Michael Schulz ▪ Uwe Neuhaus ▪ Jens Kaufmann ▪ Stephan Kühnel ▪ Emal M. Alekozai ▪ Heiko Rohde ▪ Sayed Hoseini ▪ René Theuerkauf

DASC-PM v1.1

Ein Vorgehensmodell für Data-Science-Projekte

Daniel Badura ▪ Ulrich Kerzel ▪ Carsten Lanquillon ▪ Stephan Daurer ▪ Maik Günther ▪ Lukas Huber ▪ Lukas-Walter Thiée ▪ Philipp zur Heiden ▪ Jens Passlick ▪ Jonas Dieckmann ▪ Florian Schwade ▪ Tobias Seyffarth ▪ Wolfgang Badewitz ▪ Raphael Rissler ▪ Stefan Sackmann ▪ Philipp Gölzer ▪ Felix Welter ▪ Jochen Röth ▪ Julian Seidelmann ▪ Uwe Haneke

dasc°pm v1.1

ISBN: 978-3-9824465-0-9

Elmshorn 2022

info@dasc-pm.org

Herausgeber:

NORDAKADEMIE gAG Hochschule der Wirtschaft
Köllner Chaussee 11
25337 Elmshorn

Gefördert durch die
NORDAKADEMIE-Stiftung

Grafiken: Mit freundlicher Unterstützung von Fritjof Wild

Inhaltsverzeichnis

Inhaltsverzeichnis ..... **IV**
Abbildungsverzeichnis ..... **VI**
Tabellenverzeichnis ..... **VII**
Vorwort zur Version 1.1 ..... **1**
Vorwort zur Version 1.0 ..... **2**

Teil A Allgemeine Überlegungen und Gesamtmodell **3**

1 Data Science **4**
2 Data Scientists **8**
3 Schlüsselbereiche zur Strukturierung der Aufgaben eines Data-Science-Projekts **13**
3.1 Grundlegende Anforderungen an Data-Science-Vorgehensmodelle 13
3.2 Vorgehensmodelle aus dem Bereich der Data Science 14
3.3 Schlüsselbereiche der Data Science 15
4 Data-Science-Vorgehensmodell DASC-PM **19**

Teil B Phasen im Modell **22**

5 Projektauftrag **24**
5.1 Merkmalstragender Bereich „Auslöser“ 26
5.2 Kernaufgabe „Use-Case-Entwicklung“ 27
5.3 Begleitende Aufgabe „Eignungsprüfung“ 32
5.4 Begleitende Aufgabe „Sicherstellung der Umsetzbarkeit“ 33
5.5 Kernaufgabe „Projektausgestaltung“ 34
5.6 Merkmalstragender Bereich „Projektskizze“ 35
6 Datenbereitstellung **36**
6.1 Merkmalstragender Bereich „Ursprungsdatenquellen“ 38
6.2 Kernaufgabe „Datenaufbereitung“ 40
6.3 Begleitende Aufgabe „Datenmanagement“ 42
6.4 Begleitende Aufgabe „Explorative Datenanalyse“ 43
6.5 Merkmalstragender Bereich „Analytische Datenquelle“ 45
7 Analyse **46**
7.1 Merkmalstragender Bereich „Analytische Datenquelle“ 49
7.2 Merkmalstragender Bereich „Anforderungen an Analyseverfahren“ 50
7.3 Kernaufgabe „Identifikation geeigneter Analyseverfahren“ 51
7.4 Kernaufgabe „Anwendung von Analyseverfahren“ 53
7.5 Begleitende Aufgabe „Werkzeugauswahl“ 55
7.6 Kernaufgabe „Entwicklung von Analyseverfahren“ 57
7.7 Begleitende Aufgabe „Evaluation“ 59
7.8 Merkmalstragender Bereich „Analyseergebnisse“ 61
8 Nutzbarmachung **62**
8.1 Merkmalstragender Bereich „Analyseergebnisse“ 64
8.2 Merkmalstragender Bereich „Analytische Datenquelle“ 64
8.3 Kernaufgabe „Technisch-methodische Bereitstellung“ 65
8.4 Begleitende Aufgabe „Sicherstellung technischer Umsetzbarkeit“ 67
8.5 Begleitende Aufgabe „Anwendbarkeitssicherstellung“ 69
8.6 Kernaufgabe „Fachliche Bereitstellung“ 70
8.7 Merkmalstragender Bereich „Analyseartefakte“ 71
9 Nutzung **72**
9.1 Merkmalstragender Bereich „Analyseartefakte“ 74
9.2 Begleitende Aufgabe „Monitoring“ 74
9.3 Merkmalstragender Bereich „Nutzungserkenntnisse“ 75

Teil C Übergreifende Schlüsselbereiche ... 76

10 Domäne ... 77

11 Wissenschaftlichkeit ... 78

12 IT-Infrastruktur ... 81

Teil D Schlussbemerkungen und Anhang ... 83

Schlussbemerkungen ... 84

Literatur ... 86

Verzeichnis der Autor:innen ... 88

Anhang ... 89

Abbildungsverzeichnis

Abbildung 1: Merkmale der Data Science 4
Abbildung 2: Nötige Kompetenzen von Data Scientists, in Anlehnung an Conway (2010) 8
Abbildung 3: Notwendige Kompetenzen in einem Data-Science-Projekt 9
Abbildung 4: Rollen in einem Data-Science-Projekt 9
Abbildung 5: Sieben Schlüsselbereiche der Data Science 15
Abbildung 6: Data-Science-Vorgehensmodell DASC-PM 19
Abbildung 7: Verwendete Nomenklatur und Notation in den Phasen 23
Abbildung 8: Kurzübersicht der Phase „Projektauftrag“ 24
Abbildung 9: Kompetenz- und Rollenprofil der Phase „Projektauftrag“ 24
Abbildung 10: Detaildarstellung der Phase „Projektauftrag“ 25
Abbildung 11: Kompetenz- und Rollenprofil der Aufgabe „Use-Case-Entwicklung“ 28
Abbildung 12: Kompetenz- und Rollenprofil der Aufgabe „Eignungsprüfung“ 32
Abbildung 13: Kompetenz- und Rollenprofil der Aufgabe „Sicherstellung der Umsetzbarkeit“ 33
Abbildung 14: Kompetenz- und Rollenprofil der Aufgabe „Projektausgestaltung“ 34
Abbildung 15: Kurzübersicht der Phase „Datenbereitstellung“ 36
Abbildung 16: Kompetenz- und Rollenprofil der Phase „Datenbereitstellung“ 36
Abbildung 17: Detaildarstellung der Phase „Datenbereitstellung“ 37
Abbildung 18: Kompetenz- und Rollenprofil der Aufgabe „Datenaufbereitung“ 41
Abbildung 19: Kompetenz- und Rollenprofil der Aufgabe „Datenmanagement“ 42
Abbildung 20: Kompetenz- und Rollenprofil der Aufgabe „Explorative Datenanalyse“ 44
Abbildung 21: Kurzübersicht der Phase „Analyse“ 46
Abbildung 22: Kompetenz- und Rollenprofil der Phase „Analyse“ 46
Abbildung 23: Detaildarstellung der Phase „Analyse“ 48
Abbildung 24: Kompetenz- und Rollenprofil der Aufgabe „Identifikation geeigneter Analyseverfahren“ 52
Abbildung 25: Kompetenz- und Rollenprofil der Aufgabe „Anwendung von Analyseverfahren“ 53
Abbildung 26: Kompetenz- und Rollenprofil der Aufgabe „Werkzeugauswahl“ 55
Abbildung 27: Kompetenz- und Rollenprofil der Aufgabe „Entwicklung von Analyseverfahren“ 58
Abbildung 28: Kompetenz- und Rollenprofil der Aufgabe „Evaluation“ 60
Abbildung 29: Kurzübersicht der Phase „Nutzbarmachung“ 62
Abbildung 30: Kompetenz- und Rollenprofil der Phase „Nutzbarmachung“ 62
Abbildung 31: Detaildarstellung der Phase „Nutzbarmachung“ 63
Abbildung 32: Formen der „Technisch-methodischen Bereitstellung“ 65
Abbildung 33: Kompetenz- und Rollenprofil der Aufgabe „Technisch-methodische Bereitstellung“ 66
Abbildung 34: Kompetenz- und Rollenprofil der Aufgabe „Sicherstellung technischer Umsetzbarkeit“ 68
Abbildung 35: Kompetenz- und Rollenprofil der Aufgabe „Anwendbarkeitssicherstellung“ 69
Abbildung 36: Kompetenz- und Rollenprofil der Aufgabe „Fachliche Bereitstellung“ 70
Abbildung 37: Kurzübersicht der Phase „Nutzung“ 72
Abbildung 38: Kompetenz- und Rollenprofil der Phase „Nutzung“ 72
Abbildung 39: Detaildarstellung der Phase „Nutzung“ 73

Tabellenverzeichnis

Tabelle 1: Beschreibung der Merkmale des Bereichs „Auslöser“ 26
Tabelle 2: Häufig genannte Teilaufgaben der Aufgabe „Use-Case-Entwicklung“ 27
Tabelle 3: Häufig genannte Vor- und Nachteile allgemeiner Methoden 29
Tabelle 4: Häufig genannte Vor- und Nachteile der Best Practices 31
Tabelle 5: Beschreibung der Teilaufgaben des Bereichs „Eignungsprüfung“ 32
Tabelle 6: Beschreibung der Teilaufgaben des Bereichs „Sicherstellung der Umsetzbarkeit“ 33
Tabelle 7: Beschreibung der Merkmalskategorien im Bereich Ursprungsdatenquellen 38
Tabelle 8: Häufig genannte Datenqualitätskriterien, aus Helfert et al. (2001) 39
Tabelle 9: Häufig genannte Teilaufgaben der Aufgabe „Datenaufbereitung“ 40
Tabelle 10: Häufig genannte Teilaufgaben der Aufgabe „Datenmanagement“ 42
Tabelle 11: Häufig genannte Teilaufgaben der Aufgabe „Explorative Datenanalyse“ 43
Tabelle 12: Häufig genannte Merkmale des Bereichs „Anforderungen an Analyseverfahren“ 50
Tabelle 13: Häufig genannte Teilaufgaben der Aufgabe „Identifikation geeigneter Analyseverfahren“ 51
Tabelle 14: Häufig genannte Teilaufgaben der Aufgabe „Anwendung von Analyseverfahren“ 54
Tabelle 15: Häufig genannte Teilaufgaben der Aufgabe „Werkzeugauswahl“ 56
Tabelle 16: Häufig genannte Teilaufgaben der Aufgabe „Entwicklung von Analyseverfahren“ 57
Tabelle 17: Häufig genannte Teilaufgaben der Aufgabe „Evaluation“ 59
Tabelle 18: Häufig genannte Merkmale des Bereichs „Analyseergebnisse“ 61
Tabelle 19: Häufig genannte Teilaufgaben der Aufgabe „Technisch-methodische Bereitstellung“ 66
Tabelle 20: Häufig genannte Teilaufgaben der Aufgabe „Sicherstellung technischer Umsetzbarkeit“ 67
Tabelle 21: Häufig genannte Teilaufgaben der Aufgabe „Anwendbarkeitssicherstellung“ 69
Tabelle 22: Häufig genannte Teilaufgaben der Aufgabe „Fachliche Bereitstellung“ 70
Tabelle 23: Häufig genannte Merkmale des Bereichs „Analyseartefakte“ 71
Tabelle 24: Häufig genannte Teilaufgaben der Aufgabe „Monitoring“ 74
Tabelle 25: Häufig genannte Merkmale des Bereichs „Nutzungserkenntnisse“ 75

Vorwort zur Version 1.1

Im Februar 2020 erschien mit dem Data Science Process Modell (DASC-PM) die erste Version eines umfassenden Vorgehensmodells für Data-Science-Projekte. Die vielen positiven Rückmeldungen, die wir erhalten haben, zeigen uns, dass wir den erhofften Beitrag zur Diskussion rund um Data-Science-Aktivitäten leisten konnten. Das DASC-PM hat in den letzten zwei Jahren seinen Weg in die Praxis, in Buchbeiträge (z.B. Alekozai et al., 2021) und auf wissenschaftliche Konferenzen (z.B. Schulz et al., 2020) gefunden.

