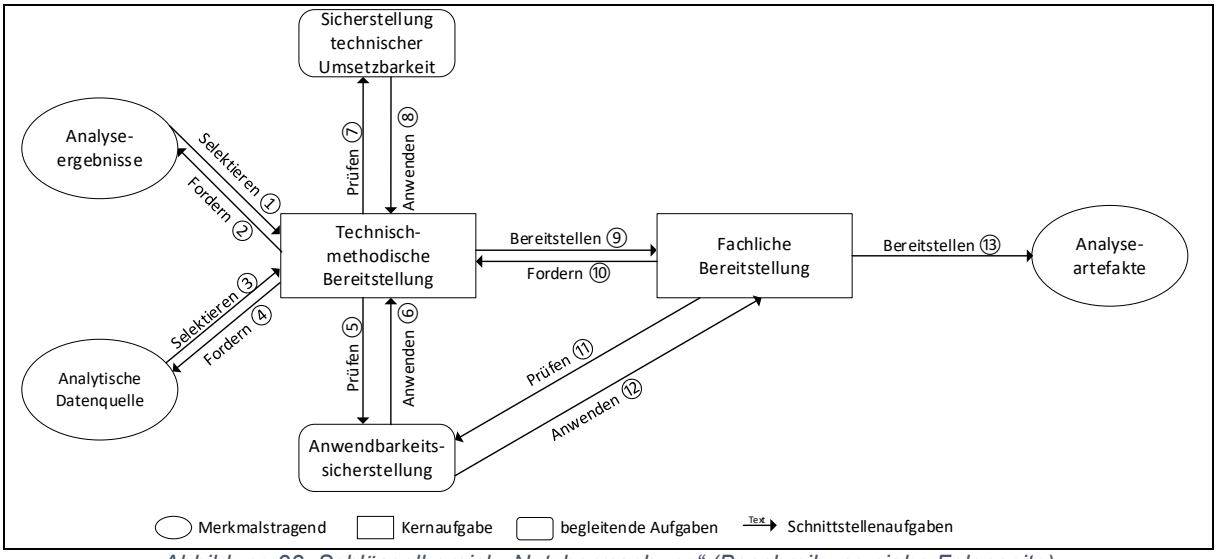


Abbildung 33: Schlüsselbereich „Nutzbarmachung“ (Beschreibung siehe Folgeseite)

①	Die Analyseergebnisse werden für die technisch-methodische Bereitstellung ausgewählt.
②	Sollten die Analyseergebnisse für die angedachte technisch-methodische Bereitstellung nicht geeignet sein, können Änderungen gefordert werden.
③	Wenn für die technisch-methodische Bereitstellung der Zugriff auf Daten benötigt wird, sind diese zu selektieren. Werden neue Daten verwendet, ist sicherzustellen, dass sie die gleichen Eigenschaften besitzen wie die Daten, mit denen das Modell entwickelt wurde.
④	Sollten die selektierten Daten für die angedachte technisch-methodische Bereitstellung nicht geeignet sein, können Änderungen gefordert werden.
⑤	Bei der technisch-methodischen Bereitstellung muss die Anwendbarkeit durch die Zielgruppe der Analyse geprüft werden.
⑥	Identifizierte Möglichkeiten zur Sicherstellung der Anwendbarkeit sind bei der technisch-methodischen Bereitstellung zu berücksichtigen.
⑦	Bei der technisch-methodischen Bereitstellung müssen technische Anforderungen geprüft werden.
⑧	Identifizierte technische Umsetzungsmöglichkeiten sind bei der technisch-methodischen Bereitstellung zu berücksichtigen.
⑨	Die Ergebnisse der technisch-methodischen Bereitstellung bilden die Grundlage zur fachlichen Bereitstellung der aufbereiteten Analyseergebnisse für die Anwenderzielgruppe. Die Auswirkungen können von einer Unterstützung bestehender Prozesse über eine Prozessanpassung bis hin zu einer kompletten Neuentwicklung von (nun ggf. automatisierten) Prozessen reichen.
⑩	Sollte die fachliche Bereitstellung nicht zu den gewünschten Ergebnissen führen, können Änderungen der technisch-methodischen Bereitstellung gefordert werden.
⑪	Bei der fachlichen Bereitstellung muss die Anwendbarkeit durch die Zielgruppe der Analyse geprüft werden.
⑫	Identifizierte Möglichkeiten zur Sicherstellung der Anwendbarkeit sind bei der fachlichen Bereitstellung zu berücksichtigen.
⑬	Durch die fachliche Bereitstellung entstehende Analyseartefakte müssen in geeigneter Form in die praktische Anwendung überführt werden.

In Abbildung 34 ist das Kompetenzprofil von Personen dargestellt, die sich im Schlüsselbereich *Nutzbarmachung* spezialisieren. Abbildung 35 stellt diejenigen Rollen dar, die bei der Betrachtung dieses Schlüsselbereichs eine Relevanz besitzen.

Abbildung 34: Kompetenzprofil des Schlüsselbereichs „Nutzbarmachung“

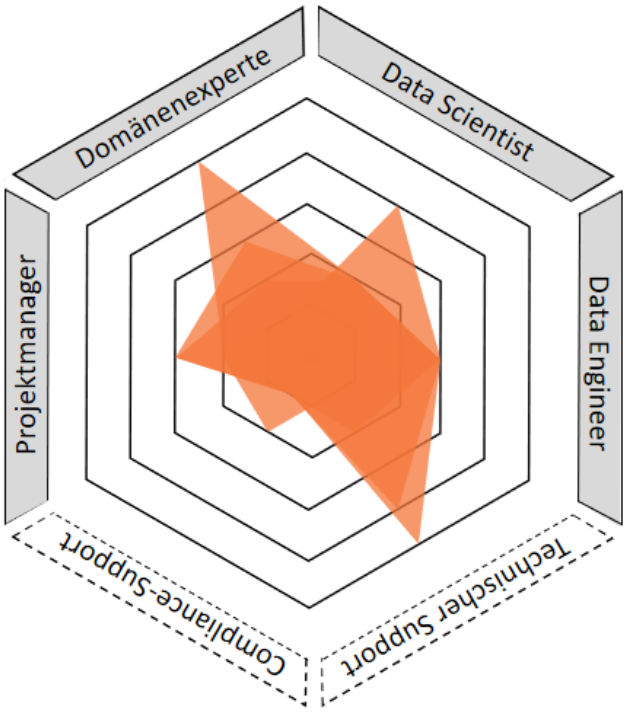


Abbildung 35: Rollen im Bereich „Nutzbarmachung“

8.1 Merkmal „Analyseergebnisse“

Die Nutzbarmachung fußt auf den Merkmalen der in Kapitel 7.8 beschriebenen Analyseergebnisse.

8.2 Merkmal „Analytische Datenquelle“

Für die Nutzbarmachung der Analyseergebnisse kann es nötig sein, auf die analytische Datenquelle zuzugreifen (vgl. Kapitel 6.5).

8.3 Aufgabe „Technisch-methodische Bereitstellung“

Die Ergebnisse der Analyse müssen für die Implementierung in einer geeigneten Form aufbereitet werden. Unterschieden werden können dabei:

- Eine manuelle Verwendung der Ergebnisse, bei der die Ergebnisse für die Zielgruppe aufbereitet und beispielsweise in Seminaren oder Workshops vermittelt werden
- Eine Umsetzung der Ergebnisse etwa in Form eines Berichtes, in dem die Ergebnisse einmalig aufbereitet werden
- Die Anwendung des trainierten Modells, um dieses auch auf unbekannte Daten anwenden zu können
- Kontinuierliches Lernen, bei dem sich das Modell durch wiederholte Anwendung auf unbekannte Daten selbsttätig anpassen kann
- Eine (ggf. nur organisationsinterne) Veröffentlichung des entwickelten Analyseverfahrens, um Dritten dessen Anwendung zu ermöglichen. So können Modellergebnisse unabhängig überprüft und Schwachstellen frühzeitig identifiziert werden

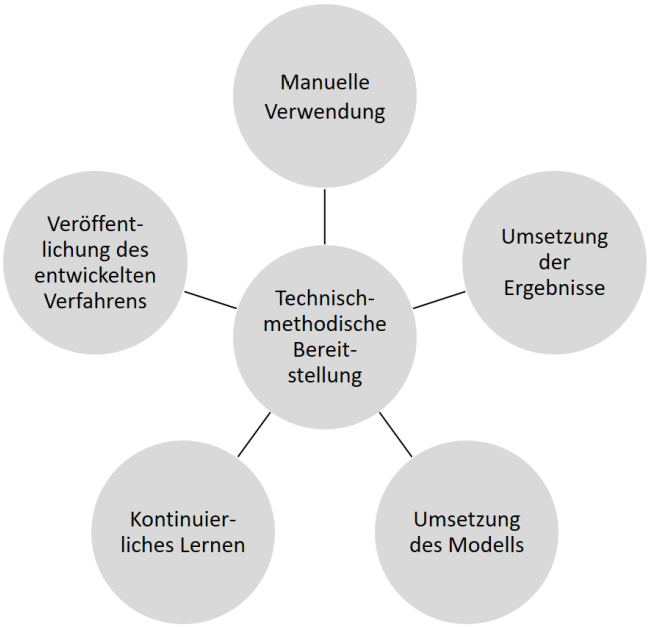


Abbildung 36: Formen der „Technisch-methodischen Bereitstellung“

Je nach Projekt ist auch die Auswahl mehrerer Implementierungsmöglichkeiten denkbar.

Das Modell muss in eine operative Produktivumgebung eingebettet werden. Einmalige Ergebnisse sind in Ausnahmefällen (z. B. für eine Pilotstudie) relevant, ansonsten wird der Wert der Modelle i. d. R. dadurch geschöpft, dass sie kontinuierlich oder on demand in eine Produktivumgebung eingebettet werden.

In Tabelle 13 werden von Teilnehmerinnen und Teilnehmern häufig genannte Aufgaben der technisch-methodischen Bereitstellung aufgeführt und beschrieben.

Tabelle 13: Häufig genannte Aufgaben des Bereichs „Technisch-methodische Bereitstellung“

Aufgabe	Beschreibung
Adressatengerechte Aufbereitung der Ergebnisse	Geeignete technisch-methodische Aufbereitung und Möglichkeit der Interpretation durch die Anwender
Aufbau der Produktivumgebung	Gegebenenfalls kann es nötig sein, eine neue Infrastruktur aufzubauen, in der die Ergebnisse laufend aktualisiert und berücksichtigt werden können.
Transfer der Ergebnisse	Für den laufenden Betrieb kann es nötig sein, die Ergebnisse aus der Analyseumgebung in ein operatives System zu transferieren.
Kontextschaffung	Die Art und Weise und der Zeitraum der Gewinnung der Ergebnisse sollten ersichtlich sein.
Automatisierung von Prozessen	Berücksichtigung allgemeiner Herausforderungen bei der Automatisierung von Prozessen, z. B.: • Was passiert im Fehlerfall? • Wie ist mit Medienbrüchen umzugehen, können sie vermieden oder kompensiert werden? • Wie kann die Ausführung in geeigneter Form protokolliert werden?
Umgang mit IT-Ressourcen	Eine effiziente Nutzung von IT-Ressourcen ist sicherzustellen.
Technischer Test des aufgesetzten Systems	Die technisch fehlerfreie Arbeitsweise des Analyse-Systems muss überprüft werden, insbesondere, wenn es in die Produktivumgebung der Organisation integriert und an reale Datenquellen angeschlossen wurde.

In Abbildung 37 ist das Kompetenzprofil von Personen dargestellt, die sich in dem Bereich *Technisch-methodische Bereitstellung* und den mit diesem direkt verbundenen Aufgaben spezialisieren, in Abbildung 38 sind beteiligte Rollen zu erkennen.

Abbildung 37: Kompetenzprofil des Bereichs „Technisch-methodische Bereitstellung“

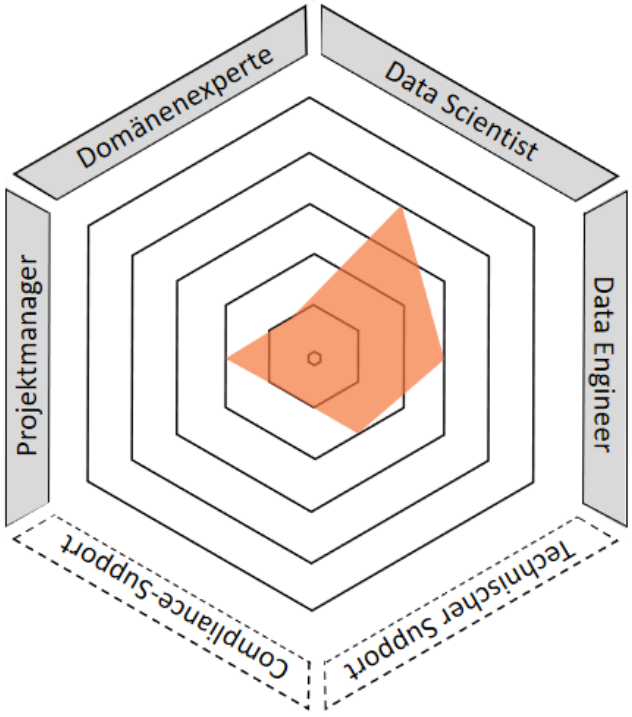


Abbildung 38: Rollen im Bereich „Technisch-methodische Bereitstellung“

8.4 Begleitende Aufgabe „Sicherstellung technischer Umsetzbarkeit“

In der Regel sollte die technisch-methodische Bereitstellung eine vollständige Automatisierung der Verfahren bedeuten. In einigen Fällen kann es aber auch sinnvoll oder notwendig sein, manuelle Schritte miteinzubeziehen. Die begleitende Aufgabe *Sicherstellung technischer Umsetzbarkeit* soll die initiale Einrichtung und den dauerhaften Betrieb der Analyseanwendung unter den definierten wirtschaftlichen Rahmenbedingungen gewährleisten. Dazu gehört auch die Sicherstellung der (technischen) Bedienbarkeit der Anwendung, der Durchführung von Wartungsarbeiten und der Umsetzung von technischen Anpassungen.

In Tabelle 14 werden von Teilnehmerinnen und Teilnehmern häufig genannte Aufgaben innerhalb des Bereichs *Sicherstellung technischer Umsetzbarkeit* aufgeführt und beschrieben.

Tabelle 14: Häufig genannte Aufgaben des Bereichs „Sicherstellung technischer Umsetzbarkeit“

Aufgabe	Beschreibung
Berücksichtigung von Zeitkritikalitäten	Muss die Analyse in Echtzeit durchgeführt werden oder handelt es sich um eine nicht zeitkritische Analyse, die z. B. über Nacht im Batchbetrieb durchgeführt werden kann?
Berücksichtigung von Laufzeiten	Wie rechenaufwendig ist der Algorithmus? Skaliert er z. B. gut mit der Datenmenge?
Umgang mit den angebundenen Datenquellen	Wie kann auf Änderungen bei den Datenquellen (Formate, Qualität, Rechte usw.) reagiert werden? Wer ist zuständig? Wie ist der Informationsfluss?
Identifikation des Hardware-Stacks	Welche Hardware wird zum Betrieb der Analyselösung benötigt? Welche Realisierungsform (on premise, private Cloud, Cloud, IaaS, PaaS, SaaS usw.) ist geeignet?
Identifikation des Software-Stacks	Ist der zu verwendende Software-Stack von der Organisation bereits vorgegeben oder muss er als Teil des Projekts noch evaluiert werden? Auch die Kompetenzen der beteiligten Personengruppen sind hier zu berücksichtigen.
Identifikation technischer Möglichkeiten und Gegebenheiten	Eine Berücksichtigung der gegebenen IT-Infrastruktur bzw. der Möglichkeit einer Beschaffung ist zu prüfen.
Prüfung von Software-Lizenzen	Werden für das Produktivsystem weitere oder zusätzliche Lizenzen benötigt?
Rechtliche Rahmenbedingungen	Wurden die rechtlichen Rahmenbedingungen für die Nutzung der Analyseanwendung (Datenschutz, Compliance usw.) geklärt, definiert und dokumentiert?
Zugriffskonzept erstellen	Ist es möglich, den Zugriff auf Analyseergebnisse auf berechtigte Anwendergruppen einzuschränken? Wurden Vorkehrungen getroffen, um die Sicherheit aller Daten zu gewährleisten?
Betrieb und Support sicherstellen	Wer ist für den Produktivbetrieb der Analyseanwendung zuständig? Wer kann bei technischen/methodischen Fragen und Problemen unterstützen?
Automatisierung	Wie weit können die Auswertung der Daten und die Integration der Ergebnisse automatisiert werden? In welchen Zeitintervallen werden die Analysen wiederholt?

In Abbildung 39 ist das Kompetenzprofil von Personen dargestellt, die sich in dem Bereich *Technisch-methodische Bereitstellung* und den mit diesem direkt verbundenen Aufgaben spezialisieren, in Abbildung 40 sind beteiligte Rollen zu erkennen.

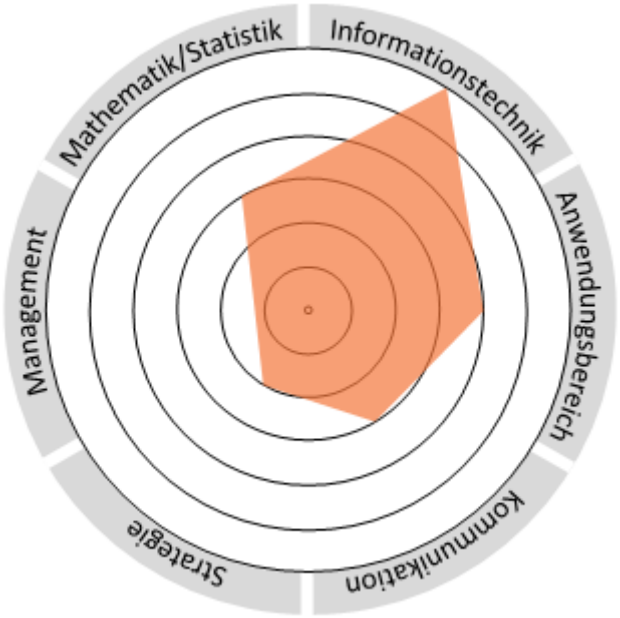


Abbildung 39: Kompetenzprofil des Bereichs „Sicherstellung technischer Umsetzbarkeit“

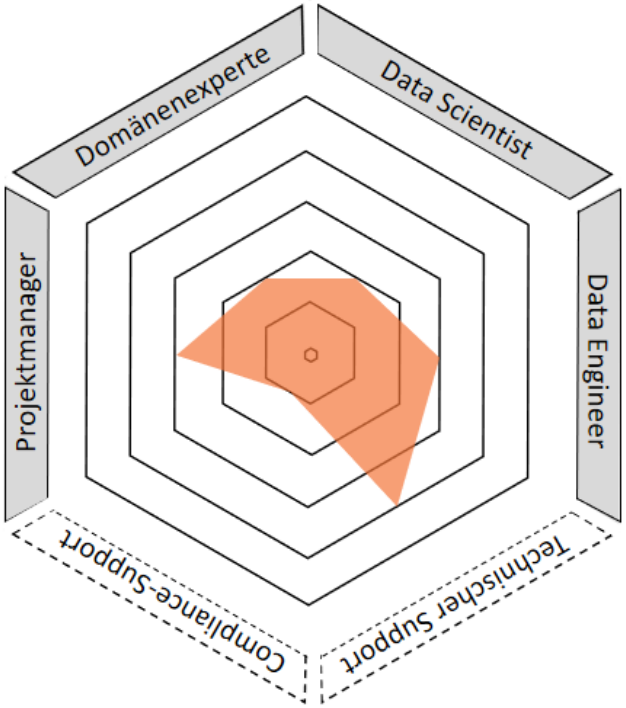


Abbildung 40: Rollen im Bereich „Sicherstellung technischer Umsetzbarkeit“

8.5 Begleitende Aufgabe „Anwendbarkeitssicherstellung“

Die Analyseergebnisse müssen in einer Form vorliegen, die von der Zielgruppe genutzt werden kann bzw. der Zielgruppe zu vermitteln ist. Die Anwendbarkeitssicherstellung sollte im Zusammenspiel von Personen mit methodischen Fachkenntnissen und Personen aus der Domäne erfolgen.

In Tabelle 15 werden von Teilnehmerinnen und Teilnehmern häufig genannte Aufgaben innerhalb der Anwendbarkeitssicherstellung beschrieben.

Tabelle 15: Häufig genannte Aufgaben des Bereichs „Anwendbarkeitssicherstellung“

Aufgabe	Beschreibung
Adressaten identifizieren	Um eine Anwendbarkeit sicherzustellen, müssen die Adressaten der Analyse bekannt sein.
UI/UX-Design festlegen	Die Oberfläche sollte für alle Benutzergruppen einfach zu verstehen und zu nutzen sein, aber trotzdem Flexibilität bieten und die Komplexität des Themas abdecken. Analyseergebnisse sollten verständlich aufbereitet werden, bspw. durch Visualisierungen.
Zugriff sicherstellen	Berechtigungsstrukturen und Zugänge sind zu definieren. Die Gewährleistung der Umsetzbarkeit ist Teil der begleitenden Aufgabe *Sicherstellung technischer Umsetzbarkeit.*
Anwender beteiligen	Im Vorfeld des Einsatzes der Analyseergebnisse können z. B. Workshops abgehalten werden, um Feedback zur Sicherstellung der Anwendbarkeit einzuholen.
Dokumentationskonzept erstellen	Neben einer technisch-methodischen Dokumentation sind auch geeignete Anwenderdokumentationen zu planen, bspw. als Interpretationshilfe oder zur Beschreibung verwendeter Kennzahlen.
Schulungskonzept erstellen	Abhängig vom Umfang der entwickelten Analyseartefakte und von der Form der Nutzbarmachung ist ein geeignetes Schulungskonzept zu konzipieren.

In Abbildung 41 ist das Kompetenzprofil von Personen dargestellt, die sich in dem Bereich *Anwendbarkeitssicherstellung* und den mit ihm direkt verbundenen Aufgaben spezialisieren, in Abbildung 42 sind beteiligte Rollen zu erkennen.

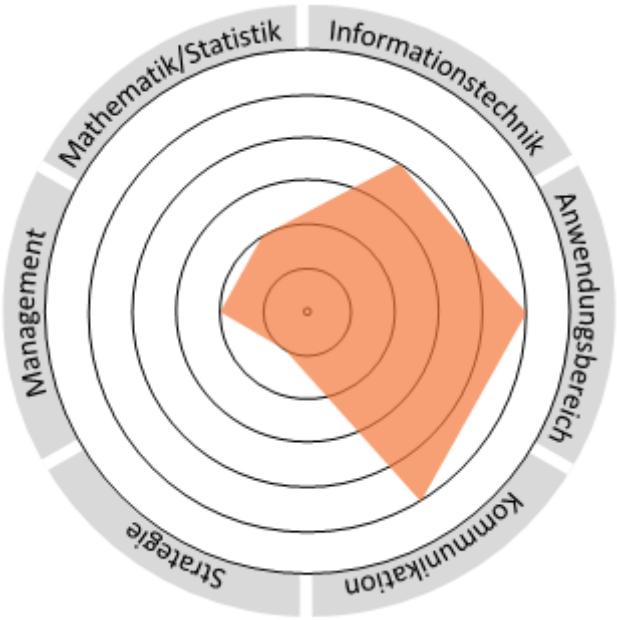


Abbildung 41: Kompetenzprofil des Bereichs „Anwendbarkeitssicherstellung“

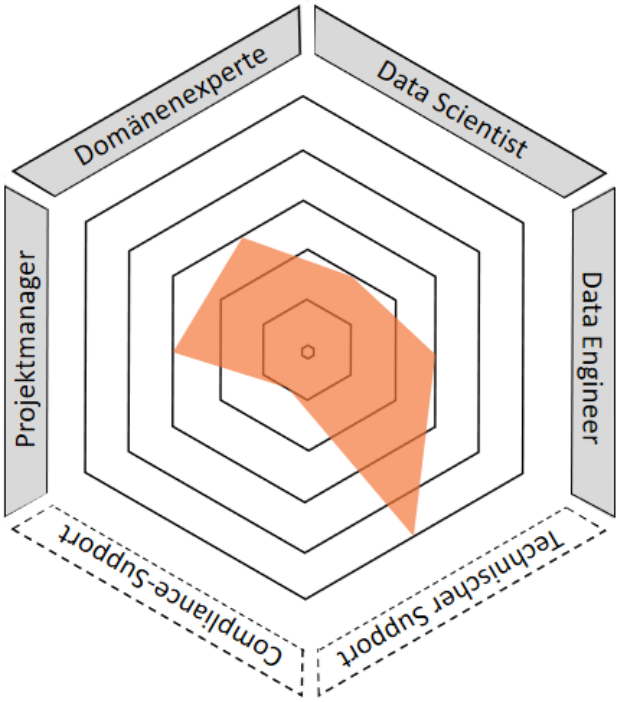


Abbildung 42: Rollen im Bereich „Anwendbarkeitssicherstellung“

8.6 Aufgabe „Fachliche Bereitstellung“

Die Aufgaben der fachlichen Bereitstellung hängen sehr stark von der Bereitstellungsform und der Domäne ab, in der das Projekt durchgeführt wird. Dargestellt werden daher ausschließlich allgemeingültige Aufgaben.

In Tabelle 16 werden von Teilnehmerinnen und Teilnehmern häufig genannte Aufgaben der fachlichen Bereitstellung beschrieben.

Tabelle 16: Häufig genannte Aufgaben des Bereichs „Fachliche Bereitstellung“

Aufgabe	Beschreibung
Sicherstellung der Nachhaltigkeit	Nachhaltigkeit bedeutet die Sicherstellung einer dauerhaften Nutzung bzw. Relevanz.
Berücksichtigung von Reichweite und Auswirkungen	Bevor die Ergebnisse jenseits des Projektteams veröffentlicht werden, sollten ihre möglichen Auswirkungen unter anderem unter moralischen und wirtschaftlichen Gesichtspunkten eingeschätzt werden.
Berücksichtigung rechtlicher Fragestellungen	Der Datenschutz und rechtliche Fragestellungen sind einzuschätzen, bevor die Analyseergebnisse verwendet werden.
Ansprechpartner festlegen	Es muss fachliche Ansprechpartner für Fragen während der laufenden Nutzung geben. Eine definierte Möglichkeit, Kontakt aufzunehmen, ist dabei ebenfalls festzulegen.
Integration in bestehende Prozesse	Eine fachliche Integration der Analyseartefakte in bestehende Prozesse ist notwendig.
Internes Kostenverrechnungsmodell	Für den Betrieb der Analyseartefakte sind Personal- und IT-Kosten zu ermitteln und ggf. auf die Anwender zu verteilen.
Schulung durchführen	Die im Zuge der Anwendbarkeitssicherstellung konzipierten Schulungen sind in geeigneter Form durchzuführen (Präsenzschulungen, Online-Schulungen, Webinare etc.).
Benutzerhandbuch erstellen	Die im Zuge der Anwendbarkeitssicherstellung konzipierte Benutzerdokumentation ist zu erstellen.
Problembehandlung festlegen	Es müssen Prüfmechanismen und Verhaltensweisen für den Fall festgelegt werden, dass das Analyseartefakt keine sinnvollen Ergebnisse (mehr) liefert.

In Abbildung 43 ist das Kompetenzprofil von Personen dargestellt, die sich in dem Bereich *Fachliche Bereitstellung* und den mit diesem direkt verbundenen Aufgaben spezialisieren, in Abbildung 44 sind beteiligte Rollen zu erkennen.