Wir möchten uns an dieser Stelle herzlich bei allen Leserinnen und Lesern bedanken, die ihre Erfahrungen mit uns geteilt haben, die uns Stärken und Verbesserungspotenziale des Modells aufgezeigt haben, und besonders natürlich bei denen, die aktiv an der Weiterentwicklung mitgearbeitet haben. Ohne alle diese Personen wäre der Weg zur jetzt vorliegenden Version 1.1 nicht möglich gewesen.

Mit dieser Version greifen wir zahlreiche Rückmeldungen aus Praxis und Wissenschaft auf sowie einige Themen, die uns besonders am Herzen liegen. So haben wir die Lesbarkeit des gesamten Dokuments durch eine stringentere Strukturierung und kürzere Einführungstexte erhöht. Das Modell selbst stellt nun klarer heraus, was Schlüsselbereiche und Phasen sind, was sie kennzeichnet und wie ihr Zusammenspiel in unterschiedlichen, auch agilen, Projektkonstellationen aussehen kann. Alle verwendeten Begrifflichkeiten haben wir dokumentübergreifend mit kritischem Auge geprüft und, wo notwendig, angepasst und vereinheitlicht. Dabei haben wir auch Vorschläge für eine weniger formale und in der Praxis eingängigere Visualisierung aufgegriffen und sowohl das Dokument als auch das eigentliche Modell in eine grafisch ansprechendere Form überführt (so hoffen wir zumindest). Eine mögliche geringfügige Reduktion der wissenschaftlichen Exaktheit der Gesamtmodelldarstellung erscheint uns vor dem Hintergrund, dass das DASC-PM „von vielen für viele“ geschaffen ist, als angemessener Trade-off zugunsten einer erhöhten Zugänglichkeit.

Inhaltlich haben wir uns bei der Version 1.1 insbesondere auf die Phase des Projektauftrags fokussiert. Wichtige Entscheidungen und Rahmenbedingungen werden zu Beginn von Data-Science-Aktivitäten festgelegt. Hierfür bieten wir mit einer umfassenderen Beschreibung der Phase und einem praktisch anwendbaren Fragenkatalog jetzt eine konkrete Basis sowohl für neue als auch erfahrene Anwender der Data Science an. Genau wie in der Version 1.0 sind die Ergebnisse als Zusammenführung der Erfahrung sämtlicher Teilnehmerinnen und Teilnehmer dieser Arbeitsgruppe zu verstehen. Für die nächsten Monate ist zudem erstmalig auch eine englischsprachige Veröffentlichung des Hauptdokuments vorgesehen, sodass die der Data Science innewohnende Interdisziplinarität auch in internationalen Projekten leichter durch das DASC-PM unterstützt werden kann.

Alle vorgestellten Ergebnisse des DASC-PM basieren weiterhin maßgeblich auf den Rückmeldungen einer breit aufgestellten Arbeitsgruppe und stellen einen Diskussionsstand dar, der Anregung und Hilfestellung sein soll, aber nie den Anspruch erhebt, das lebendige Feld der Data Science abschließend zu erfassen. Wir freuen uns, dass uns diese Lebendigkeit auch weiterhin motivieren wird, das DASC-PM zu diskutieren, zu verändern und einem breiten Publikum zugänglich zu machen. Wenn Sie Interesse an der Mitarbeit haben oder von uns über aktuelle Entwicklungen rund um das Modell informiert werden möchten, melden Sie sich gerne unter der unten angegebenen Kontaktadresse.

Elmshorn, Halle (Saale), Hamburg, Krefeld, Mönchengladbach und Stuttgart im März 2022

DASC-PM-Kernteam

Kontakt: info@dasc-pm.org

Vorwort zur Version 1.0

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

Teil A
Allgemeine Überlegungen und Gesamtmodell

1 Data Science

Trotz gestiegener Aufmerksamkeit fehlt derzeit eine allgemein akzeptierte und einheitliche Definition der Data Science. Während der Begriff in wissenschaftlichen Veröffentlichungen vieler Disziplinen bereits in unterschiedlichem Verständnis verwendet wird, fallen Definitionen aus der Praxis noch stärker durch ihre Heterogenität auf. Dieser Sachverhalt wiederum führt zu sehr unterschiedlichen Erwartungen und möglichen Missverständnissen bei den beteiligten Personengruppen.

Von den Teilnehmerinnen und Teilnehmern der Arbeitsgruppe wurden bei der Erstellung von DASC-PM v1.0 wiederholt zwei Definitionen bzw. Zusammenfassungen zentraler Aspekte der Data Science genannt: van der Aalst (2016) und Provost & Fawcett (2013). Beide führen allerdings nur einige, aber nicht alle Aspekte auf, die von den Mitgliedern der Arbeitsgruppe als relevant für eine umfassende Data-Science-Definition genannt wurden. Sie betrachten zudem Aspekte unterschiedlichen Detailgrads.

Auf Basis der Teilnehmerbeiträge unserer Arbeitsgruppe empfehlen wir daher 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.***

Abbildung 1 stellt die Merkmale dieser Data-Science-Definition vor, die im Folgenden genauer betrachtet werden.

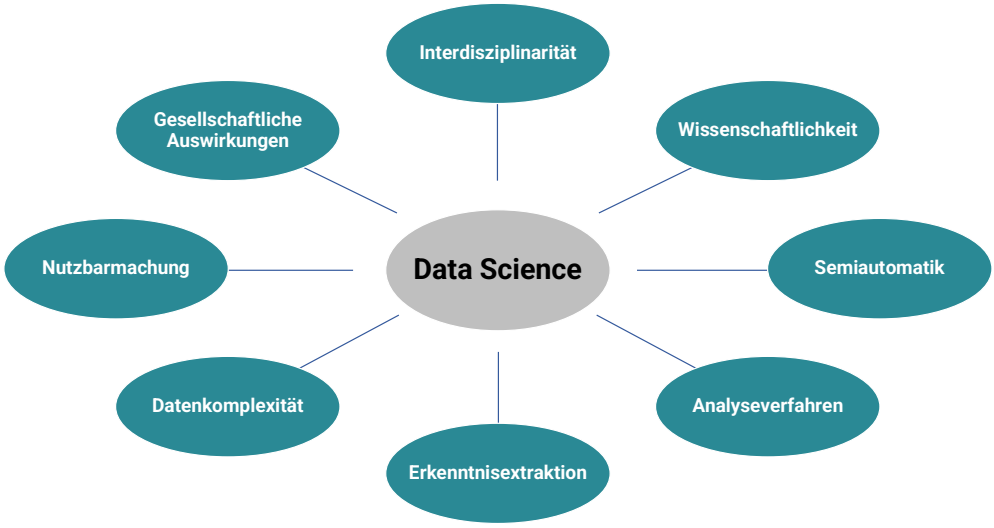


Abbildung 1: Merkmale der Data Science

Interdisziplinarität

Die Interdisziplinarität wird als 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. größere, häufig unstrukturierte oder sich dynamisch ändernde 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. Dass in vielen Texten zu diesem Thema das Domänenumfeld traditionell auf die Wirtschaft beschränkt wird, ist 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 bereits vorzufinden.

Wissenschaftlichkeit

Der Begriff *Data Science* 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 z. T. auch 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 ausgewählten 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 Business-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) und *Cross Industry Standard Process for Data Mining* (CRISP-DM) (Wirth & Hipp, 2000) begründet, deren Anwendung zumindest im Business-Kontext 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.

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.

Analyseverfahren

Die explizite Nennung einzelner Algorithmen(-gruppen) in einer Data-Science-Definition ist einem Fachgebiet mit einer hohen Entwicklungsgeschwindigkeit nicht angemessen. Sie würde zudem eine implizite Einschränkung des Fachgebietes vornehmen, auf die an dieser Stelle verzichtet wird. Zudem ist der Begriff des Algorithmus an sich im gegebenen Zusammenhang nicht adäquat, da nicht alle Analysen tatsächlich Algorithmen-basiert sind und die in Frage kommenden Algorithmen wiederum nicht notwendigerweise auf die Data Science beschränkt oder ihr zugehörig sein müssen.

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. Hierbei können der Zweck von Data Science das Aufdecken von Mustern, Trends und Zusammenhängen sowie die Optimierung sein.

Abhängig von den gegebenen Use Cases (Anwendungsfällen / Problemstellungen) können bestehende Analyseverfahren verwendet werden. Es kann jedoch auch nötig sein, Analyseverfahren weiterzuentwickeln oder vollständig neu zu entwickeln, da keine geeigneten Ansätze existieren.

Semiautomatik

Die Anwendung von Analyseverfahren erfolgt semi-automatisch, 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. Dafür sind aber vorbereitende manuelle Arbeitsschritte notwendig. Auch ist ein Erkenntnisgewinn letztendlich nur durch menschliche Beteiligung zu erreichen.

Erkenntnisextraktion und Datenkomplexität

Ein Ziel der Data Science ist die Extraktion von Erkenntnissen aus meist komplexen Daten. Diese unterscheiden sich in ihrer Struktur, ihrer Qualität, ihrer Vollständigkeit, ihrer Größe und ihrer Dimensionalität. Es kann sich um statische Daten oder Datenströme handeln, außerdem können Daten in komplexen Beziehungen zueinanderstehen.

Die Entwicklung der Data Science ist stark durch den Anstieg verfügbarer Datenmengen getrieben (Dhar, 2013). Die Analyse sehr großer und heterogener Datenbestände erforderte die Schaffung neuer Verfahren, die häufig unter dem Begriff *Big Data* aufgeführt werden. Data Science ist allerdings 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.

Nutzbarmachung

Data Science beinhaltet nicht nur die Extraktion von Erkenntnissen, sondern zusätzlich auch deren Nutzbarmachung. 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 automatisierten 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 in Form eines Software-Systems 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.

Gesellschaftliche Auswirkungen

Die Auseinandersetzung mit den gesellschaftlichen Auswirkungen der Data Science mit Hilfe 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 Data Scientists

Artikel wie der von Davenport und Patil (2012), in dem der Beruf des Data Scientists mit *The Sexiest Job of the 21st Century* betitelt wird, können den Anschein erwecken, dass alle auf diesem Gebiet nötigen Kompetenzen in einer einzelnen Person vereint sein können, beziehungsweise müssen. Diese Sichtweise wurde in vielen Publikationen übernommen, ist aber problematisch. Eine Übersicht zu existierenden Data-Scientist-Definitionen ist in Chatfield et al. (2014) zu finden.

In verschiedenen Quellen werden aufbauend auf dem Beitrag Conways (2010) von einem Data Scientist Kompetenzen in drei Bereichen gefordert:

- *Mathematisch-statistisches Wissen*
- *Informationstechnisches Wissen*
- *Anwendungsspezifisches Wissen*

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 2 – legt dies durch die geringen Überschneidungen der einzelnen Fachgebiete nahe.