Abbildung 43: Kompetenzprofil des Bereichs „Fachliche Bereitstellung“

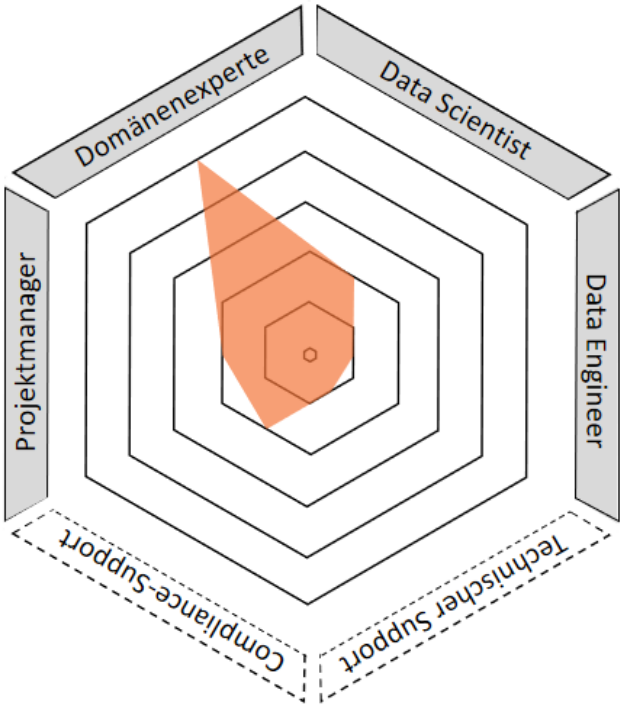


Abbildung 44: Rollen im Bereich „Fachliche Bereitstellung“

8.7 Merkmal „Analyseartefakte“

Die Merkmale der Analyseartefakte sind abhängig von der Form der Ergebnisbereitstellung (vgl. Kapitel 8.1). In Tabelle 17 werden von Teilnehmerinnen und Teilnehmern häufig genannte Merkmale von Analyseartefakten beschrieben.

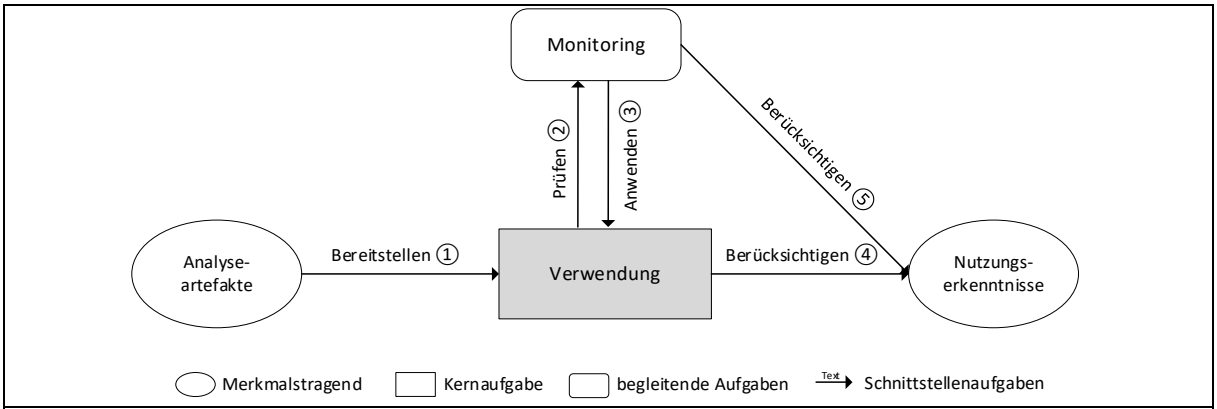
Tabelle 17: Häufig genannte Merkmale des Bereichs Analyseartefakte

Merkmal	Beschreibung
Benutzerdokumentation	Den Nutzern des Analysesystems muss ein Benutzerleitfaden/Benutzerhandbuch zur Verfügung gestellt werden, in dem die vorhandenen Berichte, Dashboards, Datenbanken etc. inklusive ihrer Zugriffsrechte beschrieben sind. Ferner sind fachliche Ansprechpartner zu benennen.
Technische Dokumentation	Zur Wartung und Weiterentwicklung des Analysesystems muss eine detaillierte Beschreibung der eingesetzten/entwickelten Software (Code-Basis, Ein- und Ausgabe, ausgeführte Zwischenschritte, Abhängigkeiten von anderen Komponenten) vorliegen. Außerdem ist die technische Infrastruktur, die für das Analysesystem geschaffen wurde bzw. in die es eingebettet ist, zu dokumentieren. Auch hier sind technische Ansprechpartner zu benennen.
Modelldokumentation	Zur Anpassung und künftigen Weiterentwicklung der Analysemodelle müssen diese detailliert beschrieben sein (inklusive der Prämissen für den Modelleinsatz).
Handlungsempfehlungen	Zumindest im Fall einer manuellen Verwendung von Ergebnissen sind Handlungsempfehlungen für die Empfänger der Analyseartefakte zu definieren.
Modelle	Die aus der Analyse heraus entstehenden Modelle können auf neue Daten angewendet werden.
Berichte	Die aus der Analyse hervorgehenden Daten sind zielgruppengerecht in Form von Berichten darzustellen.
Analyseinfrastruktur	Häufig muss zur dauerhaften Nutzung der Analysemodelle eine spezifische Analyseinfrastruktur bereitgestellt werden, die selbst wiederum in die IT-Infrastruktur der Organisation eingebettet ist.
Support	Es wird ein definierter fachlicher und technischer Support, sowohl zur Betreuung des Betriebes als auch zur Behebung von Problemfällen, benötigt.

9. SCHLÜSSELBEREICH *NUTZUNG*

Der Schlüsselbereich *Nutzung* wird in die in Abbildung 45 dargestellten und in Verbindung stehenden Teilbereiche untergliedert, die in den folgenden Unterkapiteln einzeln betrachtet werden.

Die Aufgabe *Verwendung* selbst gehört nicht zum Tätigkeitsbereich eines Data Scientists, vielmehr liegt sie im Bereich der Domänenexperten. Zudem existiert ein so breites Spektrum potenzieller Verwendungen, dass eine vollständige Darstellung nicht möglich ist und hier deshalb auf sie verzichtet wird. Trotz der genannten Einschränkungen ist die Verwendung ein relevanter Aspekt eines Data-Science-Projekts. Zum einen ist die kontinuierliche Überwachung des Betriebs der Analyseartefakte notwendig, um sicherzustellen, dass die Analyseergebnisse ihre Gültigkeit behalten. Zum anderen werden bei großen Projekten die Analyseartefakte häufig in mehreren Schritten zur Verfügung gestellt, wobei bei einem Folgeschritt die Erfahrungen aus der Verwendung der vorhergehenden Schritte berücksichtigt werden.



① Die Analyseartefakte werden zur Verwendung bereitgestellt.
② Die Verwendung muss wiederkehrend überwacht werden.
③ Ergebnisse der Überwachung sind bei der Verwendung zu berücksichtigen.
④ Domänenspezifische Erkenntnisse aus der Verwendung sind bei Weiter- und Neuentwicklungen zu berücksichtigen.
⑤ Datenwissenschaftliche Erkenntnisse aus der Verwendung sind bei Weiter- und Neuentwicklungen zu berücksichtigen.

Abbildung 45: Schlüsselbereich Nutzung

In Abbildung 46 ist das Kompetenzprofil von Personen dargestellt, die sich im Schlüsselbereich *Nutzung* spezialisieren, Abbildung 47 stellt diejenigen Rollen dar, die bei der Betrachtung dieses Schlüsselbereichs eine Relevanz besitzen.

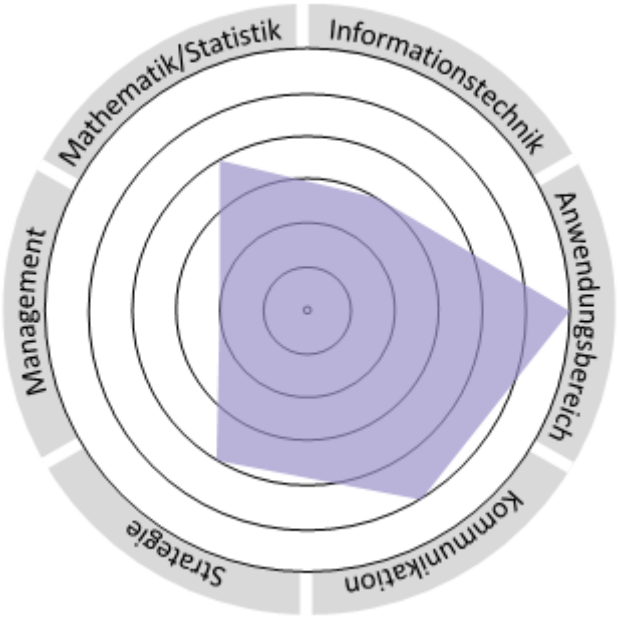


Abbildung 46: Kompetenzprofil des Bereichs „Monitoring“

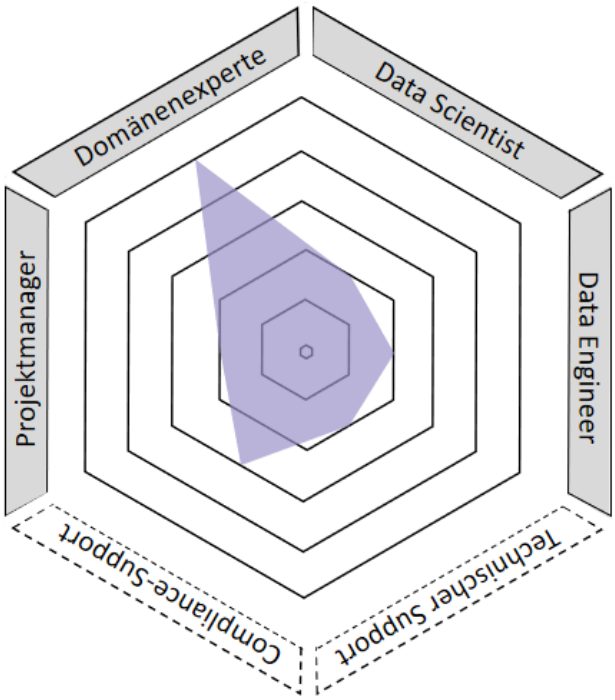


Abbildung 47: Rollen im Bereich „Monitoring“

9.1 Merkmal „Analyseartefakte“

Die Nutzung der Analyseartefakte fußt auf deren Merkmalen (vgl. Kapitel 8.7).

9.2 Begleitende Aufgabe „Monitoring“

Innerhalb des Monitorings muss der Regelbetrieb, für den das Analyseartefakt langfristig ausgelegt ist, überwacht werden. Dabei ist insbesondere die Qualität der Analyseergebnisse kontinuierlich zu überprüfen und die ständige Anwendbarkeit des Modells zu verifizieren.

In Tabelle 18 werden von Teilnehmerinnen und Teilnehmern häufig genannte Aufgaben beim Monitoring der Verwendung aufgeführt und beschrieben.

Tabelle 18: Häufig genannte Aufgaben des Bereichs „Monitoring“

Aufgabe	**Beschreibung**
Analyseartefakte allgemein	
Sicherstellung der korrekten Anwendungsdomäne	Die Analyseartefakte sind für eine bestimmte Domäne erstellt worden. Diese Spezialisierung muss gewahrt werden.
Bewertung der Analyseartefakte	Die Ergebnisse der Analyse sollten wiederkehrend hinsichtlich ihrer Aussagekraft und Vorhersagegüte bewertet werden.
Nachhaltigkeit der Analyseartefakte prüfen	Es ist zu prüfen, ob die Analyseartefakte gepflegt werden und wie schnell Ergebnisse veralten.
Anwendung von Analyseartefakten	
Prüfung der Daten	Das Modell wird möglicherweise auf Daten angewendet, die zum Zeitpunkt der Erstellung noch nicht existieren. Es ist so weit wie möglich sicherzustellen, dass die Anwendung korrekte Ergebnisse liefert. Dies sollte sowohl von Daten- als auch von Domänenexperten verifiziert werden.
Überwachung von Fehlern	Fehlerberichte müssen gesammelt und ausgewertet werden, darunter fallen u. a. das unerwartete Verhalten von Modellen oder neue Formen von Datenfehlern.
Metadaten zur Anwendung	
Erkennen von Performance-Herausforderungen	Die Identifikation von Performance-Herausforderungen bei der Nutzbarmachung ist limitiert. Daher sollte dieser Aspekt auch bei der Verwendung überwacht werden.
Auswerten von Nutzungsdaten	Es ist zu prüfen, ob die Analyseartefakte weiterhin verwendet werden sollen. Dafür müssen die Nutzungsdaten aufgezeichnet werden.

Als Artefakt dieses Aufgabenbereichs entsteht ein Evaluationsbericht, der eine Bewertung der Nützlichkeit von Analyseartefakten ermöglicht.

Da der Schlüsselbereich Nutzung mit dem Monitoring nur eine Teilaufgabe beinhaltet, die in der Verantwortung eines Data Scientists liegt, ist die Abbildung 46 für das Kompetenzprofil und die Abbildung 47 für die beteiligten Rollen zu Rate zu ziehen.

9.3 Merkmal „Nutzungserkenntnisse“

Auf Basis der Nutzungserkenntnisse kann entschieden werden, ob die Nutzung von Analyseartefakten eingestellt werden sollte oder ob Letztere zu überarbeiten sind. Das kann entweder auf Grund veränderter Gegebenheiten nötig werden oder weil sich die erarbeitete Lösung im produktiven Einsatz nicht bewährt hat.

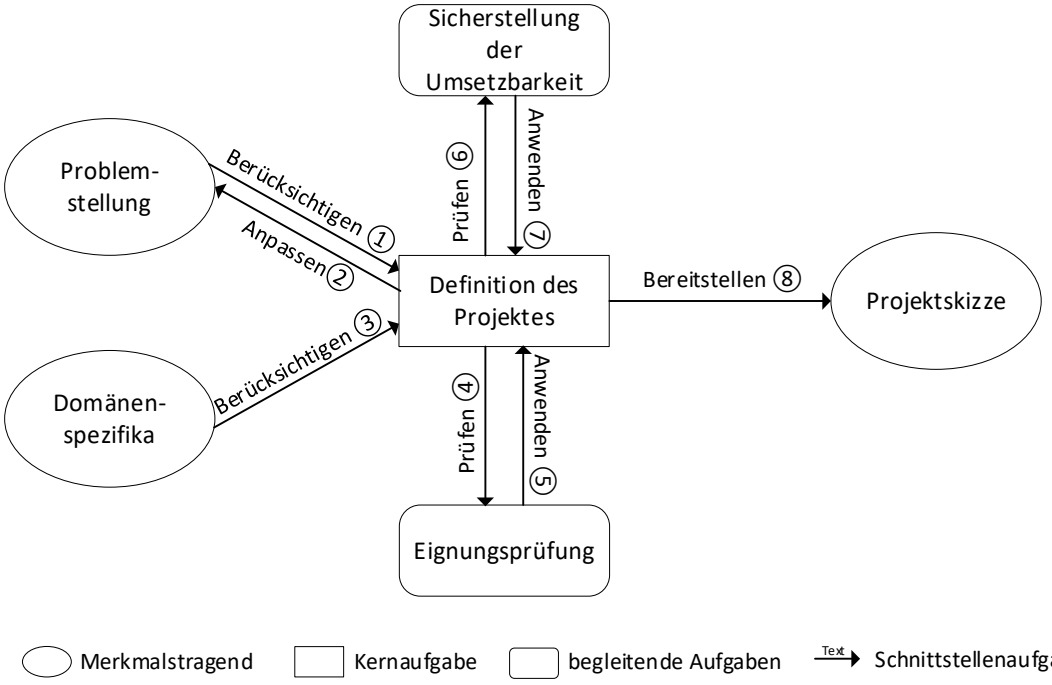
In Tabelle 19 werden von Teilnehmerinnen und Teilnehmern häufig genannte Merkmale beschrieben, nach denen Nutzungserkenntnisse untergliedert werden können.

Tabelle 19: Häufig genannte Merkmale des Bereichs „Nutzungserkenntnisse“

Merkmal	**Beschreibung**
Fehlerberichte	Die Berichte ermöglichen eine Bewertung dahingehend, ob die Analyseartefakte ausreichend stabil betrieben werden können.
Nutzungshäufigkeit	Werden Analyseartefakte von den Domänenexperten nicht genutzt, kann der Betrieb unnötig sein und ggf. auf Weiterentwicklungen verzichtet werden.
Performance der Analyseartefakte	Eine Betrachtung der Performance ermöglicht eine Bewertung der Eignung der verwendeten technischen Infrastruktur.
Nutzungsart	Die Art der Nutzung kann eine mögliche Weiterentwicklung aus Domänenperspektive beeinflussen.

10. SCHLÜSSELBEREICH DOMÄNE

Der Schlüsselbereich *Domäne* beinhaltet Aufgaben, die am Beginn eines Data-Science-Vorhabens stehen – dargestellt in Abbildung 48. Die abgebildeten und in Verbindung stehenden Teilbereiche werden in den folgenden Unterkapiteln betrachtet.



① Auf Basis einer in der Domäne identifizierten Problemstellung kann die Definition eines Projekts vorgenommen werden.

② Sollten bei der Definition eines Projekts Hindernisse identifiziert werden, die eine geeignete Umsetzung fraglich machen, ist die Anpassung der Problemstellung zu prüfen.

③ Die Spezifika der Domäne, in der das Projekt durchgeführt wird, sind bei der Definition zu berücksichtigen.

④ Die Eignung des Problems für ein Data-Science-Projekt ist zu prüfen.

⑤ Die Ergebnisse der Eignungsprüfung sind bei der Definition des Projekts zu berücksichtigen.

⑥ Bei der Definition des Projekts ist immer wieder die Umsetzbarkeit sicherzustellen.

⑦ Werden besondere Anforderungen identifiziert, die die Projektumsetzbarkeit sicherstellen können, sind diese umzusetzen.

⑧ Eine Projektskizze ist als Ergebnis dieses Teilprozesses bereitzustellen.

Abbildung 48: Schlüsselbereich „Domäne“

In Abbildung 49 ist das Kompetenzprofil von Personen dargestellt, die sich im Schlüsselbereich *Domäne* und dabei auf Aufgaben, die zu Beginn eines Data-Science-Vorhabens bearbeitet werden müssen, spezialisieren. Abbildung 50 stellt nach demselben Prinzip diejenigen Rollen dar, die bei der Betrachtung dieses Schlüsselbereichs eine Relevanz besitzen.

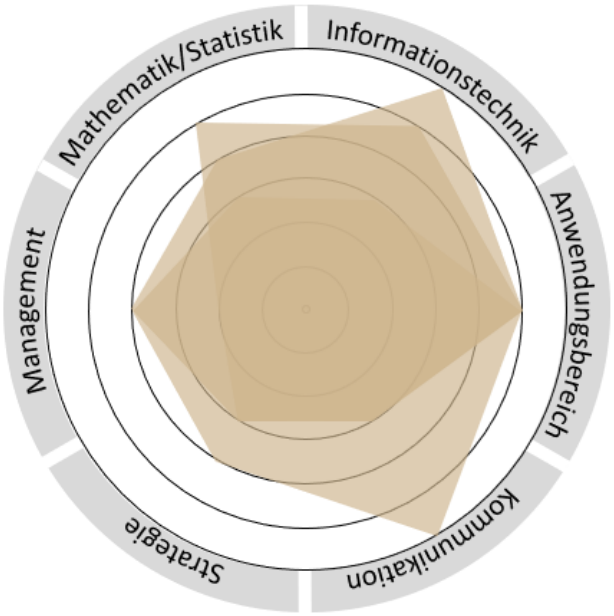


Abbildung 49: Kompetenzprofil des Schlüsselbereichs „Domäne“

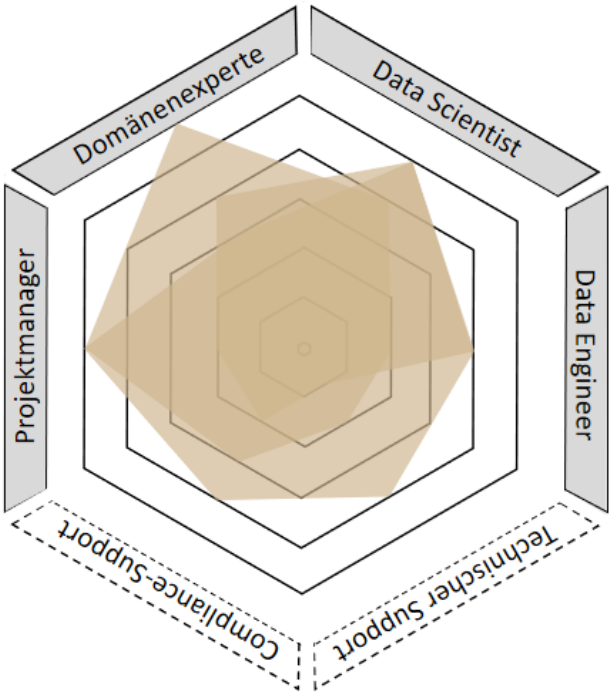


Abbildung 50: Rollen im Bereich „Domäne“

Neben den zuvor adressierten expliziten Aufgaben beeinflusst die Domäne alle anderen Schlüsselbereiche, allerdings in unterschiedlichem Ausmaß. Nachfolgend werden die Teilbereiche nach Schlüsselbereich dargestellt, die von den Teilnehmerinnen und Teilnehmern am häufigsten als relevant in Bezug auf die Domäne genannt wurden.

Schlüsselbereich Daten

- *Ursprungsdatenquellen*: Die Datenrelevanz ist nur domänenspezifisch zu bewerten, ein Datenverständnis kann ebenfalls nur unter Berücksichtigung der Domäne aufgebaut werden.
- *Datenaufbereitung*: Datenstandards innerhalb der Domäne (z. B. Datenschutz) sind zu berücksichtigen. Transformationen (z. B. Vergleichbarmachung von Messwerten, Identifikation von Datenfehlern oder von Ausreißern, die zwar weit vom Kern der Verteilung der Daten entfernt liegen, aber richtige Werte darstellen und keine Messfehler sind) sind ebenfalls im Domänenkontext zu sehen.
- *Explorative Datenanalyse*: Die Domänenspezifik wird über die Problemstellung in die explorative Datenanalyse eingebracht. Ziel ist die Schaffung eines Mehrwertes bezogen auf die Domäne.

Schlüsselbereich Analyseverfahren

- *Anforderungen an die Analyseverfahren*: Domänenspezifische Rahmenbedingungen (z. B. rechtliche oder regulatorische) schließen ggf. ganze Kategorien von Verfahren für die Verwendung aus. Darüber hinaus gibt es oft wünschenswerte, aber nicht zwingend erforderliche Anforderungen, die für die Domänenexperten in der Anwendung von Bedeutung sind. So gibt es z. B. viele Bereiche, in denen erklärbare Modelle wünschenswert sind oder kausale Abhängigkeiten berücksichtigt werden müssen.
- *Identifikation geeigneter Analyseverfahren*: In vielen Domänen existieren häufig verwendete Analyseverfahren, die als Vergleichsmaßstab herangezogen werden können. Zudem beeinflusst die Form der gewünschten Analyseergebnisse die Auswahl.
- *Evaluation*: Ergebnisse müssen mit Hintergrundwissen in den Domänenkontext eingeordnet und in einer für Domänenexperten geeigneten Form dargestellt werden. Je nach Anwendung ergeben sich weitergehende Anforderungen an die Evaluation bzw. werden die relevanten Metriken zur Evaluation aus der Domäne heraus definiert.

Schlüsselbereich Nutzbarmachung

- *Anwendbarkeitssicherstellung*: Der Domänenhintergrund der Anwender ist zu berücksichtigen.
- *Fachliche Bereitstellung*: Analyseergebnisse werden im Domänenkontext angewandt.
- *Technisch-methodische Bereitstellung*: Die Rahmenbedingungen der Nutzbarmachung müssen berücksichtigt werden.
- *Sicherstellung technischer Umsetzbarkeit*: Existierende nichtfunktionale Anforderungen der Domäne müssen bekannt sein.

10.1 Merkmal „*Problemstellung*“

Am Beginn von Data-Science-Projekten steht eine Problemstellung, die der Domäne entstammt. In Tabelle 20 werden von Teilnehmerinnen und Teilnehmern häufig genannte Merkmale von Problemstellungen aufgeführt und beschrieben.

Tabelle 20: Beschreibung der Merkmale des Bereichs „Problemstellung“

Merkmal	Beschreibung
Ziel	Es ist zu entscheiden, wie die Ergebnisse des Projekts später verwendet werden sollen, z. B. ob es sich bei dem Projektziel um einen Erkenntnisgewinn oder den produktiven Betrieb von Modellen handelt.
Fachlicher Zweck	Durch die Definition des fachlichen Zwecks der Lösung kann der Projektrahmen festgelegt werden. Weiterhin ist es möglich, die Relevanz einer Problemlösung festzustellen.
Anforderungen	Es ist zu beschreiben, welche Anforderungen zu erarbeitende Lösungen erfüllen müssen.
Messbarkeit	An welchen Kriterien ist festzumachen, wie die Projektergebnisse letztendlich zu bewerten sind?
Daten	Datenquellen, die dem Data-Science-Projekt zugrunde liegen, sind anzugeben.
Beteiligte Bereiche	Die domänenseitig an der Projektdurchführung beteiligten Bereiche sind zu benennen.
Komplexität	Eine Einschätzung der Komplexität der Problemstellung ermöglicht eine geeignete Einordnung.
Handlungsalternativen	Dies betrifft sowohl Alternativen in der Durchführung des Data-Science-Projekts als auch Alternativen zur Durchführung.

Eine Schärfung und Konkretisierung der Problemstellung ist häufig erst in den Folgeschritten möglich, weshalb an dieser Stelle keine besonderen formalen Anforderungen an Formulierungen oder Dokumentationsart gestellt werden. Auch ein spezifisches Abstraktionsniveau der Beschreibungen muss nicht vorgegeben werden, da sich Problemstellungen stark voneinander unterscheiden können.

10.2 Merkmal „*Domänenspezifika*“

Bei der Definition des Projekts ist die Domäne, in der die Problemstellung entstanden ist, unbedingt zu berücksichtigen. Es sind domänenspezifische Anforderungen an das Projekt zu definieren und die Vermittlung der für die Durchführung notwendigen Domänenkenntnisse an das Projektteam ist zu gewährleisten. Die domänenspezifischen Anforderungen führen häufig dazu, dass etablierte Data-Science-Methoden in der Anwendung oder bei der Auswertung der Ergebnisse (leicht) angepasst werden müssen.

In Tabelle 21 werden von Teilnehmerinnen und Teilnehmern häufig genannte Merkmale von Problemstellungen aufgeführt und beschrieben.

Tabelle 21: Beschreibung der Merkmale des Bereichs „Domänenspezifika“

Merkmal	Beschreibung
Fachliche Domäne	Eine Beschreibung der fachlichen Domäne, innerhalb derer die Problemstellung zu bearbeiten ist, muss erfolgen.
Anwendungsrahmen	Das Abstraktionsniveau ist festzulegen. Handelt es sich bei der zu entwickelnden Lösung z. B. nur um Handlungsempfehlungen für eine Abteilung oder geht es um strategische Entscheidungen eines Konzerns?
Ressourcenverfügbarkeit	Es ist festzulegen, welches Domänenwissen für die Durchführung des Projekts nötig ist, ob dieses Wissen innerhalb der Organisation vorhanden ist oder ob externe Experten hinzugezogen werden müssen.