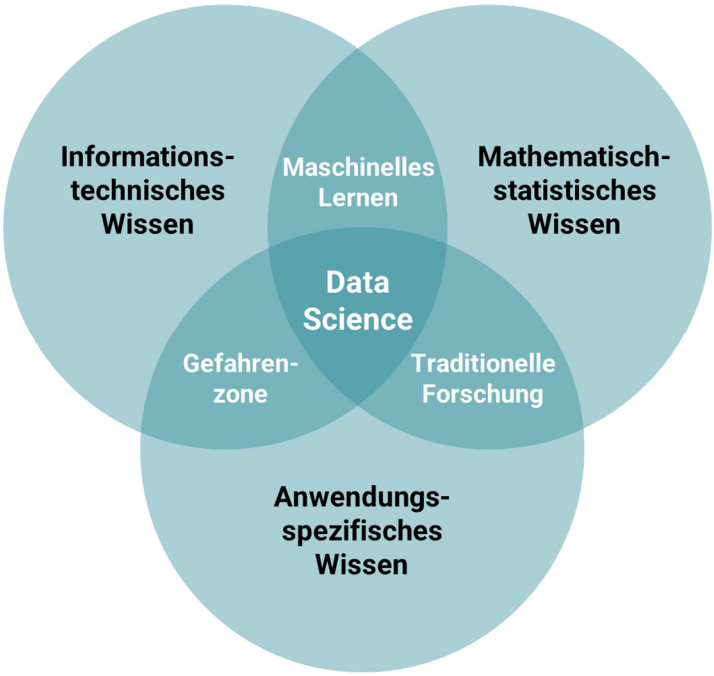


Abbildung 2: Nötige Kompetenzen von Data Scientists, in Anlehnung an Conway (2010)

Eine oberflächliche Kenntnis ist jedoch i. d. R. nicht ausreichend: Abhängig vom konkreten Anwendungsfall können vertiefte Kompetenzen in einem oder mehreren der genannten drei Bereiche nötig sein. Data Scientists müssen zudem in der Lage sein,

- *Kommunikation mit allen Anspruchsgruppen in einer geeigneten Sprache zu gestalten (Davenport & Patil, 2012),*
- *das Management eines Data-Science-Projekts zu übernehmen und*
- *die strategische Einordnung von Aktivitäten vorzunehmen.*

Abbildung 3 fasst sämtliche Kompetenzen zusammen, die für die Durchführung von Data-Science-Projekten benötigt werden.

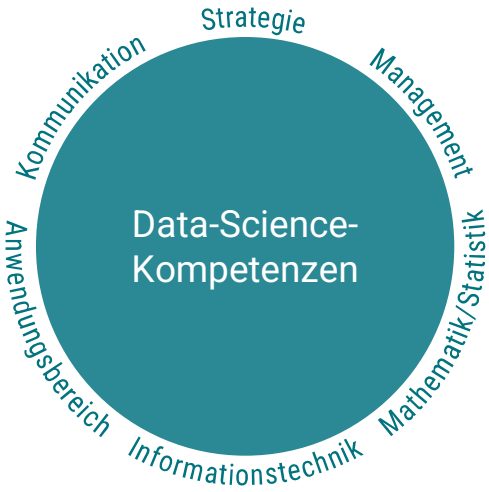


Abbildung 3: Notwendige Kompetenzen in einem Data-Science-Projekt

Für eine einzelne Person ist es in der Regel nicht möglich, weitreichende Fähigkeiten in allen genannten Bereichen aufzubauen (Zschech et al., 2018). Data Scientists können sich daher entweder in einer Disziplin bzw. in wenigen Disziplinen spezialisieren oder übergeordnete bzw. weniger datenorientierte Rollen übernehmen.

Bei der Spezialisierung 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. Bei großen Projekten werden diese Rollen häufig noch in Unterrollen aufgeteilt, was in der Grafik der Übersicht halber nicht weiter ausgeführt wird. Entsprechende Hinweise finden sich bei den nachfolgenden Beschreibungen. Rollen und Personen müssen dabei nicht deckungsgleich sein. Mehrere Personen können eine Rolle ausfüllen, eine Person kann mehrere Rollen übernehmen. In Abbildung 4 sind die Überlegungen zu Rollen zusammengefasst. Es werden dabei vier Kernrollen und zwei ergänzende Rollen unterschieden.

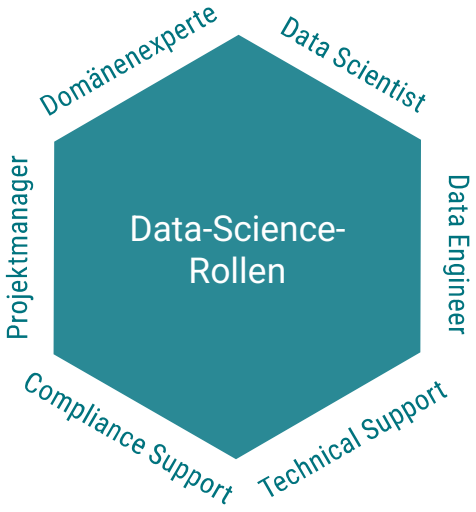


Abbildung 4: Rollen in einem Data-Science-Projekt

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 sind Data Scientists zuständig für die Durchführung aller Aspekte eines Data-Science-Projekts. Sie arbeiten mit Domänenexperten zusammen, sind aber für alle methodischen, technischen und organisatorischen Fragen verantwortlich. Obwohl dieses Verständnis des Data Scientists 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.

In diesem Dokument wird im Weiteren unter *Data Scientist* immer eine spezifischer definierte Rolle verstanden:

> ***Im spezifischeren Sinne versteht man unter Data Scientists Spezialisten für den Analysebereich eines Data-Science-Projekts, die insbesondere für die Auswahl der Analysemethoden und -werkzeuge, die Durchführung von Analysen und die Interpretation der Ergebnisse zuständig sind.***
>
> ***In den übrigen Bereichen eines Data-Science-Projekts sind sie nur beratend tätig. Weitere Aufgaben müssen dementsprechend von anderen Rollen übernommen werden.***

Die Rolle des 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 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 im spezifischen Sinne, 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 und Prognoseerstellung tätig. Hierbei bedient er 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 dabei relevanten Parametern 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 domänenspezifisches Wissen, 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.

Kernrolle „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.

Kernrolle „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 bzw. des Anwendungsfalls. Domänenexperten können Prioritäten für zu modellierende oder analysierende Aspekte setzen und sind Bindeglieder zu den methodischen und technischen Experten.

Innerhalb der Domänenexperten kann es wieder Unterrollen geben. Im Business-Kontext häufig anzutreffen sind etwa die *Business Developer*, welche den, einem Projekt zugrundeliegenden domänenspezifischen Use Case 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.

Kernrolle „Projektmanager“

Projektmanager planen, steuern und koordinieren den Gesamtablauf eines Data-Science-Projekts. Dazu benötigen sie – 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.

Ergänzende Rolle „Technischer Support“

Neben den vier 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.

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 *IT Infrastructure Architect*, verantwortlich für den Entwurf einer geeigneten IT-Infrastruktur für das Projekt, und *IT-Techniker/IT-Administratoren*, welche die benötigte Hard- und Software bereitstellen sowie die zugrundeliegenden Systeme konfigurieren. Aber auch Anwendungsentwickler, die sich mit der Implementierung von Anwendungssoftware/-werkzeugen zur produktiven Nutzung der Analyseergebnisse befassen, werden hier dem technischen Support zugeordnet.

Ergänzende 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.

3 Schlüsselbereiche zur Strukturierung der Aufgaben eines Data-Science-Projekts

In diesem Kapitel werden zunächst die grundlegenden Anforderungen an Data-Science-Vorgehensmodelle betrachtet und mit Erfahrungen weiterer Modelle kombiniert. Abschließend werden Schlüsselbereiche von Data-Science-Projekten herausgearbeitet, die zur Strukturierung eines Data-Science-Vorgehensmodells verwendet werden können.

3.1 Grundlegende Anforderungen an Data-Science-Vorgehensmodelle

Generell soll durch die Verwendung eines Vorgehensmodells die Qualität von Data-Science-Projekten erhöht werden. Das Durchlaufen sämtlicher Schritte – von der Projektkonzeption bis zur Nutzung 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 wurden. Dadurch kann eine Reproduzierbarkeit, Wiederverwendbarkeit und Generalisierbarkeit der Ergebnisse sichergestellt werden. Zudem muss das Vorgehensmodell skalierbar sein, um Projekte unterschiedlicher Größe zu unterstützen. Hierzu wird eine Unterscheidung zwischen auszuführenden Projektaktivitäten und qualitativen Anforderungen an die Projektkoordination und -organisation vorgenommen.

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 wichtig ist, als auch die Verständlichkeit, was wiederum die Akzeptanz des Modells gefährdet.

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 Anwendung können irrelevante Modellbausteine somit übersprungen werden. Alternativ zu einer Modularisierung können spezialisierte Varianten des Vorgehensmodells entstehen – abhängig von der betrachteten Domäne und/oder den eingesetzten Analyseverfahren.

Für die Erarbeitung des Vorgehensmodells in diesem Dokument ergeben sich damit folgende Rahmenbedingungen:

- ***Abstraktionsebene***
  Es wird zunächst ein Modell auf hoher Abstraktionsebene erarbeitet, das für den breiten Einsatz in Data-Science-Projekten geeignet ist. Der Fokus der Arbeitsgruppe liegt initial nicht auf

der abschließenden Festlegung und Beschreibung aller Details. Vielmehr sollen durch den Einsatz des Modells in realen Projekten Erkenntnisse gewonnen und Probleme identifiziert werden, die im Rahmen eines kontinuierlichen Weiterentwicklungsprozesses Berücksichtigung finden.

- ***Rollenbilder und Kommunikation***
  Gerade in großen Projekten sollen die Projektbeteiligten 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.

- ***Berücksichtigung der Projektarbeit und des Projektumfelds***
  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.

3.2 Vorgehensmodelle aus dem Bereich der Data Science

Zusätzlich zu den bereits in Kapitel 1 aufgeführten Vorgehensmodellen KDD und CRISP-DM, deren zugrundeliegende Logik eine konzeptionelle Verbundenheit zur Data Science aufweist sind weitere verwandte Modelle zu betrachten, die speziell für den Data-Science-Bereich entwickelt wurden. Vor allem im Zusammenhang mit dem CRISP-DM sind aber auch vielversprechende Bestrebungen zu identifizieren, die dieses Modell an die Anforderungen von Data-Science-Projekten anpassen sollen, z.B. das CRISP-ML(Q) (Studer et al., 2021).

Der KDD-Prozess weist eine klare Verständlichkeit und einen eindeutigen Arbeitsweg von der Datenquelle hin zum erlangten Wissen auf. Rücksprünge sind möglich, werden aber nicht im Sinne eines iterativen Vorgehens gefordert. Der Prozess ist 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 fokussiert berücksichtigt.

Beim CRISP-DM wird die Möglichkeit zur Iteration der einzelnen Prozessschritte innerhalb eines Analyseprojekts stärker deutlich. Das Modell bleibt dennoch einfach und klar verständlich. Die schwächer ausgeprägte Differenzierung der einzelnen Prozessschritte erfordert eine engere Zusammenarbeit der beteiligten Personengruppen bei geringerer Abgrenzung der einzelnen Aufgaben. Das Modell enthält neben daten-, analyse- und auswertungsbezogenen Prozessschritten mit dem *Business Understanding* einen Prozessschritt mit sehr starkem Fokus auf der Domäne. CRISP-DM wurde aus der Industrie heraus entwickelt. Der Prozessschritt der Nutzbarmachung ist vorhanden, allerdings nur wenig ausgeprägt, was im Hinblick auf die heute gängigen datengetriebenen Produkte und Dienstleistungen problematisch sein kann.

Ein weiteres Modell ist der *Team Data Science Process* (TDSP), der von der Firma Microsoft entwickelt und publiziert 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 auf Data-Science-Projekte im Business-Kontext und beinhaltet viele Aspekte, die sich auch in den anderen, älteren Vorgehensmodellen wiederfinden. Er expliziert, im Vergleich zu den genannten Modellen, Rollen und zugeordnete Aufgaben und ergänzt die Bereiche *IT-Infrastruktur* und *Customer Acceptance*. Durch Letztere wird die Domänenrelevanz stärker hervorgehoben als im CRISP-DM, fokussiert aber auch hier auf den Unternehmenskontext. Das Modell ist mit Bezug auf die Bereiche der Data Science vollständiger als die zuvor genannten, definiert den Ablauf der Prozessschritte aber nicht immer ausführlich. In der Dokumentation wird häufig – dem Ursprung des Modells mit der damit verbundenen Besetzung eines Marktsegments geschuldet – auf Microsoft-Technologien verwiesen.