10.3 Aufgabe „Definition des Projekts“

Ziel der Definition des Projekts ist es, auf der Grundlage der festgelegten Anforderungen an und Informationen über die Datengrundlage sowie der Miteinbeziehung der Domänenspezifika die notwendigen Arbeitsschritte festzulegen, die zur Erfüllung der festgelegten Anforderungen führen. Da sich der Vorgang kaum von der Projektdefinition bei anderen Projekten unterscheidet, sei an dieser Stelle auf Standardliteratur zum Projektmanagement verwiesen. Einzig der Ausgang von Data-Science-Projekten ist häufig ungewisser als bei Standardprojekten und muss ggf. intensiver betrachtet werden. Unter Umständen muss zwischen explorativen Forschungs- und Entwicklungsprojekten und solchen Projekten, die konkret auf eine Umsetzung bzw. einen Regelbetrieb abzielen, unterschieden werden.

In Abbildung 51 ist das Kompetenzprofil von Personen dargestellt, die sich in dem Bereich *Definition des Projekts* und den mit diesem direkt verbundenen Aufgaben spezialisieren, in Abbildung 52 sind beteiligte Rollen zu erkennen.

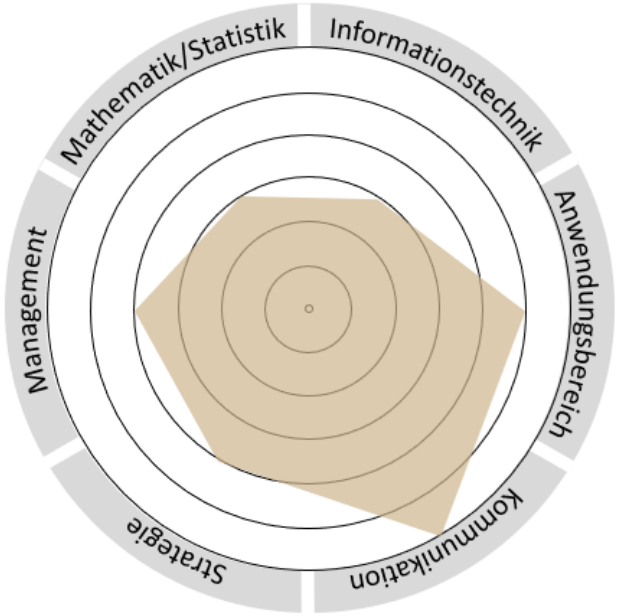


Abbildung 51: Kompetenzprofil des Bereichs „Definition des Projekts“

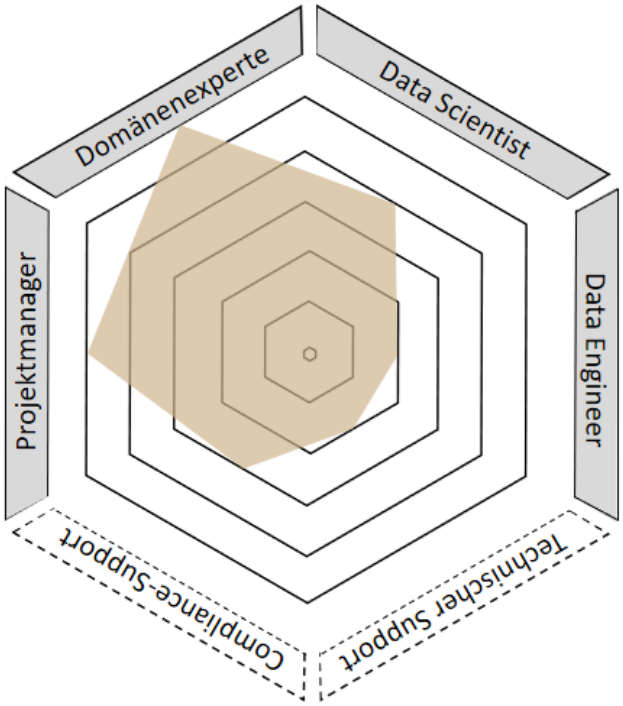


Abbildung 52: Rollen im Bereich „Definition des Projekts“

10.4 Begleitende Aufgabe „Eignungsprüfung“

Ziel der Eignungsprüfung ist es, zu entscheiden, ob sich das Projekt erfolgreich durchführen lässt. Dafür ist zu prüfen, ob die festgelegten Anforderungen unter Verwendung der eingeplanten Ressourcen erfüllt werden können. Die Aufgabe besteht darin, die zur Verfügung stehende Datengrundlage zu bewerten und einzuschätzen, ob durch die Anwendung von Analyseverfahren mit angemessener Wahrscheinlichkeit ein geeignetes Ergebnis erzielt werden kann. In Tabelle 22 werden von Teilnehmerinnen und Teilnehmern häufig genannte Aufgaben der Eignungsprüfung aufgeführt und beschrieben.

Tabelle 22: Beschreibung der Merkmale des Bereichs „Eignungsprüfung“

Aufgabe	Beschreibung
Eignung der Problemstellung	Die Durchführung einer Prüfung, ob es sich tatsächlich um eine Problemstellung handelt, bei der der Einsatz von Data Science geeignet erscheint, ist nötig.
Methodeneignung	Es ist zu prüfen, ob grundsätzlich Analyseverfahren existieren oder entwickelt werden können, die mit angemessener Wahrscheinlichkeit ein geeignetes Ergebnis erzielen. Hierfür sind ggf. auch bereits erste Tests durchzuführen.
Eignung des Ziels	Ein Abgleich der erwarteten Projektergebnisse mit der Problemstellung ist durchzuführen.
Berücksichtigung vergangener Projekte	Ein Abgleich von vergangenen Projekten mit dem derzeit in Definition befindlichen Projekt ist durchzuführen.
Vergleich von Projektalternativen	Unter Berücksichtigung knapper Ressourcen ist zu prüfen, ob die Bearbeitung der Problemstellung sinnvoll ist oder ob anderen Problemstellungen Vorzug gegeben werden sollte.
Durchführung von Experteninterviews	Durch die Einbindung von (weiteren) Experten kann die Projektdefinition validiert werden.

In Abbildung 53 ist das Kompetenzprofil von Personen dargestellt, die sich in dem Bereich *Eignungsprüfung* und den mit diesem direkt verbundenen Aufgaben spezialisieren, in Abbildung 54 sind beteiligte Rollen zu erkennen.

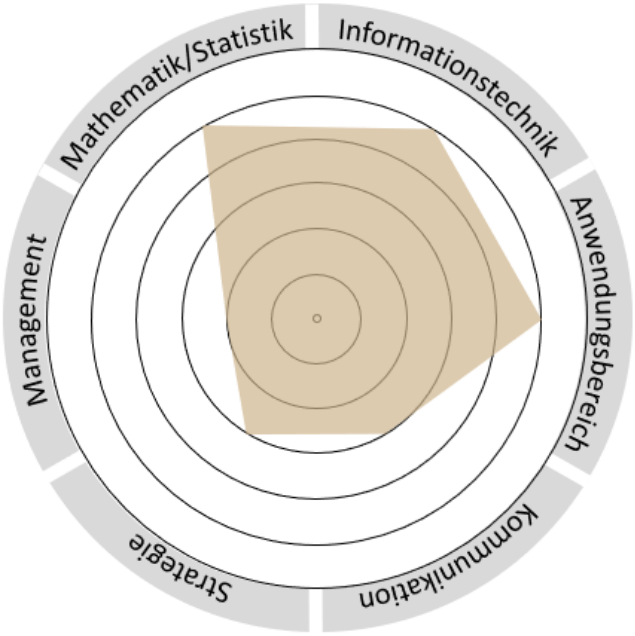


Abbildung 53: Kompetenzprofil des Bereichs „Eignungsprüfung“

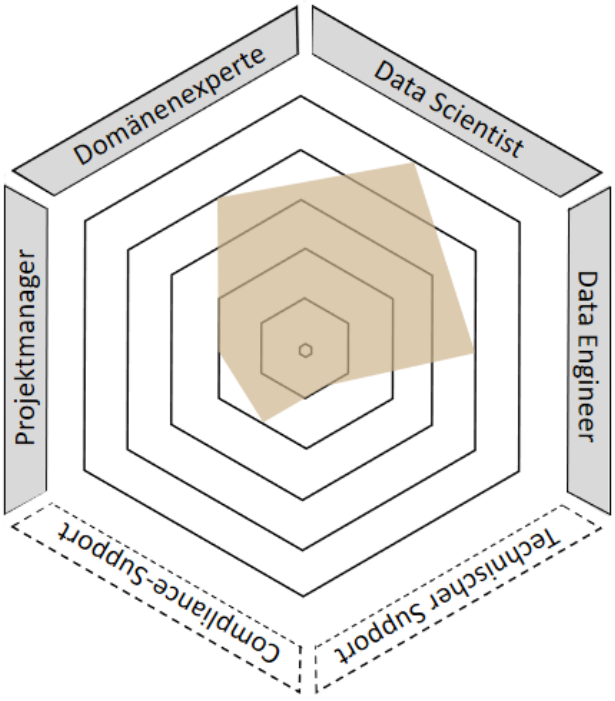


Abbildung 54: Rollen im Bereich „Eignungsprüfung“

10.5 Begleitende Aufgabe „Sicherstellung der Umsetzbarkeit“

In diesem Schritt ist zu prüfen, welche Projektideen sich umsetzen lassen. Oft handelt es sich dabei um einen iterativen Prozess mit allen Interessengruppen. In Tabelle 23 werden von Teilnehmerinnen und Teilnehmern häufig genannte Aufgaben zur Sicherstellung der Umsetzbarkeit aufgeführt und beschrieben.

Tabelle 23: Beschreibung der Merkmale des Bereichs „Sicherstellung der Umsetzbarkeit“

Aufgabe	**Beschreibung**
Prüfung der IT-Infrastruktur	Es ist zu prüfen, ob die vorhandene IT-Infrastruktur dazu geeignet ist, das definierte Projekt umzusetzen. Alternativ ist zu prüfen, ob andere technische Möglichkeiten existieren und ggf. weitere Infrastruktur angeschafft werden kann.
Bewertung der Expertise	Die Expertise der beteiligten Personen ist bzgl. ihrer Eignung für das gegebene Projekt zu prüfen.
Risikoeinschätzung	Das Projektrisiko (Eintrittswahrscheinlichkeiten des Risikos, Schwere der Konsequenzen) ist einzuschätzen.
Kosten-Nutzen-Analyse	Eine Analyse des Nutzens ist zwar häufig nur sehr schwer durchzuführen, die Kosten sollten aber grundsätzlich bewertet werden.

In Abbildung 55 ist das Kompetenzprofil von Personen dargestellt, die sich in dem Bereich *Sicherstellung der Umsetzbarkeit* und den mit diesem direkt verbundenen Aufgaben spezialisieren, in Abbildung 56 sind beteiligte Rollen zu erkennen.

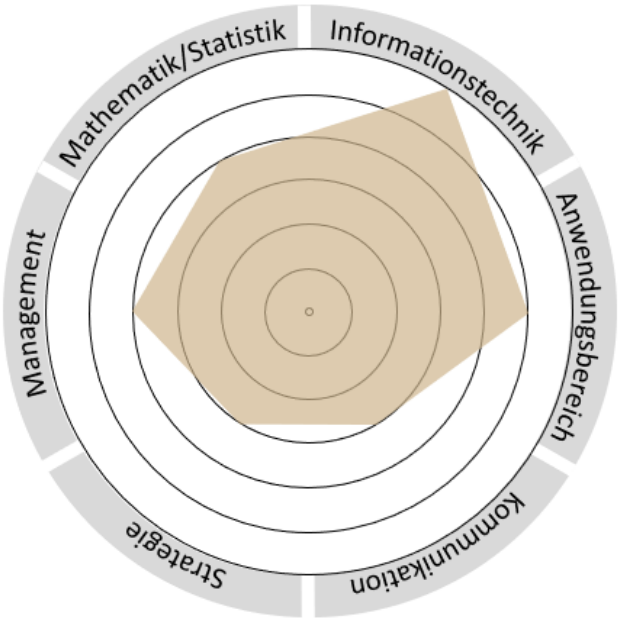


Abbildung 55: Kompetenzprofil des Bereichs „Sicherstellung der Umsetzbarkeit“

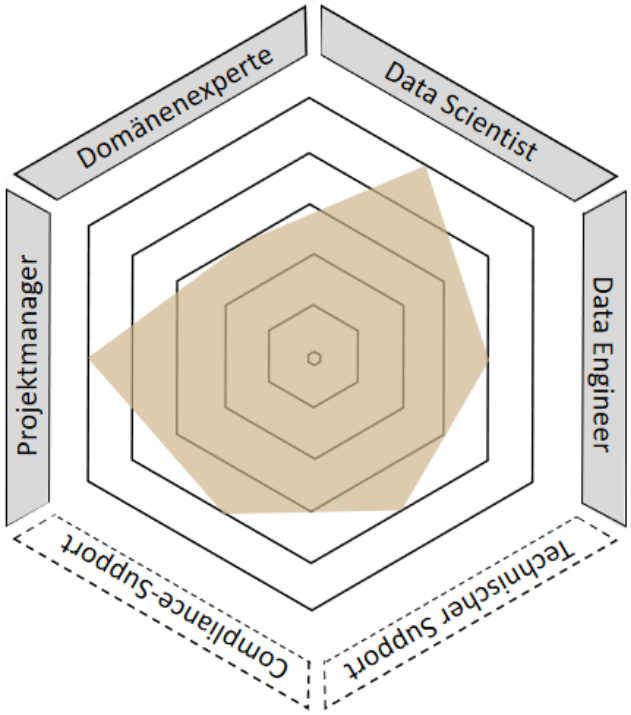


Abbildung 56: Rollen im Bereich „Sicherstellung der Umsetzbarkeit“

10.6 Merkmal „Projektskizze“

Eine vollständige Planung und Beschreibung des Ablaufes von Projekten ist im Data-Science-Kontext im Gegensatz zu vielen anderen Projekten in der Regel nicht möglich, wodurch als Ergebnis dieses Teilprozesses nur eine Projektskizze entstehen kann. Insbesondere sollte beachtet werden, dass die Projekte in der Regel in agiler Arbeitsweise durchgeführt werden. Beim Team Data Science Process (TDSP) ist beispielsweise die Scrum-Methodik bereits im Vorgehensmodell tief verankert. Genau wie die *Definition des Projekts* unterscheiden sich auch die Merkmale der Ergebnisskizze dabei nicht von denen anderer Projekte, weshalb auch

hier auf die Standardliteratur verwiesen wird. Wichtig ist, dass bei der Beschreibung des Projekts ein Abstraktionslevel gewählt wird, durch das alle relevanten Anforderungen und Informationen aus Daten-, Domänen- und Analysesicht prägnant dargestellt werden. Außerdem sollten durch die Beschreibung die zu diesem Zeitpunkt bereits identifizierbaren Arbeitsschritte aufgezeigt werden, die zur Erfüllung der festgelegten Anforderungen führen. Sollten sich bei der Projektdurchführung Änderungen ergeben, ist die Projektskizze anzupassen.

11. SCHLÜSSELBEREICH IT-INFRASTRUKTUR

Die IT-Infrastruktur ist in allen Schlüsselbereichen zu berücksichtigen, allerdings in unterschiedlichem Ausmaß. Zudem sind Art und Größe des Projekts ein wichtiger Faktor dahingehend, welche Rolle die Infrastruktur tatsächlich spielt. Dabei sind etwa die Form der geplanten Nutzbarmachung oder die Komplexität der verwendeten Daten und Analyseverfahren zu berücksichtigen. Auch die bestehende Infrastruktur der Organisation muss in der Regel bei einer Betrachtung berücksichtigt werden. Dies gilt hauptsächlich für solche Systeme, die direkt mit dem Data-Science-Projekt in Verbindung stehen, aber auch für Infrastruktur, die für das Projektmanagement oder die Zusammenarbeit im Team genutzt werden kann.

Nachfolgend werden die Teilbereiche – aufgegliedert nach Schlüsselbereichen – dargestellt, die von den Teilnehmerinnen und Teilnehmern am häufigsten als relevant in Bezug auf die IT-Infrastruktur genannt wurden.

Schlüsselbereich Daten

- *Analytische Datenquelle*: Es muss berücksichtigt werden, welche Zugriffsmöglichkeiten und Schnittstellen zu der analytischen Datenquelle bestehen. Möglicherweise schränkt das Zielsystem die verwendbaren Technologien ein.
- *Datenaufbereitung*: Die zur Verfügung stehende Rechenleistung ist zu berücksichtigen, auch unter Betrachtung der Software, die für die Datenaufbereitung verwendet wird.
- *Explorative Datenanalyse*: Die Rechenleistung der IT-Infrastruktur ist genauso zu berücksichtigen wie der Aspekt, ob Daten direkt in der Datenbank analysiert werden können oder ob zunächst ein Abzug von ihnen gemacht werden muss.
- *Ursprungsdatenquellen*: Es muss berücksichtigt werden, welche Zugriffsmöglichkeiten und Schnittstellen zu den Datenquellen bestehen. Möglicherweise schränken die Quellsysteme die verwendbaren Technologien ein.

Schlüsselbereich Analyseverfahren

- *Evaluation*: Mögliche Analyseverfahren sind unter Berücksichtigung der vorhandenen oder beschaffbaren Technologien zu evaluieren.
- *Identifikation geeigneter Analyseverfahren*: Es ist zu bewerten, welche IT-Infrastruktur benötigt wird, um die nötigen Analysen durchzuführen. Weiterhin ist zu prüfen, ob die Daten in der analytischen Datenquelle untersucht werden können oder ob sie zunächst heruntergeladen werden müssen.

Schlüsselbereich Nutzbarmachung

- *Sicherstellung technischer Umsetzbarkeit*: Innerhalb dieser begleitenden Aufgabe sind die Anforderungen an die IT-Infrastruktur sinnvoll zu detaillieren.
- *Technisch-methodische Bereitstellung*: Die IT-Infrastruktur muss dazu geeignet sein, das Modell in der angedachten Form zu betreiben. Dabei sind auch Möglichkeiten des Updates, des Backups und des Zugriffs zu berücksichtigen.

Schlüsselbereich Nutzung

- *Monitoring*: Eine wiederholte Überprüfung, ob die gewählte IT-Infrastruktur und ihre Dimensionierung für den Betrieb geeignet und effizient sind, ist durchzuführen.

Schlüsselbereich Domäne

- *Sicherstellung der Umsetzbarkeit*: Es ist zu prüfen, ob das Projekt mit der vorhandenen oder unter Berücksichtigung des Projektbudgets beschaffbaren IT-Infrastruktur umzusetzen ist.

12. SCHLÜSSELBEREICH WISSENSCHAFTLICHES VORGEHEN

Bei der Durchführung von Data-Science-Projekten ist ein wissenschaftliches Vorgehen in jeder einzelnen Phase nötig, was sich bereits in der Verwendung eines Vorgehensmodells widerspiegelt. Der Grad der Wissenschaftlichkeit kann variieren, die Mindestanforderungen vollständige Replizierbarkeit und statistische Validität müssen jedoch gewährleistet sein. Der Grad der Variation betrifft insbesondere die Theorieverankerung der Forschungsfrage, welche bei sehr praxisnahen Projekten kurz ausfallen kann. Dafür ist eine Kosten-Nutzen-Abwägung unter Berücksichtigung möglicher Risiken zu treffen, z. B. des Risikos, dass durch eine fehlende oder nur kurze Aufarbeitung der Literatur und ein dadurch bedingtes Übersehen wichtiger früherer Befunde oder Methoden auch der Lösungsraum eingeschränkt werden könnte. Unabhängig von Merkmalen des individuellen Projekts sind, neben einem strukturierten Vorgehen, eine geeignete Dokumentation und eine statistisch fundierte Evaluation bzw. Validierung der Ergebnisse in jedem Fall unabdingbar.

Zunächst sollen die wissenschaftlichen Anforderungen kurz beleuchtet werden, die das gesamte Data-Science-Projekt betreffen. Nachfolgend werden die Teilbereiche, aufgegliedert nach Schlüsselbereichen, dargestellt, welche von den Teilnehmerinnen und Teilnehmern am häufigsten als relevant in Bezug auf das wissenschaftliche Vorgehen genannt wurden.

Alle Schlüsselbereiche betreffende wissenschaftliche Anforderungen

Grundsätzlich gelten für ein Data-Science-Projekt dieselben grundlegenden Standards, denen auch andere praxisnahe wissenschaftliche Arbeiten genügen müssen. Dies sind vor allem vier Punkte:

1. Der Forschungsgegenstand (im vorliegenden Fall der Projektauftrag) muss so genau umrissen sein, dass er auch für Dritte erkennbar ist. Dies ist wichtig, um die Aussage des wissenschaftlichen Beitrags eingrenzen und einordnen zu können, jedoch auch, um die passenden Methoden zu wählen.
2. Das Resultat des Projekts muss eine Aussage sein, die so bisher noch nicht (aus diesem Blickwinkel) getroffen werden konnte. Anderenfalls wäre das Projekt obsolet.
3. Das Resultat muss nützlich sein, was aber häufig schon durch den Projektauftrag gegeben ist.
4. Das Projekt muss so dokumentiert sein, dass es einer „wissenschaftlichen Öffentlichkeit“ möglich ist, anhand der bestehenden Angaben die getroffenen Aussagen/Hypothesen nachzuprüfen. Gerade in Unternehmenskontexten kann die „wissenschaftliche Öffentlichkeit“ aber auch nur eine unternehmensinterne sein. Dieser letzte Punkt geht aber auch mit dem Grundsatz der Replizierbarkeit einher, die sicherstellt, dass die Methode so gut beschrieben ist, dass eine andere Partei mit Zugriff auf dieselbe Infrastruktur und dieselben Daten zum gleichen Ergebnis kommt. Zudem hat dies auch Implikationen für die statistische Belastbarkeit der Ergebnisse, welche durch Rigorosität in der Auswertung sichergestellt werden muss.

Zudem gilt es, die drei technischen Ansprüche Objektivität, Reliabilität und Validität zu beachten.

Schlüsselbereich Domäne

- *Definition des Projekts*: Für das Projekt wird es typischerweise, unabhängig davon, ob es im wirtschaftlichen oder wissenschaftlichen Bereich angesiedelt ist, einen wissenschaftlichen Kontext geben, welcher eine Aufarbeitung existierender Verfahren und wissenschaftlicher Publikationen abhängig vom Projekt notwendig machen kann.

Schlüsselbereich Daten

- *Explorative Datenanalyse*: Daten müssen möglichst umfänglich verstanden werden und es muss eine fundierte statistische Evaluation bzw. Validierung der Ergebnisse sowie des Zustandekommens von potenziellen Fehlern in den Daten durchgeführt werden. Die Eignung der Daten zur Untersuchung der Problemstellung muss geprüft werden, Datenbereinigungsanforderungen sind zu identifizieren. Ein Nachweis über die korrekte Anwendung einer geeigneten explorativen Datenanalyse muss erbracht werden.
- *Datenaufbereitung*: Die Datentransformation muss transparent und replizierbar sein, es sind korrekte Verfahren zu verwenden und Aufbereitungsschritte in geeigneter Weise zu dokumentieren. Die Rohdaten sind für die Reproduzierbarkeit der Datenaufbereitung langfristig zu archivieren.

Schlüsselbereich Analyseverfahren

- *Identifikation geeigneter Analyseverfahren*: Die Anforderungen an das zu entwickelnde Analyseverfahren sind zu prüfen, Ziele und Rahmenparameter nachvollziehbar sind festzulegen. Eine Übersicht über vorhandene Verfahren gemäß diesen Kriterien inkl. der Berücksichtigung aktueller wissenschaftlicher Veröffentlichungen ist zu erstellen und die Auswahl der Analyseverfahren ist zu begründen. Auch die Erkenntnis, dass kein geeignetes Verfahren existiert, muss in geeigneter Form dargelegt werden.
- *Anwendung von Analyseverfahren*: Bei der Parametrisierung von Analyseverfahren ist zielgerichtet vorzugehen. Die korrekte Anwendung des Analyseverfahrens ist genauso zu gewährleisten wie die Durchführungsobjektivität. Gerade auch zur korrekten Anwendung ist die wissenschaftliche Literatur heranzuziehen. Insbesondere muss sichergestellt sein, dass die Grundannahmen für das Analyseverfahren gegeben sind. Eine Dokumentation von Analyseergebnissen inklusive deren Interpretation ist anzufertigen. Von Beginn an muss die Evaluation mitgedacht werden, Test- sowie Validierungsdatensätze müssen in geeigneter Form vorgehalten werden. Nur so kann verhindert werden, dass die Analyseergebnisse statistische Artefakte der betrachteten Daten widerspiegeln und keine allgemeingültigen Zusammenhänge.
- *Entwicklung von Analyseverfahren:* Bei der Integration bestehender Verfahren muss dargelegt werden, an welchen Stellen a.) bisherige Methoden eingebaut werden, b.) bisherige Methoden Schwachstellen aufweisen und wie diese durch Veränderung der Verfahren beseitigt werden und c.) nachweisbar noch keine Verfahren existieren, sodass eine Neuentwicklung notwendig ist. Hierbei ist auf die Ergebnisse aus der Identifikation zurückzugreifen. Auch bei der Entwicklung eines Analyseverfahrens ist eine Interaktion mit der Fach-Community, um geeignete Verfahren nach definierten Standards zu entwickeln und ggf. auch überprüfen zu lassen, sinnvoll.
- *Evaluation*: Eine systematische Aufbereitung von Bewertungen und Tests ist durchzuführen, eine korrekte Anwendung geeigneter Evaluationsverfahren unter Verwendung einer gleichbleibenden Testumgebung ist sicherzustellen. Die korrekte Funktionsweise der Verfahren ist nachzuweisen, die Ergebnisse sind kritisch zu bewerten und die Evaluation ist vollständig zu dokumentieren.

Schlüsselbereich Nutzbarmachung

- *Technisch-methodische Bereitstellung*: Nach Abschluss des Analyseverfahrens wird dieses technisch und methodisch zur Verfügung gestellt. Dies kann in Form von abgeschlossenen Software-Modulen oder -Paketen geschehen oder als nutzbarer Service innerhalb einer IT-Infrastruktur. Letzteres beinhaltet beispielsweise die Bereitstellung

einer Programmierschnittstelle oder eines (Web-)Services. Damit Nutzer das Analyseverfahren nutzen können, sollten eine vollständige Beschreibung und Dokumentation bereitgestellt werden.

Schlüsselbereich Nutzung

- *Verwendung der Analyseartefakte*: Die Nutzung erfolgt im Wesentlichen nach wirtschaftlichen und domänenspezifischen Erfordernissen. Wissenschaftliche Erfordernisse sind dagegen vernachlässigbar und ein wissenschaftliches Vorgehen ist dementsprechend nicht notwendig.
- *Monitoring & Auswertung der Nutzungskenntnisse:* Mit der Übergabe der Analyseartefakte an die Nutzer sind Hypothesen über die Leistung der Artefakte verbunden. Inwieweit die Analyseartefakte diesen Hypothesen gerecht werden, muss erfasst und wissenschaftlich korrekt bewertet werden. Eine wesentliche Anforderung ist die Dokumentation von Unterschieden zwischen der Herkunftsumgebung von Trainingsdaten und der tatsächlichen Nutzungsumgebung der Analyseartefakte sowie von Änderungen in der Nutzungsumgebung.