3.3 Schlüsselbereiche der Data Science

Unter Berücksichtigung der bisherigen Überlegungen und genannten Vorgehensmodelle lassen sich Schlüsselbereiche identifizieren, auf deren Basis strukturiert ein Vorgehensmodell abgeleitet werden kann. Basierend auf der kritischen Diskussion von KDD, CRISP-DM und TDSP können sieben Schlüsselbereiche der Data Science abgeleitet werden, deren allgemeiner Zusammenhang in Abbildung 5 dargestellt ist.

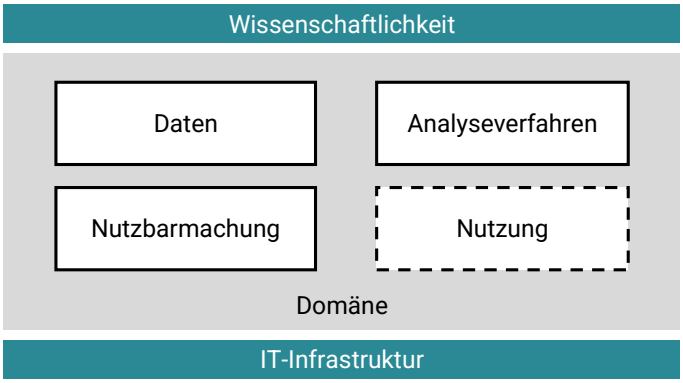


Abbildung 5: Sieben Schlüsselbereiche der Data Science

Im Zentrum stehen die Bereiche *Daten* und *Analyseverfahren*, die im Rahmen unterschiedlicher Prozessschritte auch 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 dagegen 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. Um der ambivalenten Rolle der Nutzung in Data-Science-Projekten Rechnung zu tragen, wird sie in der Abbildung den vorherigen drei Schlüsselbereichen gegenüber grafisch abgegrenzt und stärker mit der Domäne verknüpft.

Im Schlüsselbereich *Domäne* sind die vier zuvor genannten Schlüsselbereiche eingebettet. Dieser Bereich ist, unter Berücksichtigung der Definition des Data-Science-Begriffes, nicht, wie bei anderen Vorgehensmodellen, auf den Business-Kontext beschränkt.

Die bisher beschriebenen Bereiche werden durch den übergreifenden Schlüsselbereich *Wissenschaftlichkeit* flankiert, der bisher in keinem Vorgehensmodell explizite Berücksichtigung findet, jedoch einen Kernbestandteil der Data-Science-Disziplin bildet.

Der begleitende Schlüsselbereich *IT-Infrastruktur* spielt in vielen Data-Science-Projekten eine zunehmend wichtigere Rolle (siehe 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 und wie 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“

Die Anwendung eines geeigneten Datenanalyseverfahrens ist der zentrale Schritt im Data-Science-Prozess, da diese (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 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 allgemein Methoden aus dem Bereich *Künstliche Intelligenz*. Stehen keine passenden Analyseverfahren zur Verfügung, so müssen bestehende Verfahren angepasst oder sogar neue Verfahren entwickelt werden. Vorbereitende Schritte sind die verfahrensspezifische Datenaufbereitung und die Merkmalskonstruktion. Eine weitere Aufgabe in diesem Schlüsselbereich ist die Evaluierung.

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“). Die Nutzung wird daher als Schlüsselbereich eines umfassenden Data-Science-Vorgehensmodells betrachtet.

Schlüsselbereich „Domäne“

Die vier zentralen Schlüsselbereiche können nur im Kontext der Domäne, also des Anwendungsfeldes, konkretisiert werden. Ein breites Hintergrundwissen auf diesem Anwendungsfeld ist an vielen Stellen des Data-Science-Prozesses relevant, z. B. 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 „Wissenschaftlichkeit“

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.

Im Forschungs-, aber auch im Business-Kontext gilt es zudem, eine angemessene Arbeitsweise sicherzustellen, etwa das Aufstellen einer Hypothese, die Evaluation angewandter Methoden im Hinblick auf Angemessenheit und Effizienz, die Überprüfung der Validität der Ergebnisse, die Sicherstellung ihrer Reproduzierbarkeit und das fundierte 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 oder die Evaluation und Nutzbarmachung der Analyseergebnisse werden mit Hilfe von spezialisierten Software-Produkten umgesetzt. Die Bandbreite und Komplexität dieser Produkte sind insbesondere bei größeren Projekten 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 Data-Science-Vorgehensmodell DASC-PM

Aufbauend auf den zuvor beschriebenen Ausarbeitungen wird in diesem Kapitel das Data-Science-Process-Model (DASC-PM) als Vorgehensmodell für Data-Science-Projekte eingeführt – siehe Abbildung 6. Die Visualisierung des DASC-PM leitet sich dabei aus den zuvor identifizierten Schlüsselbereichen eines Data-Science-Projekts und auch aus den bereits angedeuteten Abhängigkeiten zwischen ihnen ab.

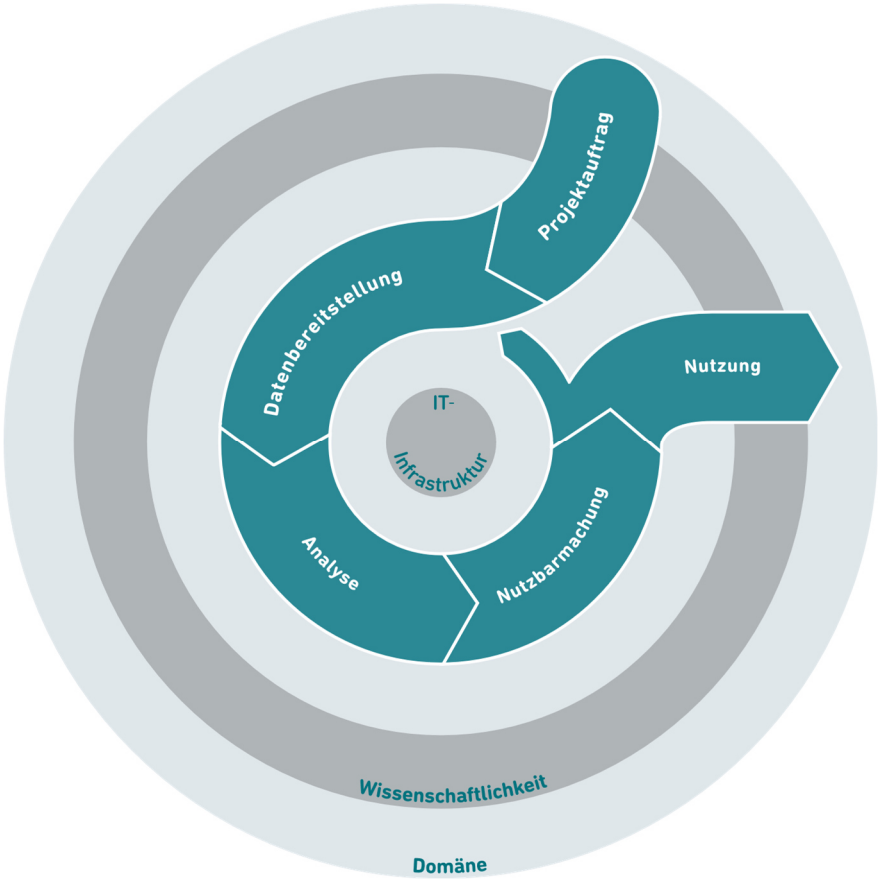


Abbildung 6: Data-Science-Vorgehensmodell DASC-PM

Die im Modell dargestellten Pfeile zeigen den primären Pfad bei der Verwendung des DASC-PM.
Sie stellen einzelne Phasen dar, die sich aus den vier zentralen Schlüsselbereichen ableiten, sowie den Projektauftrag als phasenähnliche Aktivität aus dem übergreifenden Schlüsselbereich Domäne.
Unter Nutzung von IT-Infrastruktur und Berücksichtigung von Wissenschaftlichkeit können Projektphasen mehrfach durchlaufen werden, bevor eine umfangreiche Nutzung außerhalb des eigentlichen Data-Science-Projektes erfolgt und/oder zu einem neuen/veränderten Projektauftrag führt.

Die zuvor vorgestellten Schlüsselbereiche gehen im vorgestellten Modell auf. Die vier zentralen Schlüsselbereiche und die aus ihnen abgeleiteten Aufgaben werden dabei in iterativ durchlaufbare Phasen überführt. Die drei übergreifenden Schlüsselbereiche formen den Rahmen dieses so entstehenden Prozesses. Eine Sonderstellung nimmt der Schlüsselbereich *Domäne* ein, der sowohl den weit umfassenden Rahmen des Projektes bildet als auch Aufgaben in sich trägt, die phasenartig zu betrachten sind und daher hier als Projektauftrag separat in das Modell eingefasst werden. Die folgende Aufstellung fasst die Bestandteile des DASC-PM prägnant zusammen.

Phasen

- ***Projektauftrag***
  Innerhalb einer Domäne bestehende Probleme lösen eine Use-Case-Entwicklung aus. Die vielversprechendsten Use Cases werden anschließend zu einer Data-Science-Projektskizze ausgestaltet. Alle zugehörigen Aufgaben finden sich in der Phase des Projektauftrags wieder. Durch die frühe, relativ umfassende Betrachtung des Projekts sind hier häufig auch umfassende Fähigkeiten in fast allen Kompetenzbereichen erforderlich.
- ***Datenbereitstellung***
  Innerhalb der Phase der Datenbereitstellung werden alle Aktivitäten zusammengefasst, die dem Schlüsselbereich Daten zuzuordnen sind, weshalb der verwendete Begriff weit gefasst ist. Die Phase beinhaltet die Datenaufbereitung (von der Erfassung bis zur Speicherung), das Datenmanagement und eine explorative Analyse. Als Ergebnis dieser Phase entsteht eine für die weitere Analyse geeignete Datenquelle.
- ***Analyse***
  In einem Data-Science-Projekt können entweder bestehende Verfahren angewendet oder zunächst neue Verfahren entwickelt werden – die entsprechende Entscheidung ist eine eigene Herausforderung. Die Phase umfasst daher nicht nur die Analysedurchführung, sondern auch angrenzende Tätigkeiten. Das Artefakt der Phase ist ein Analyseergebnis, das eine methodische und fachliche Evaluation durchlaufen hat.
- ***Nutzbarmachung***
  In der Phase der Nutzbarmachung wird eine anwendbare Form der Analyseergebnisse geschaffen. Projektspezifisch kann dies entweder eine umfangreiche Betrachtung technischer, methodischer und fachlicher Aufgaben nach sich ziehen oder eher pragmatisch gehandhabt werden. Die Analyseartefakte können sowohl Resultate als auch Modelle oder Verfahren selbst umfassen und werden den Adressaten in unterschiedlicher Form zur Verfügung gestellt.
- ***Nutzung***
  Die sich an die Projektdurchführung anschließende Verwendung von Analyseartefakten ist nicht als primärer Teil eines Data-Science-Projekts anzusehen. Ein Monitoring ist aber abhängig von der 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 (auch im Sinne iterativer Vorgehensweisen) zu erlangen.

Übergreifende Schlüsselbereiche

- ***Domäne***
  Neben dem Projektauftrag als Hauptbestandteil stellen die expliziten Anforderungen oder Umstände der anderen Phasen häufig domänenspezifische Rahmenbedingungen dar, welche die dortigen Aufgaben beeinflussen. Die Domäne muss daher durchgängig berücksichtigt werden.

- ***Wissenschaftlichkeit***
  Die Wissenschaftlichkeit erhebt in Data-Science-Projekten keinen allgemeinen Anspruch auf ein vollständig formalisiertes akademisches und durchweg forschungsorientiertes Vorgehen. Während dies im Kontext von Forschungsprojekten durchaus so sein kann, bezieht sich der Aspekt der Wissenschaftlichkeit im Business-Kontext vor allem auf eine saubere Methodik, wie sie typischerweise als Eigenschaft bzw. Mindestanforderung wissenschaftlichen Arbeitens erwartet wird.
  Der definierte Projektauftrag ist in jeder einzelnen Projektphase entsprechend methodisch 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.
- ***IT-Infrastruktur***
  Sämtliche Schritte, die ein Data-Science-Projekt durchlaufen muss, sind von der zu Grunde liegenden IT-Infrastruktur abhängig; das tatsächliche Ausmaß der IT-Unterstützung ist allerdings projektindividuell zu bewerten. Auch wenn die Nutzung spezifischer Hard- und Software häufig bereits organisationsintern festgelegt ist, sollten, die limitierenden und befähigenden Merkmale der IT-Infrastruktur (oder ggf. auch die Möglichkeit der Infrastrukturerweiterung) in sämtlichen Projektphasen berücksichtigt werden.

Iterationen und Abbruch bei der Modellnutzung

Obwohl aus Gründen der Übersichtlichkeit auf eine Visualisierung dieser Tatsache in Abbildung 6 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 für den Aufbau eines Verständnisses des betrachteten Problems bzw. der betrachteten Probleme hilfreich sein.

Sofern Iterationen vorgesehen sind, müssen die einzelnen Phasen nur insoweit neu durchlaufen werden, wie es der jeweiligen Iteration erforderlich und förderlich ist. So kann beispielsweise eine Nutzbarmachung und (projektinterne) Nutzung eines erstellten Modells dazu führen, dass, unter Auslassen einer erneuten Datenbereitstellung, die Analyse-Phase neu und ggf. in geringerem Umfang durchlaufen wird, um das Modell zu verbessern. Im Projektverlauf sind selbstverständlich auch Rücksprünge zwischen den Phasen möglich, sofern das Projektteam sie für notwendig hält.

Teil B
Phasen im Modell

Hinweise zu Teil B

In den folgenden Kapiteln werden die einzelnen Phasen des DASC-PM detailliert betrachtet. Die Darstellungen der Phasen und die ihnen zugeordneten Aufgaben basieren dabei auf der Expertise der Teilnehmerinnen und Teilnehmer der Arbeitsgruppe. Ein Anspruch auf Vollständigkeit der beschriebenen Inhalte je Phase besteht nicht, vielmehr soll die Ausarbeitung dazu dienen, den Leserinnen und Lesern ein Gefühl für relevante Aspekte innerhalb der verschiedenen Projektschritte zu vermitteln.

Einführend zeigt Abbildung 7 die verwendete Nomenklatur sowohl als strukturierte Übersicht, als auch in einer Beispielnotation für die Detaildarstellung, die in den einzelnen Kapiteln verwendet wird, um die Zusammenhänge darzustellen. Die in der Übersicht kursiv dargestellten Begriffe werden im Rahmen der jeweiligen Phase näher beschrieben; die fett markierten Begriffe je Phase in Unterkapiteln erläutert.

Schlüsselbereich

Phase
- Bereich
  - ***Merkmalstragend***
  - Aufgabe
    - ***Kernaufgabe***
    - ***Begleitende Aufgabe***
- *Schnittstellen(aufgabe)*

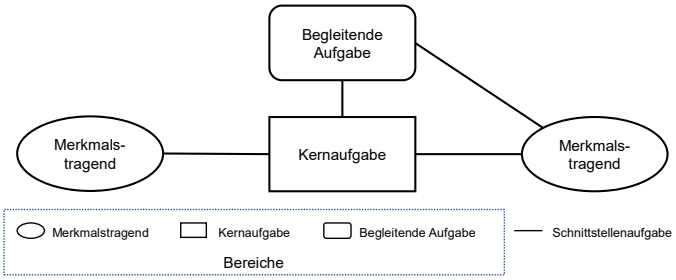


Abbildung 7: Verwendete Nomenklatur und Notation in den Phasen

Alle Kapitel beginnen mit einer einführenden und zusammenfassenden Übersichtsseite, auf der in vereinfachter Form die einzelnen Bereiche der jeweiligen Phase aufgeführt sind. Unterschieden werden dabei merkmalstragende und aufgabentragende Bereiche. Letztere werden weiterhin unterschieden in Kernaufgaben des Bereichs und begleitende Aufgaben. Die verbindenden Kanten stellen Schnittstellen dar und beinhalten ebenfalls Aufgaben, die durch die Verzahnung der einzelnen Bereiche entstehen.

Die einführende Übersicht je Phase wird ergänzt durch die kombinierte Darstellung der ermittelten notwendigen Kompetenz- und Rollenprofile, die in den einzelnen Bereichen der Phase zum Tragen kommen. Diese Darstellungen basieren auf Abbildung 3 und Abbildung 4, wie sie in Kapitel 2 hergeleitet wurden. In der ersten Darstellung wird in Form eines Netzdiagramms das Kompetenzprofil von Personen dargestellt, die sich im jeweiligen Aufgabenbereich spezialisieren, der zweiten Darstellung ist die Rollenrelevanz für den jeweiligen Aufgabenbereich zu entnehmen. Die Darstellungen sind dabei durch eine Aggregation der Rückmeldungen von Teilnehmerinnen und Teilnehmern entstanden und sollen eine ungefähre Einschätzung des Kompetenzprofils für ein typisches Data-Science-Projekt ermöglichen, die Ausprägungen können sich im individuellen Projekt und Kontext aber stark davon unterscheiden.

Im Anschluss folgt je Kapitel eine detailliertere Darstellung der Zusammenhänge zwischen den Bereichen der Phase und ihren Schnittstellen. Merkmale, Kernaufgaben und begleitende Aufgaben werden darauffolgend in jeweils einem eigenen Unterkapitel detailliert.

5 Projektauftrag

Innerhalb einer Domäne bestehende Probleme lösen eine Use-Case-Entwicklung aus. Die vielversprechendsten Use Cases werden anschließend zu einer Data-Science-Projektskizze ausgestaltet. Alle zugehörigen Aufgaben finden sich in der Phase des Projektauftrags wieder. Durch die frühe und relativ umfassende Betrachtung des Projekts sind hier häufig auch umfassende Fähigkeiten in fast allen Kompetenzbereichen erforderlich.

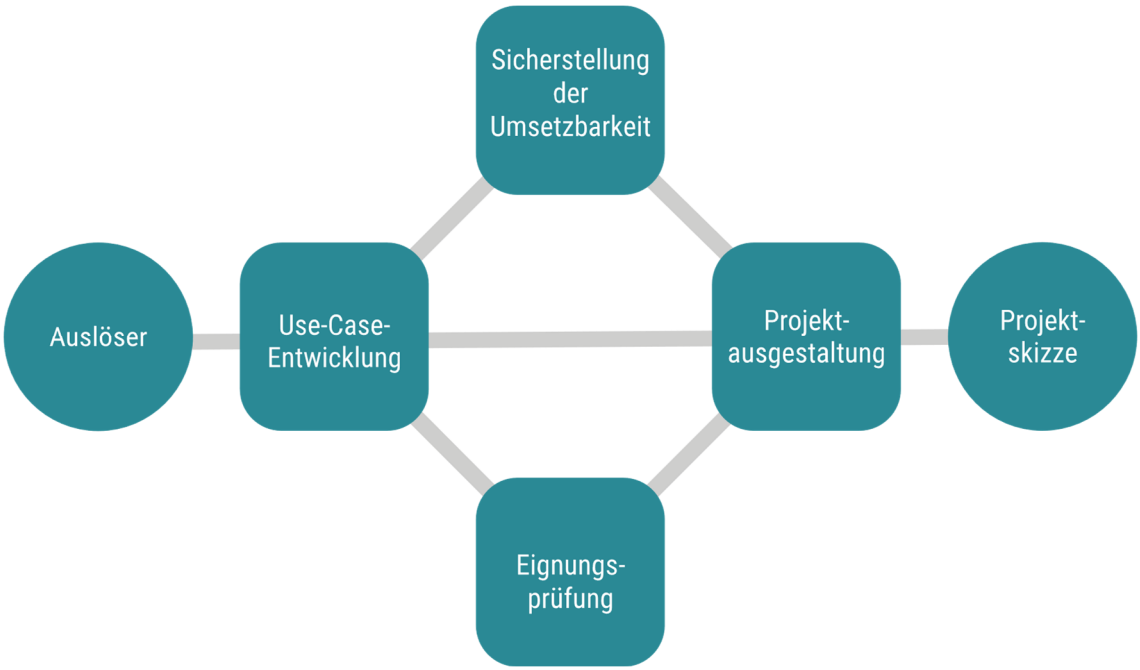


Abbildung 8: Kurzübersicht der Phase „Projektauftrag“

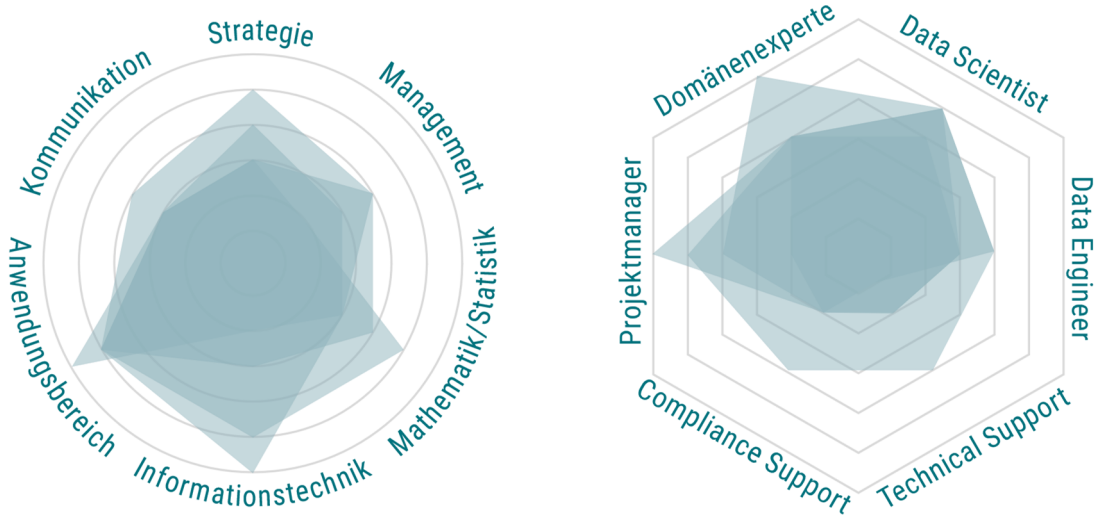
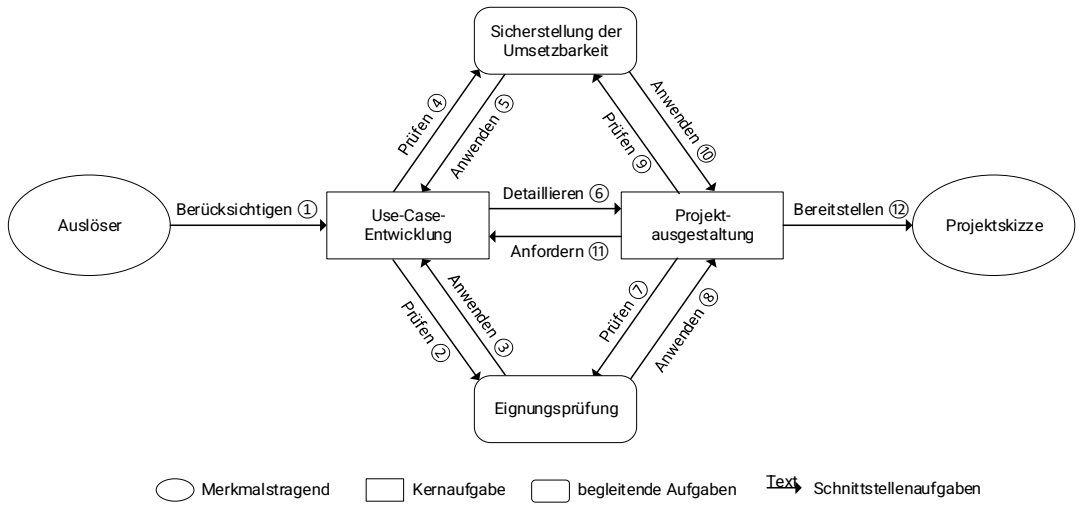