13. FAZIT

Ein umfangreiches Thema zu strukturieren, um es in Gänze erfassen zu können und einzelne Teile dann gezielt zu nutzen, ist eine in Wissenschaft und Praxis gleichermaßen verbreitete Vorgehensweise. Dass insbesondere jene, die sich professionell mit Strukturen, Mustern und analytischer Aufbereitung befassen, den Drang haben, ein komplexes Themenfeld wie Data Science zu durchdringen und für eine größere Leserschaft aufzubereiten, ist daher keinesfalls verwunderlich. Das vorliegende Ergebnis ist das Ende eines solchen Aufbereitungsprozesses und stellt auf unterschiedlichen Ebenen und in vielen verschiedenen Facetten vor, wie Praktiker und Forscher das Thema Data Science wahrnehmen, umsetzen und in ihrem Alltag verankern. Leser dieser Ausarbeitung erhalten so einen gleichermaßen strukturiert aufbereiteten wie direkt auf ihren eigenen beruflichen Kontext anwendbaren Katalog an Erkenntnissen.

Um dies zu erreichen, wurde zunächst ein umfassenderes Bild von Data Science und den zugehörigen Themenfeldern sowie den verwandten Begriffen gezeichnet. Auch in den Umfragen innerhalb der Arbeitsgruppe zeigt sich, was die vorhandene Literatur vermuten lässt: Data Science ist ein vielschichtiges und stark interdisziplinär geprägtes Arbeits- und Forschungsgebiet. Mit der erarbeiteten Definition liegt eine umfangreiche und dennoch präzise Beschreibung der wesentlichen Merkmale vor.

Ganz im Sinne einer praxisrelevanten Ausarbeitung wurde mit DASC-PM ein Prozessmodell entwickelt, das die relevanten Schritte in der projektgetriebenen Anwendung von Data Science darlegt und für die Durchführung von Data-Science-Aktivitäten detailliert beschreibt. Erfahrene Anwender auf dem Gebiet der Datenanalyse finden dabei ein Modell vor, das in der Struktur Ähnlichkeiten zu den seit vielen Jahren erprobten Modellen, wie z. B. CRISP-DM, aufweist, sodass eine Überführung von bereits etablierten Aktivitäten mit überschaubarem Aufwand in DASC-PM gelingt. Neueinsteiger in die Thematik wiederum erhalten ein Modell, dass die Komplexität von Data-Science-Initiativen auf die Kernthematiken reduziert und sukzessive ausformuliert, sodass bei einer ersten Durchführung von Data-Science-Projekten schwerpunktmäßig dort vertieft werden kann, wo dies nötig erscheint. In beiden Fällen hebt DASC-PM als Ergebnis eines intensiven Austauschs zwischen Wissenschaftlern und Praktikern dabei auch ein wissenschaftliches Vorgehen als Kerneigenschaft hervor und unterstützt die Anwender des Modells dabei, nachvollziehbar und methodisch vorzugehen, damit die Ergebnisse gleichermaßen mehrwertstiftend wie belastbar sind.

Das Prozessmodell führt für jede definierte Kernkompetenz auf, welche Aktivitäten und Ergebnisse im Rahmen von Data-Science-Initiativen relevant sind und wie sie ausgestaltet werden können. Dabei werden zu jedem Komplex die wichtigsten Aufgaben beschrieben und definiert. Die umfangreichen Aufzählungen von Merkmalen zu den einzelnen Aufgaben erlauben es jedem Anwender des Modells, eine kritische Betrachtung der eigenen Vorgehensweise durchzuführen oder sich auf Basis der dargestellten Möglichkeiten die für sein Unternehmen relevantesten Merkmale herauszusuchen. Dabei kann eine bewusste Selektion durchgeführt werden, die ein mühsames Zusammensuchen aus diversen Quellen erspart und – ganz im Sinne wissenschaftlichen Handelns – eine umfangreiche Referenz bietet, sodass die verwendeten Vorgehensweisen nicht wie zufällig gewählt erscheinen, sondern auf den Erfahrungen und dem Austausch einer großen Gruppe von Fachexperten beruhen. Ergänzend dazu stellt DASC-PM auch diverse Hinweise zu weiterführenden Ausarbeitungen oder Kriterienkatalogen bereit, so z. B. zum Thema Datenqualität.

Neben den methodischen Betrachtungen stellt DASC-PM mit dem „Data Scientist“ aber auch die wichtigste Komponente einer erfolgreichen Data Science in den Vordergrund. Wie bereits zu Beginn dieser Ausarbeitung dargelegt, begann der (begriffliche) Siegeszug der Data Sci-

ence nicht mit der „Datenwissenschaft“ als solcher, sondern mit den Analysten, die sie anwendeten. Der „Data Scientist“ ist die markante Figur im diskutierten Themenbereich (Davenport und Patil, 2012). Seine Tätigkeiten und die konkrete Abgrenzung seines Themenfeldes sind bereits seit einigen Jahren Gegenstand einer dynamischen Diskussion, die auch entsprechende Publikationen hervorgebracht hat (Harris et al., 2013; Zschech et al., 2018).

Auch in der hier durchgeführten Betrachtung der Kernkompetenzen zeigt sich, dass es „den einen“ Data Scientist in Reinform nicht gibt – und auch nicht geben muss. Während für die initiale Bereitstellung, Aufbereitung und explorative Analyse von Daten vorrangig Kenntnisse in Mathematik, Statistik und Informationstechnik vorhanden sein müssen, verschiebt sich das notwendige Kompetenzprofil bei der Betrachtung der Analyseverfahren leicht und nimmt vor allem ein größeres Verständnis des Anwendungsbereichs als Anforderung mit auf. Dass das Kernprofil des Schlüsselbereichs Analyseverfahren dabei weiterhin ein Verständnis der Informationstechnik fordert, zusätzlich grundlegende Kommunikationsfähigkeiten verlangt und damit umfangreicher ist als das der meisten anderen Bereiche, liegt darin begründet, dass Data Science wie hier dargestellt im Kern die Datenanalyse beschreibt. Dort liegt entsprechend auch das komplexeste Anforderungsprofil vor.

Je stärker die Schlüsselbereiche die fachliche Sicht betreffen, desto mehr treten Mathematik, Statistik und Informationstechnik als Kompetenzen in den Hintergrund. Entscheidend für die Nutzbarmachung und den allgemeinen Auftritt in der Domäne sind vielmehr Kommunikationsfähigkeiten, strategisches Verständnis und maßgeblich ein hohes Verständnis des Anwendungsbereichs. Personalverantwortliche können auch aus den umfangreichen Darstellungen zu den einzelnen Schlüsselbereichen ablesen, dass Projektmanagement als Kompetenz tatsächlich nur wenige Personen betrifft. Es zeigt sich, analog zu vielen anderen Überlegungen zu dem Thema, dass auch ein Data-Science-Projekt ein orchestrierendes Element benötigt, aber nicht jeder Data Scientist dazu ein Projektmanager sein oder überhaupt über umfangreiches Wissen in allen Bereichen verfügen muss.

Für Unternehmen ist all dies eine gute Nachricht. Die hohe Nachfrage nach Data Scientists auf dem Arbeitsmarkt macht es in Verbindung mit dem großen Umfang an prinzipiell relevanten Skills nahezu unmöglich, einen Tausendsassa, eine eierlegende Wollmilchsau, ein „Einhorn“ zu finden. Wird aber DASC-PM angewendet, um den Analyseprozess zu strukturieren, können die einzelnen Stationen gezielt mit Personen besetzt werden, die in ihrem Gebiet über hohe Expertise verfügen und diese auch an den geeigneten Stellen einsetzen können. Solche Experten sind in Unternehmen vielfach vorhanden oder können entsprechend gesucht oder ausgebildet werden, ohne dass eine Masse sehr unterschiedlicher Qualifikationen auf einmal in einer Person vereint werden muss.

Das vorliegende Prozessmodell ist – wie alle Modelle – eine vereinfachte Version der Wirklichkeit. Weder muss es sklavisch befolgt werden, noch erhebt es den Anspruch, jede Variante und Eventualität eines Vorgehens oder einer Methodik darzulegen. Es bietet auch keine Anleitung zur vollständigen Abarbeitung jedes einzelnen abgebildeten Bausteins. Vielmehr ist das Modell eine solide Grundlage zur Durchführung von Data-Initiativen, da es auf mehr als nur die Erfahrungen eines einzelnen Unternehmens oder einer einzelnen Forschungsgruppe zurückgreifen kann. DASC-PM ist daher mehr als ein Best-Practice-Ansatz. Es ist eine strukturierte, fundierte und umsetzbare Aufbereitung eines der relevantesten Themen der Wirtschaft und Wissenschaft, nämlich der planvollen und ergebnisorientierten Nutzbarmachung von Daten, der Data Science.

LITERATUR

Chatfield, A. T., Shlemoon, V. N., Redublado, W., & Rahman, F. (2014). Data scientists as game changers in big data environments. In*: Proceedings of the 25th Australasian Conference in Information Systems*.

Conway, D. (2010). The data science venn diagram. Dataists, drewconway.com/zia/2013/3/26/the-data-science-venn-diagram, aufgerufen am 03.02.2020.

Davenport, T. H., & Patil, D. J. (2012). Data scientist. *Harvard Business Review*, 90 (5), 70-76.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56 (12), 64-73.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17 (3), 37.

Harris, H., Murphy, S., & Vaisman, M. (2013). Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work. O'Reilly.

Helfert, M., Hermann, C., & Strauch, B. (2001). Datenqualitätsmanagement. Institut für Wirtschaftsinformatik, Universität St. Gallen.

McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management Revolution. *Harvard Business Review*, 90 (10), 60-66.

Microsoft (2017): What is the Team Data Science Process?, https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview, aufgerufen am 16.05.2019.

Olivotti, D., Passlick, J., Axjonow, A., Eilers, D., & Breitner, M. H. (2018). Combining machine learning and domain experience: a hybrid-learning monitor approach for industrial machines. In: *International Conference on Exploring Service Science* (261-273). Springer, Cham.

Palmer, M. (2006). Data is the New Oil, https://ana.blogs.com/maestros/2006/11/data_is_the_new.html, letzter Zugriff: 09.12.2019.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data*, 1 (1), 51-59.

van der Aalst, W. (2016). Data science in action. In: *Process Mining* (3-23). Springer, Berlin, Heidelberg.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In: *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (29-39).

Zschech, P., Fleißner, V., Baumgärtel, N., & Hilbert, A. (2018). Data Science Skills and Enabling Enterprise Systems. *HMD Praxis der Wirtschaftsinformatik*, 55 (1), 163-181.

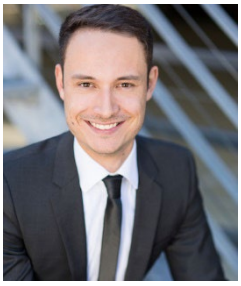
Dr. Emal M. Alekozai holds a MSc in physics, a PhD in computer science from Heidelberg University (Interdisciplinary Center for Scientific Computing) in collaboration with the research center MATHEON and Oak Ridge National Laboratory and an MBA from the Mannheim Business School. He has over a decade of work experience in IT and risk management consulting for corporates, national & supranational financial institution and regulators. Currently he is with Bosch in-house consulting responsible for SAP finance and data science topics.

Wolfgang Badewitz (M.Sc.) ist seit 2019 wissenschaftlicher Mitarbeiter am FZI Forschungszentrum Informatik und promoviert am Institute of Information System and Marketing (IISM) des KIT. Dabei beschäftigt er sich mit der ökonomischen Ausgestaltung von kollaborativen Data Value Chains im Industrie-4.0-Kontext. Seine besondere Aufmerksamkeit gilt Mechanismen zur Incentivierung einer erfolgreichen Datenweitergabe zwischen Partnern in Wertschöpfungsnetzwerken.

Daniel Badura (M.A.) arbeitet als Consultant bei der valantic Business Analytics GmbH in den Bereichen Data Science und Business Intelligence. Zuvor studierte er Angewandte Wirtschaftssprachen und internationale Unternehmensführung an der Hochschule Bremen und der East China Normal University Shanghai sowie Betriebswirtschaft an der Hochschule Coburg.

Prof. Dr. Harald Binder ist Direktor des Instituts für Medizinische Biometrie und Statistik am Universitätsklinikum Freiburg. Er studierte von 1996 bis 2001 Psychologie und Mathematical Behavioral Sciences (Regensburg und University of California, Irvine) und wurde anschließend an der LMU München in Statistik promoviert. Nach seiner Postdoc-Zeit in Freiburg übernahm er 2011 die Professur und Leitung der Abteilung Biometrie und Statistik an der Universitätsmedizin Mainz, bevor er 2017 die Professur in Freiburg antrat. Seine Forschungsschwerpunkte liegen auf Techniken des maschinellen Lernens für molekulare und klinische Daten.

Prof. Dr. Dorothee Brauner ist seit 2016 Professorin für BWL mit Schwerpunkt quantitative Methoden der Marktforschung an der HS Esslingen. Im Rahmen Ihrer Vorlesungen lehrt sie im Bereich quantitative Marktforschung und Advanced Data Analytics die Studierenden, wie man Business-Probleme in analytische Lösungen übersetzen kann. Vor Ihrer Tätigkeit als Professorin war sie zuletzt als Managerin im Bereich Data Science bei MHP – A Porsche Company tätig. Sie arbeitete dort hauptsächlich als Expertin in den Bereichen Big-Data-Strategie und Data Mining.



David Dann (M.Sc.) ist seit 2017 wissenschaftlicher Mitarbeiter am Institut für Wirtschaftsinformatik und Marketing am Karlsruher Institut für Technologie. Er leitet dort die Forschungsgruppe Electronic Markets & User Behavior. In seiner Forschung beschäftigt er sich mit Peer-to-Peer-Plattformen, Data Analytics, Machine Learning sowie Internet User Behavior und Psychologie.