Abbildung 9: Kompetenz- und Rollenprofil der Phase „Projektauftrag“

Detaildarstellung der Phase Projektauftrag



①	In der Domäne auftretende Ereignisse lösen die Entwicklung von Use Cases aus.
②	Die Eignung des Use Cases für die Behandlung in einem Data-Science-Projekt ist zu prüfen.
③	Die Ergebnisse der Eignungsprüfung sind bei der Use-Case-Entwicklung zu berücksichtigen.
④	Bei der Use-Case-Entwicklung ist die Umsetzbarkeit im Zuge eines Projektes sicherzustellen.
⑤	Werden besondere Anforderungen identifiziert, die die Use-Case-Entwicklung beeinflussen, sind diese zu berücksichtigen.
⑥	Ausgewählte Use Cases sind zu einem Projekt auszugestalten.
⑦	Die Eignung des ausgewählten Use Cases für die Behandlung in einem Data-Science-Projekt ist zu prüfen.
⑧	Die Ergebnisse der Eignungsprüfung sind bei der Projektausgestaltung zu berücksichtigen.
⑨	Die Umsetzbarkeit des ausgewählten Use Cases in einem Projekt ist sicherzustellen.
⑩	Werden besondere Anforderungen identifiziert, welche die Projektausgestaltung beeinflussen, sind diese zu berücksichtigen.
⑪	Sollte bei der Ausgestaltung erkannt werden, dass eine Projektdurchführung nicht sinnvoll ist, wird zur Use-Case-Entwicklung zurückgekehrt.
⑫	Eine Projektskizze ist als Ergebnis dieses Teilprozesses bereitzustellen.

Abbildung 10: Detaildarstellung der Phase „Projektauftrag“

5.1 Merkmalstragender Bereich „Auslöser“

Die Initialisierung eines Data-Science-Projektes wird durch ein Ereignis, meist ein Problem, in der Domäne ausgelöst. Im forschenden oder explorativen Kontext kann es sich jedoch auch um offene Fragestellungen handeln, die nicht per se ein ‚Problem‘ im üblichen Sprachgebrauch darstellen.

In Tabelle 1 werden von Teilnehmerinnen und Teilnehmern häufig genannte Merkmale von Auslösern aufgeführt und beschrieben.

Tabelle 1: Beschreibung der Merkmale des Bereichs „Auslöser“

Merkmal	**Beschreibung**
Ziel	Es ist zu entscheiden, wie die Ergebnisse eines Data-Science-Projekts, das durch ein Problem ausgelöst wird, 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 zu erarbeitenden 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.
Beteiligte Bereiche	Die domänenseitig an der Projektdurchführung beteiligten Bereiche sind zu benennen.
Fachliche Domäne	Eine Beschreibung der fachlichen Domäne, innerhalb derer die Probleme zu bearbeiten sind, 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 ganzen Konzerns?
Komplexität	Erst die Einschätzung der Komplexität der betrachteten Probleme 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 auslösenden Probleme zu einem oder mehreren Use Cases ist in der Regel erst in den Folgeschritten möglich, weshalb an dieser Stelle keine besonderen formalen Anforderungen an Formulierungen oder Dokumentationsarten gestellt werden. Auch ein spezifisches Abstraktionsniveau der Beschreibungen muss nicht vorgegeben werden, da sich Auslöser stark voneinander unterscheiden können.

5.2 Kernaufgabe „Use-Case-Entwicklung“

Ein häufig dargestelltes Problem bei der ersten Beschäftigung mit der Data-Science-Disziplin in Organisationen ist die Identifikation und Auswahl praktikabler Use Cases. Ein Use Case wird in diesem Dokument definiert als eine in sich geschlossene Einheit, die jedoch weiter in Arbeitspakete zerlegt werden kann. Ein Projekt kann aus mehreren Use Cases bestehen, wobei jeder Use Case einen eindeutigen Nutzen stiftet. Im Idealfall wird ein Use Case durch eine Iteration des DASC-PM bearbeitet, allerdings kann es auch nötig sein, dass mehrere Iterationen für einen Use Case durchlaufen werden müssen.

In Tabelle 2 werden von Teilnehmerinnen und Teilnehmern häufig genannte Teilaufgaben der Use-Case-Entwicklung aufgeführt und beschrieben.

Tabelle 2: Häufig genannte Teilaufgaben der Aufgabe „Use-Case-Entwicklung“

Teilaufgabe	Beschreibung
Aufbau eines Verständnisses für die Disziplin	Die Erwartungen an die Data-Science-Disziplin decken sich oftmals nicht mit deren Möglichkeiten. Zudem fehlt es Data-Science-Initiativen häufig an einem klaren Fokus. Ein Verständnis für die Disziplin ist aufzubauen.
Use-Case-Identifikation	Zum einen kann das Herunterbrechen von Organisationszielen in Use Cases, die innerhalb eines Projektes betrachtet werden können, eine Herausforderung darstellen. Zum anderen müssen sich Use Cases aus Anforderungen und Problemen in einer Organisation ergeben. Oft bleibt unklar, welche Ergebnisse bei der Umsetzung der jeweiligen Use Cases zu erwarten sind. Die Aufgabe besteht darin, sowohl kreative als auch realisierbare Ideen zu entwickeln.
Use-Case-Priorisierung	Die Auswahl der geeigneten Use Cases für die Umsetzung ist teilweise nicht möglich, da Potenziale für die eigene Organisation vor der Projektdurchführung ggf. nicht direkt ersichtlich sind. Es fehlen oder existieren möglicherweise nur ungenaue Inputdaten und (Zwischen-)Metriken zur Bewertung alternativer Use Cases, bspw. auf Basis von (Kapital-)Renditen. Es ist daher möglich, dass sich Anstrengungen zur Priorisierung von Use Cases nicht rentieren, da sie sich später als falsch herausstellen könnten.
Abstimmung beteiligter Personengruppen	Die Personengruppen in einer Organisation, denen durch die Umsetzung von Use Cases in Form von Projekten geholfen werden kann, wissen häufig nicht, welchen Nutzen die Data Science für sie erbringen kann. Die Datenspezialisten dagegen erkennen ggf. nicht die relevantesten Use Cases einer Organisation. Die Kommunikation und Abstimmung zwischen diesen beiden Personengruppen sind daher von hoher Relevanz.

Auf die Frage nach den Personengruppen oder Abteilungen in einer Organisation, welche die Entwicklung von Problemen hin zu Use Cases vorantreiben sollten, gibt es aus Sicht der Teilnehmerinnen und Teilnehmern keine eindeutige Antwort. Domänenexperten und Data Scientists wurden am häufigsten genannt, aber auch andere Personengruppen sollten berücksichtigt werden.

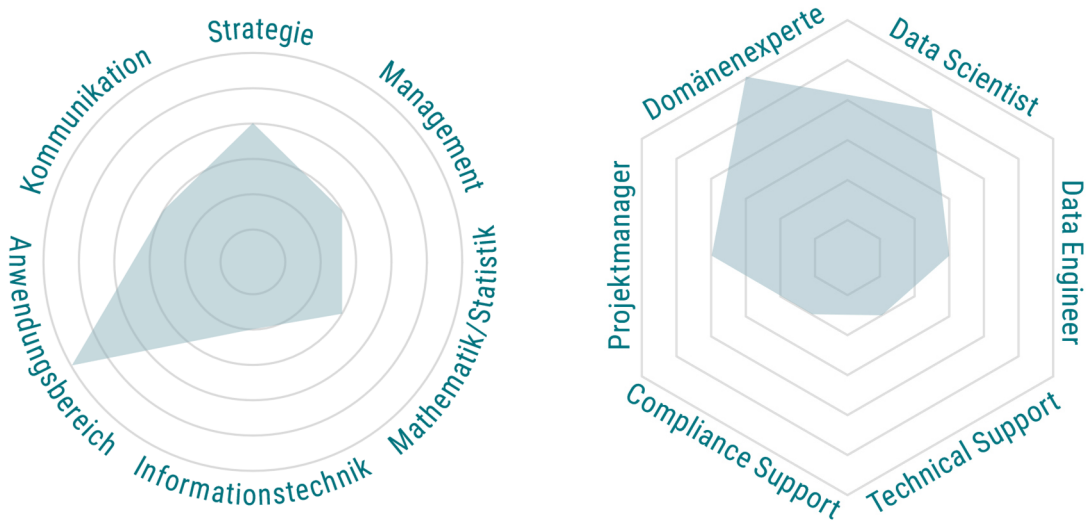


Abbildung 11: Kompetenz- und Rollenprofil der Aufgabe „Use-Case-Entwicklung“

Bei der Auswahl von geeigneten Use Cases ist es wichtig, eine vertrauensvolle Basis zwischen den beteiligten Personengruppen zu schaffen. Domänenexperten sollten offen von schwierigen, aufwendigen oder herausfordernden Aufgaben berichten können, ohne dabei durch vorgreifende Lösungsvorschläge beeinflusst zu werden. Empfehlenswert sind vorbereitende bilaterale Gespräche, um die Rahmenbedingungen abzuklopfen, gute Beziehungen zu den Ansprechpartnern aufzubauen und Interviews bzw. Workshops vorzubereiten. Die gewählten Methoden müssen dabei an das gegebene Umfeld der Organisation angepasst werden.

Die beiden von den Teilnehmerinnen und Teilnehmern am häufigsten genannten Formate zur Identifikation geeigneter Use Cases sind Interviews and Workshops. Interviews eignen sich dabei eher für kleinere Bereiche mit wenigen Beteiligten. Die Interviews sollten dabei von erfahrenden Interviewern auf Augenhöhe durchgeführt werden. Workshops eignen sich auch für größere Gruppen und Bereiche. Workshops mit vielen Teilnehmern werden insbesondere zur Ideensammlung eingesetzt. Die Konkretisierung und Priorisierung von Use Cases kann anschließend in kleineren Workshops durchgeführt werden. Um möglichst alle Facetten beleuchten zu können, sollte das Teilnehmerfeld möglichst heterogen sein, also verschiedene Bereiche (bspw. Domänenexperten, Data Scientists, Entscheider) und Senioritäts-Level abdecken. Mögliche Punkte auf der Agenda eines solchen Workshops sind die Vorstellung von exemplarischen Use Cases, ein gemeinsames Brainstorming und die Bewertung von Vorschlägen nach Relevanz und Umsetzbarkeit. Gegebenenfalls ist auch eine Unterfütterung mit ausgewählter Data-Science-Theorie angebracht. Ergebnis der Workshops sollte die Auswahl und Ausgestaltung eines möglichst konkreten Use Cases sein. Es sollte erkennbar sein, was durch die Betrachtung des Use Cases in einem Projekt konkret erreicht werden soll und welcher geschäftliche Nutzen (Business Value/Impact) dadurch entstehen könnte.