Michael Felderer is a professor of software and data engineering at the University of Innsbruck, Austria and a guest professor at the Blekinge Institute of Technology, Sweden. His fields of expertise and interest include software and data quality, testing, data-driven engineering, soft-ware analytics, requirements engineering and model-based software engineering. Michael Felderer co-authored more than 130 publications and received 10 best paper awards. His re-search has a strong empirical focus also using methods of data science and is directed towards development and evaluation of efficient and effective methods to improve quality and value of industrial software systems and processes in close collaboration with companies. For more in-formation, visit his website at mfelderer.at.

Prof. Dr. Nick Gehrke
Professor an der NORDAKADEMIE – Hochschule der Wirtschaft im Fachbereich Informatik. Nach dem Studium der Betriebswirtschaftslehre war er als wissenschaftlicher Mitarbeiter am Institut für Wirtschaftsinformatik an der Universität Göttingen tätig. Anschließend folgte eine 5-jährige Tätigkeit in der prüfungsnahen Beratung bei der PricewaterhouseCoopers AG in Hamburg. Dort legte er das Steuerberaterexamen ab und qualifizierte sich zum Certified Information System Auditor (CISA). Er ist Mitglied der Gesellschaft für Informatik, der Information Systems Audit and Control Association (ISACA), der Steuerberaterkammer Schleswig-Holstein, des Berufsverbandes für Aufsichtsgremien Financial Experts Association e.V. und des Verbandes der Hochschullehrer für Betriebswirtschaft e.V. (VHB). Co-Founder und Data Scientist zapliance GmbH. Co-Founder und Board Member von ARIC (Artificial Intelligence Center Hamburg e.V.).

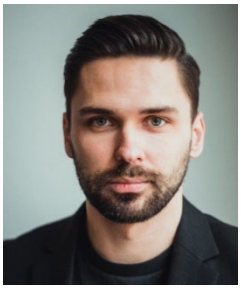
Prof. Dr.-Ing. Philipp Gölzer ist Professor für Digitale Fabrik und Materialflusssysteme an der Technischen Hochschule in Nürnberg. Nach seiner Promotion im Themenfeld Industrie 4.0 / Big Data war er mit Aufbau und Leitung der Forschergruppe Data Science & Optimization am Fraunhofer IIS/SCS betraut. Forschungsaktivitäten und -interessen liegen im Bereich modellgetriebener und datengetriebener Verfahren für Fragestellungen in Produktion und Logistik. In Nebentätigkeit ist er heute als Geschäftsfeldkoordinator Digitalisierte Produktion weiterhin für das Fraunhofer IIS/SCS tätig.

Dr. Alexander Gröschel ist Postdoktorand am FZI Forschungszentrum Informatik in Karlsruhe. Er leitet die Forschungsgruppe Business Data Analytics. Seine Arbeitsschwerpunkte liegen sowohl im Bereich von Forecast-Methoden für Predictive-Maintenance-Anwendungen als auch im Bereich der Systemvirtualisierung von Produktionsanlagen und Demand Forecasting im Industrie-4.0-Kontext.

Prof. Dr. Jens Kaufmann ist Inhaber der Professur für Wirtschaftsinformatik, insb. Data Science, an der Hochschule Niederrhein. Zuvor war er mehrere Jahre in der Beratung bei Horváth & Partners sowie im Bereich des Global CIO bei der ERGO Group AG in Düsseldorf tätig. Er dozierte als Gastprofessor an der University of North Carolina in Charlotte, NC, USA, und beschäftigt sich in Lehre und Forschung schwerpunktmäßig mit der Anwendung von Data Science und ihrem Transfer in die betriebliche Praxis.

Prof. Dr. Ulrich Kerzel
2018 ff.: Professur Data Science & AI, IUBH Internationale Hochschule GmbH
2017/18: Vice President Data Science RWTH Business School GmbH
2012-17: Principal Data Scientist Blue Yonder (jetzt: JDA), Director Data Science Academy
2010-12: Research Fellow CERN
2008-10: Senior Research Fellow Magdalene College, Cambridge, UK
2006-08: Research Associate Trinity Hall, Cambridge, UK
2006-10: PostDoc, University of Cambridge, UK
2005-05: PostDoc, KIT
2005: Promotion (Physik)

Dr. Simon Kloker wurde 2018 am Institut für Wirtschaftsinformatik und Marketing am Karlsruher Institut für Technologie promoviert. Seither leitet er dort die Forschungsgruppe Business Data Analytics. In seiner Forschung beschäftigt er sich mit der Erkennung und Prävention von süchtigem Verhalten bei der Nutzung von Informationssystemen.



Dr. Stephan Kühnel studierte Betriebswirtschaftslehre an der Martin-Luther-Universität Halle-Wittenberg und arbeitet dort seit 2013 als wissenschaftlicher Mitarbeiter am Lehrstuhl für Betriebliches Informationsmanagement. Dr. Kühnel war von 2014 bis 2017 in mehreren Praxisprojekten als Data Scientist tätig und promovierte im Jahr 2019 zum Thema wirtschaftliche Bewertung von Geschäftsprozess-Compliance. Seit 2019 arbeitet er als Postdoctoral Researcher und forscht in den Themenbereichen Process & Data Mining sowie IT-Sicherheit & IT-Compliance.

Prof. Dr. Carsten Lanquillon ist Professor für Business Intelligence und quantitative Methoden und Mitgründer des Data Science Lab an der Hochschule Heilbronn. Seine Forschungsschwerpunkte liegen in den Bereichen Machine Learning, Data Science, Text Mining und Big Data Analytics. Seit mehr als 20 Jahren beschäftigt er sich sowohl in der Praxis als auch in der Theorie erfolgreich mit der Entwicklung und Anwendung von Data-Science-Lösungen für vielfältige fachliche Fragestellungen aus unterschiedlichen Fachbereichen in großen und kleinen Unternehmen.

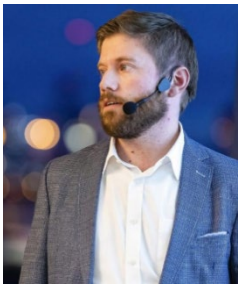
Dipl.-Inform. Uwe Neuhaus ist Dozent und wissenschaftlicher Mitarbeiter im Fachbereich Informatik an der NORDAKADEMIE. Nach seinem Informatikstudium an der TU Braunschweig und einem interdisziplinären Ergänzungsstudium an der Washington University, St. Louis, arbeitete er für die FernUniversität Hagen im Lehrgebiet Datenbanken und Informationssysteme sowie als Technical Trainer für einen führenden Entwickler von CMS-Systemen. Seine Schwerpunkte an der NORDAKADEMIE sind die Bereiche Algorithmen, Analytische Informationssysteme und Machine Learning.

Jens Passlick (M.Sc.)
Seit März 2016 ist Jens Passlick wissenschaftlicher Mitarbeiter am Institut für Wirtschaftsinformatik an der Leibniz Universität Hannover. Davor hat er ein Masterstudium mit dem Schwerpunkt Operations Management and Research in Hannover erfolgreich abgeschlossen. Das Bachelorstudium wurde in Hannover und Gent (Belgien) absolviert. Seine Forschungsinteressen liegen in den Bereichen Business Analytics – vor allem Self-Service Analytics –, Industrie-4.0-Themen und Forschungsmethoden.

Maik Prothmann (B.Sc.) hat 2014 seinen B.Sc. in internationalem Finanzmanagement an der HfWU-Nürtingen-Geislingen erlangt. Während seiner Bachelor-Thesis befasste er sich ausführlich mit automatisierten Handelssystemen für den Kapitalmarkt und entwickelte ein eigenes. Seither läuft dieses als angemeldetes Gewerbe weiter. Im Anschluss an den Bachelor studierte er noch 5 weitere Semester Wirtschaftsinformatik an der Hochschule Esslingen. Während beider Studiengänge sammelte er mittels Praktika und Werksstudententätigkeiten erste Arbeitserfahrungen in dem Bereich Finanzwesen mit dem Schwerpunkt Kapitalanlage. Zurzeit arbeitet er als In-House-Consultant bei der Robert Bosch GmbH für den Finance-Bereich mit Schwerpunkt SAP Finance und Data Science.

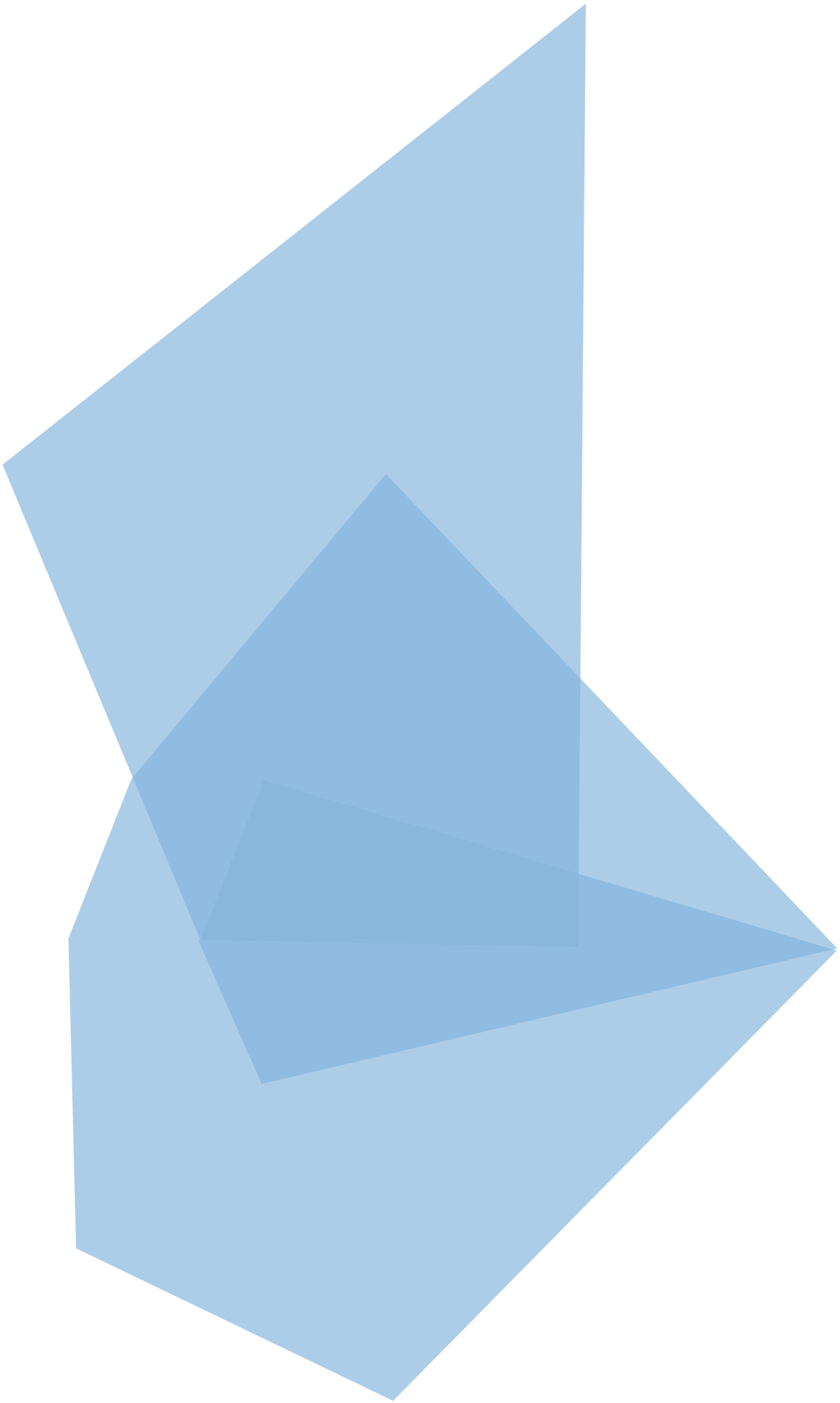
Dr. Raphael Rissler promovierte am Institute of Information System and Marketing (IISM) des KIT. Seine Forschungsinteressen umfassen die Analyse von Flow-Zuständen während der Arbeit, den Einfluss von IT-ausgelösten Unterbrechungen auf diese Zustände sowie deren Klassifikation mittels Machine Learning (ML) anhand physiologischer Indikatoren. Raphael Rissler forscht und arbeitet heute als Data Scientist bei der SAP SE und führt Kundenprojekte im Bereich ML und Künstliche Intelligenz durch.

Heiko Rohde (M.Sc.) ist Managing Consultant bei der valantic Business Analytics GmbH in den Bereichen Business Analytics, Planung und Data Science. Zuvor hat er einen Bachelor of Science in Wirtschaftsinformatik und einen Master of Science in Betriebswirtschaftslehre an der FH Wedel abgeschlossen.
Seit 2011 begleitet Herr Rohde Kunden in vielseitigen Projekten bei der digitalen Transformation und unterstützt sie mit neuen Technologien bei der Integration und Nutzung der vielfältigen Daten.



Prof. Dr. Michael Schulz hält eine Professur für Wirtschaftsinformatik, insb. analytische Informationssysteme, an der NORDAKADEMIE – Hochschule der Wirtschaft in Elmshorn. Zudem ist er als Projektmanager bei der valantic Business Analytics GmbH tätig. Herr Schulz promovierte nach berufsbegleitenden Studien zum Diplom-Kaufmann (FH) und Master of Science in Wirtschaftsinformatik an der Philipps-Universität Marburg. Seine Interessenschwerpunkte in Lehre, Forschung und Praxisprojekten liegen in der Business Intelligence und der Data Science.

Felix Welter (B.Sc.) hat sein duales Studium der Wirtschaftsinformatik an der NORDAKADEMIE in Kooperation mit der Firma Otto (GmbH & Co KG) abgeschlossen. Dabei hat er an verschiedenen Data-Science-Projekten teilgenommen und nach Ende des Bachelorstudiums im Bereich Natural Language Processing gearbeitet. Derzeit verfolgt er an der Universität Hamburg seinen Master of Science der Wirtschaftsinformatik mit dem Schwerpunkt „Entwicklung und Management von Informationssystemen“.



ISBN e-Book: 978-3-00-064898-4