Zur Ausgestaltung der Workshops können unterschiedliche Methoden eingesetzt werden. Generell sollten die verwendeten Methoden dem Workshop eine Grundstruktur und einen roten Faden geben, gleichzeitig aber auch Raum für flexible und kreative Elemente lassen. Durch eine solche Grundstruktur können die Workshop-Teilnehmer bspw. auf einen ähnlichen Wissensstand gebracht oder behandelte Fragestellungen auf einer gemeinsamen wissenschaftlichen Basis beleuchtet werden. Die kreativen Elemente können die Motivation steigern, unterschiedliche Denk- und Herangehensweisen stimulieren sowie die Gruppeninteraktion auch zwischen sehr unterschiedlichen Teilnehmerinnen und Teilnehmern fördern.

Zur Identifikation und Auswahl geeigneter Data Science Use Cases haben sich allgemeine Methoden wie Fokusgruppen, Fish Bowls, Design Thinking oder Hackathons bewehrt. Darüber hinaus existieren aber auch speziellere, auf den Kontext von *Data Science*, *Artificial Intelligence*, *Machine Learning* oder *Big Data* zugeschnittene Methoden, etwa der Enterprise AI Canvas (Kerzel, 2021), der Machine Learning Canvas (Dorard, 2015) oder das von Bill Schmarzo (2015) beschriebene Vorgehen. Wichtig ist zu beachten, dass die ausgewählten Methoden zur Fragestellung und zum jeweiligen Personenkreis passen müssen (Stichwort ‚Akzeptanz', insbesondere bei ‚exotischen' Methoden).

In Tabelle 3 werden die von den Teilnehmerinnen und Teilnehmern am häufigsten genannten Vor- und Nachteile allgemeiner Methoden.

Tabelle 3: Häufig genannte Vor- und Nachteile allgemeiner Methoden

Vorteile	**Nachteile**
Methode: **Fokusgruppen**	
• *Hilfreich bei der Identifikation von Use Cases* • *Hilfreich bei der Priorisierung von Use Cases* • *Liefern häufig schnell Ergebnisse* • *Nützlich zur Vereinheitlichung von Perspektiven und Wissensständen der beteiligten Gruppen* • *Einbezug von unterschiedlichen Ansichten/Meinungen* • *Fördern die Entwicklung neuer Ideen und geben Raum für spontane Einfälle* • *Fokusgruppen können sich im Idealfall zu Task Forces entwickeln*	• *Erfordert erfahrenen, neutralen Moderator (dieser ist daher selbst kein zentraler Ideengeber)* • *Gute Strukturierung zur Erreichung der Ergebnisse notwendig (z. B. durch Orientierung an Leitfragen), da sonst Gefahr der Ziellosigkeit besteht* • *Teilnehmerzahl begrenzt (bei mehr als zwölf Personen häufig keine fokussierte Diskussion mehr möglich)* • *Hohe Anforderung an Gruppenzusammensetzung (Diversität und dennoch vergleichbare Wissensstände, da sonst einzelne Teilnehmer die Diskussion dominieren können). Ergebnisse häufig stark abhängig von der Zusammensetzung der Gruppe.* • *Ideen stellen häufig nur eine Momentaufnahme dar und können tiefergehende Folgeüberlegungen erfordern* • *Strukturierte Nachbereitung und Analyse (zusammenfassendes Transskript, finaler Report usw.) aufwendig*
Methode: **Fish Bowl (Innen-Außenkreis-Methode)**	
• *Eine größere Personengruppe kann sich beteiligen (aktiver im Innenkreis, passiver im Außenkreis)* • *Dynamischer Wechsel der Diskussionsteilnehmer je nach Thema, Perspektive und Kompetenzen möglich* • *Geringerer Druck auf die Teilnehmer, da der Innenkreis jederzeit verlassen werden kann* • *Gut geeignet, um die Identifikation von Use Cases durch den Innenkreis gezielt voranzutreiben*	• *Ein Kern motivierter Personen für den Innenkreis notwendig* • *Kleinere Gruppen im Innenkreis können die Diskussion prägen, wodurch die Diversität der Diskussion leidet* • *Personen im Außenkreis trauen sich ggf. nicht in den Innenkreis zu wechseln (insbesondere bei physischen Treffen).*

Vorteile	Nachteile
Methode: **Design Thinking**	
• *Fokus auf die Perspektive des Endanwenders schafft größtmöglichen Nutzen und Akzeptanz* • *Interdisziplinäre Teams ermöglichen Berücksichtigung diverser Aspekte*	• *Design Thinking ist eher eine kreative Herangehensweise zur Gestaltung von Produkten für eine gewisse Zielgruppe. Nutzung zur Use-Case-Erarbeitung im Data-Science-Kontext nur in speziellen Fällen (z. B. Entwicklung eines Management-Dashboards) ohne Anpassung möglich.*
Methode: **Hackathons**	
• *Realisierung eines Prototyps oder Minimum Viable Product (MVP) in kurzer Zeit* • *Fördern tiefgehende Diskussionen über Lösungsansätze* • *Heterogene Arbeitsgruppen möglich* • *Ergebnis häufig sehr konkret und nutzbar* • *Wettbewerb fördert Anreiz zur Entwicklung innovativer Lösungen* • *Mögliche (unterstützende) Realisierungsform für Quick Wins/Proof of Concepts*	• *Hoher organisatorischer Aufwand* • *Gefahr der Fokussierung auf konkreten Use Case; breitere Sichtweise kommt ggf. zu kurz* • *Fokus auf der technischen Umsetzung und daher ggf. Vernachlässigung weiterreichender Aspekte (wie z. B. Datenbeschaffung, Compliance, gesellschaftliche Konsequenzen)*

Zusätzlich werden von den Teilnehmerinnen und Teilnehmern der Arbeitsgruppe folgende Best Practices zur Auswahl geeigneter Data Science Use Cases empfohlen:

- *Durch eine **Analyse der Organisations-/Bereichsstrategie** können strategisch passende Use Cases identifiziert werden, die dadurch eine höhere Aufmerksamkeit und Unterstützung erfahren.*
- ***Quick Wins** sind Use Cases, die mit geringem Aufwand und realistischen Erfolgschancen einen konkreten Unternehmensnutzen stiften. Durch Quick Wins lassen sich häufig anfängliche Skeptiker überzeugen und Ressourcen für nachfolgende, aufwändigere Projekte sichern.*
- ***Leuchtturmprojekte** sind mit viel Energie und Aufwand erstellte Vorzeigeprojekte. Sie sollen verdeutlichen, was durch ein gut geplantes und durchgeführtes Data-Science-Projekt erreicht werden kann. Leuchtturmprojekte sollen eine hohe Sichtbarkeit entfalten, anfängliche Skeptiker bspw. durch Testimonials überzeugen und somit andere Abteilungen/Organisationseinheiten zur Nachahmung animieren. Teile des Leuchtturmprojekts können für Folgeprojekte idealerweise angepasst und wiederverwendet werden.*
- *Data Science Use Cases, bei denen die technische Umsetzbarkeit fraglich ist, können mit einem **Proof of Concept** begonnen werden. Dabei wird mit einem vordefinierten und i. d. R. geringen Ressourceneinsatz (bspw. Zeit, Personal) versucht, die generelle Erreichbarkeit des angestrebten Ziels unter den gegebenen Rahmenbedingungen (existierende Daten, verfügbare Methoden, gegebene IT-Infrastruktur usw.) zu belegen.*

In Tabelle 4 werden die von den Teilnehmerinnen und Teilnehmern die am häufigsten genannten Vor- und Nachteile von Best Practices aufgeführt.

Tabelle 4: Häufig genannte Vor- und Nachteile der Best Practices

Vorteile	**Nachteile**
Best Practice: **Ausrichtung an Organisations-/Bereichsstrategie**	
• *Ausrichtung grundsätzlich immer empfehlenswert, zumindest aber sollten Konflikte mit der Organisationsstrategie vermieden werden* • *Hilfreich zur Maximierung des Nutzens* • *Mehr Aufmerksamkeit für das Projekt bei der Geschäfts- bzw. Bereichsleitung (je nach Strategieebene).*	• *Als einziges Kriterium u. U. problematisch, da dadurch sehr komplexe und aufwendige Projekte fokussiert werden könnten.* • *Gegebenenfalls schwierig, Unterstützung aus einzelnen Zielgruppen zu erhalten*
Best Practice: ***Quick Wins***	
• *Nutzstiftende Ergebnisse mit wenig Aufwand* • *Gut geeignet, um Aufmerksamkeit/Unterstützung für Folgeprojekte zu erhalten* • *Demonstriert pragmatisches, ressourcenschonendes Handeln* • *Schnelle Erfolge motivieren auch das Projektteam selbst*	• *Begrenzte Anzahl von Themen, die als Quick Wins realisiert werden können* • *Fokus auf schnelle Zielerreichung führt ggf. zur Vernachlässigung anderer Aspekte (z. B. Datenqualität, Einheitlichkeit der Datenbasis, Benutzerfreundlichkeit)* • *Gefahr der Entstehung von Datensilos/Insellösungen, da ganzheitliche Lösung zu aufwendig*
Best Practice: **Leuchtturmprojekt**	
• *Verdeutlichung der durch Data Science erreichbaren Ergebnisse und Vorteile* • *Viel Aufmerksamkeit auch über Organisationsgrenzen hinweg* • *Können „blueprints" liefern, wie Data Science Projekte optimal durchgeführt werden*	• *Hoher Aufwand (dieser wird nach außen aber nur begrenzt ersichtlich)* • *Hoher Ressourcenbedarf (Zeit, Geld) kann negativ wirken, falls der entstehende Nutzen nicht hoch genug ist*
Best Practice: **Proof of Concept**	
• *Schnelle und ressourcensparende Entwicklung eines Prototyps* • *Aufschlussreich für weitere Entscheidungen und Entwicklungen*	• *Verleitet zu „Quick-and-Dirty"-Lösungen, die später u. U. dennoch im produktiven Einsatz landen, da sie (zumindest in Grundzügen) funktionieren*

5.3 Begleitende Aufgabe „Eignungsprüfung“

Ziel der Eignungsprüfung ist es, zu entscheiden, ob sich die identifizierten und später ausgewählten Use Cases erfolgreich in einem Projekt umsetzen lassen. Dafür ist zu prüfen, ob die festgelegten Anforderungen unter Verwendung der vorhandenen Ressourcen erfüllt werden können. In Tabelle 5 werden von Teilnehmerinnen und Teilnehmern häufig genannte Teilaufgaben der Eignungsprüfung aufgeführt und beschrieben. Wird der betrachtete Use Case für die Umsetzung in einem Projekt ausgewählt, erfolgt eine detailliertere Prüfung in der Phase der Projektausgestaltung (vgl. Abschnitt 5.5).

Tabelle 5: Beschreibung der Teilaufgaben des Bereichs „Eignungsprüfung“

Teilaufgabe	Beschreibung
Eignung des Use Cases	Es ist zu prüfen, ob es sich tatsächlich um einen Use Case handelt, bei dem der Einsatz von Data Science als geeignet erscheint.
Eignung der Methode	Es ist zu prüfen, ob Analyseverfahren existieren oder entwickelt werden können, die mit angemessener Wahrscheinlichkeit ein geeignetes Ergebnis erzielen. Hierfür sind ggf. erste Tests durchzuführen.
Bewertung der Datengrundlage	Es ist häufig unklar, welche Daten für Data-Science-Projekte verfügbar sind bzw. beschafft werden können, in welcher Qualität sie vorliegen und inwiefern sie sich für die Verwendung in Analysen eignen. Auch der Aufwand der Datenaufbereitung und der tatsächliche Nutzen von Daten kann im Vorfeld häufig nur schwer eingeschätzt werden. Eine erste Bewertung der Datengrundlage in diesem frühen Stadium ist zwingend erforderlich.
Eignung des Ziels	Ein Abgleich der erwarteten Projektergebnisse mit dem betrachteten Use Case ist durchzuführen.
Berücksichtigung früherer Projekte	Ein Abgleich von früher bereits durchgeführten Projekten mit dem aktuell betrachteten Use Case ist durchzuführen.
Priorisierung des Use Cases	Unter Berücksichtigung knapper Ressourcen ist zu prüfen, ob die Berücksichtigung des Use Cases sinnvoll ist oder ob anderen Problemen Vorzug gegeben werden sollte.

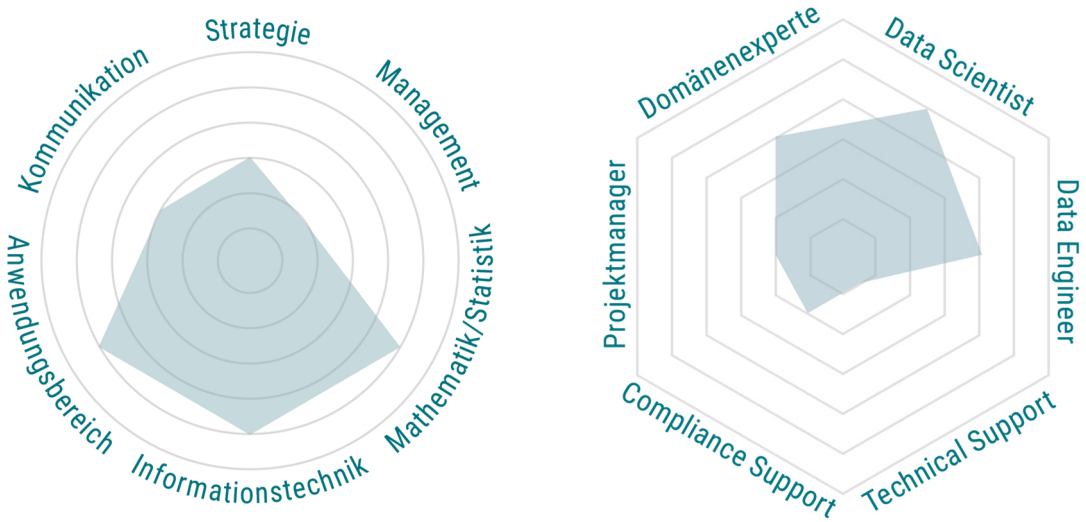


Abbildung 12: Kompetenz- und Rollenprofil der Aufgabe „Eignungsprüfung“

5.4 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 einigen Fällen kann bei der Umsetzung von Use Cases eine große Unsicherheit darin bestehen, ob die Untersuchungen zu Erkenntnissen führen und wie diese aussehen könnten. In Organisationen existiert häufig ein hoher Anteil an implizitem Wissen, bei dem zu Projektbeginn unklar ist, wie dieses in Analyseartefakten abgebildet werden kann. Hinzu kommen rechtliche Aspekte, die mögliche Use Cases einschränken. Hier kennen sich die zuständigen Ansprechpartner z. T. nicht gut aus und sind ggf. übervorsichtig. Auch der Aufwand für die Umsetzung von Use Cases und die notwendigen Ressourcen zur erfolgreichen Durchführung des Data-Science-Projektes werden häufig unterschätzt. In Tabelle 6 werden die von den Teilnehmerinnen und Teilnehmern am häufigsten genannten Teilaufgaben zur Sicherstellung der Umsetzbarkeit aufgeführt und beschrieben. Wird der betrachtete Use Case für die Umsetzung in einem Projekt ausgewählt, erfolgt eine detailliertere Prüfung in der Phase der Projektausgestaltung (vgl. Abschnitt 5.5).

Tabelle 6: Beschreibung der Teilaufgaben des Bereichs „Sicherstellung der Umsetzbarkeit“

Teilaufgabe	Beschreibung
Prüfung der IT-Infrastruktur	Es ist zu prüfen, ob die vorhandene IT-Infrastruktur dazu geeignet ist, den betrachteten Use Case 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 die Umsetzung des betrachteten Use Cases zu prüfen.
Risikoeinschätzung	Das Risiko bei der Umsetzung des Use Cases im Zuge eines Projektes ist einzuschätzen (Eintrittswahrscheinlichkeiten des Risikos, Schwere der Konsequenzen).
Kosten-Nutzen-Analyse	Eine Analyse des Nutzens ist zwar häufig nur sehr schwer durchzuführen, die Kosten sollten aber grundsätzlich bewertet werden.

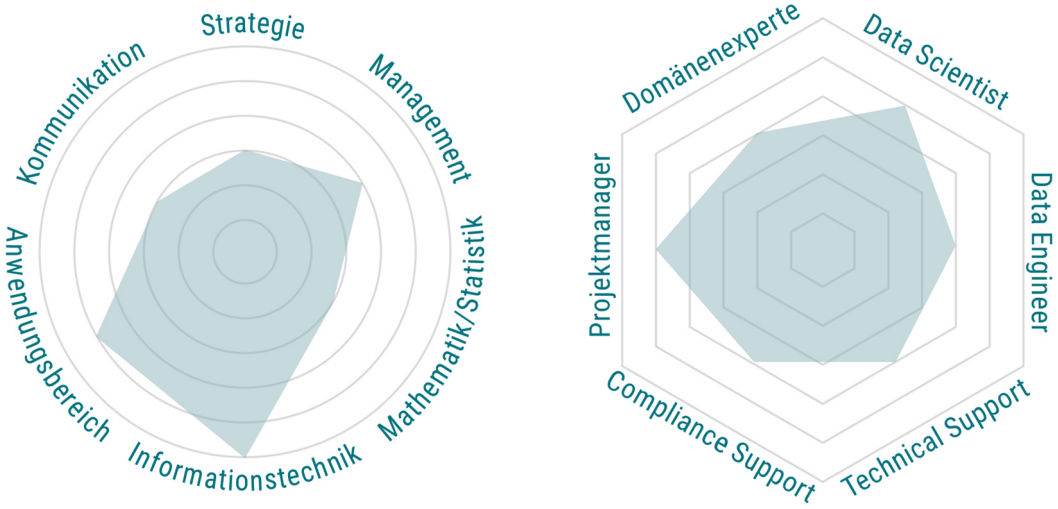


Abbildung 13: Kompetenz- und Rollenprofil der Aufgabe „Sicherstellung der Umsetzbarkeit“

5.5 Kernaufgabe „Projektausgestaltung“

Ziel der Projektausgestaltung ist es, die notwendigen Arbeitsschritte zu bestimmen, die zur Erfüllung der Anforderungen führen, die durch den jeweiligen Use Case spezifiziert werden. Dies muss auf Basis der Informationen über die Datengrundlage und unter Miteinbeziehung der Domänenspezifika geschehen. Da sich die grundlegenden Merkmale des Projektmanagements kaum von denen anderer Projekte unterscheiden, sei an dieser Stelle auf zugehörige Standardliteratur verwiesen. Data-Science-Vorhaben besitzen allerdings auch ganz spezifische Projektmerkmale. Da ihr Projekterfolg häufig schlechter abschätzbar ist als bei Vorhaben in anderen Bereichen, müssen sie intensiv betrachtet werden. Unter Umständen muss bei dieser Betrachtung zwischen explorativen Forschungs- und Entwicklungsprojekten und solchen Projekten, die konkret auf eine Umsetzung bzw. einen Regelbetrieb abzielen, unterschieden werden.

Eine Hilfestellung zur intensiven Betrachtung spezifischer Data-Science-Projektmerkmale bietet der Fragebogen im Anhang. Wird durch dessen Bearbeitung erkannt, dass der ausgewählte Use Case nur eingeschränkt als geeignet bewertet werden kann, um im Zuge eines Projektes umgesetzt zu werden, ist eine Anpassung vorzunehmen (vgl. Abschnitt 5.2). Teile des Fragebogens können während der Projektdurchführung wiederholt betrachtet werden – immer dann, wenn neue Informationen zur Verfügung stehen.

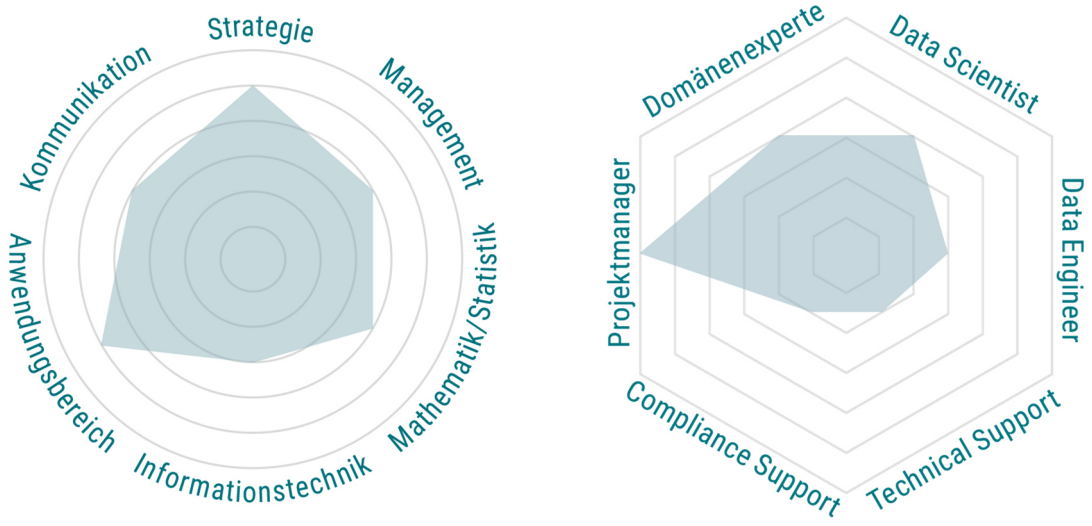


Abbildung 14: Kompetenz- und Rollenprofil der Aufgabe „Projektausgestaltung“

5.6 Merkmalstragender Bereich „Projektskizze“

Im Gegensatz zu vielen anderen Disziplinen ist eine vollständige Planung und Beschreibung des Ablaufs von Data-Science-Projekten, i. d. R. nicht möglich. Als Ergebnis dieses Teilprozesses kann deshalb nur eine Projektskizze entstehen, die im Projektverlauf immer weiter ausgestaltet werden muss. Insbesondere in agilen Vorgehensmodellen wie Scrum oder Kanban kann dies zunächst auf die Erstellung eines Backlogs oder einer vergleichbaren Sammlung an angestrebten Funktionalitäten/Strukturen für angestrebten Lösungen hinauslaufen. Der hier verwendete Begriff der Projektskizze ist entsprechend auf die Vorgehensweise zu übertragen oder anzupassen.

Wichtig ist in jedem Fall, dass bei der Beschreibung des Projekts ein Abstraktionslevel gewählt wird, mit dem sich alle relevanten Anforderungen und Informationen aus Daten-, Domänen- und Analysesicht prägnant darstellen lassen.

Außerdem sollten durch die Projektbeschreibung die zu diesem Zeitpunkt bereits identifizierbaren Arbeitsschritte/-folgen aufgezeigt werden, die zur Erfüllung der festgelegten Anforderungen führen. Sollten sich bei der Projektdurchführung Änderungen ergeben, ist die Projektskizze entsprechend anzupassen.

6 Datenbereitstellung

Innerhalb der Phase der Datenbereitstellung werden die Aktivitäten zusammengefasst, die dem Schlüsselbereich Daten zuzuordnen sind, weshalb der verwendete Begriff weit gefasst ist. Die Phase beinhaltet die Datenaufbereitung (von der Erfassung bis zur Speicherung), das Datenmanagement und eine explorative Analyse. Als Ergebnis dieser Phase entsteht eine für die weitere Analyse geeignete Datenquelle.

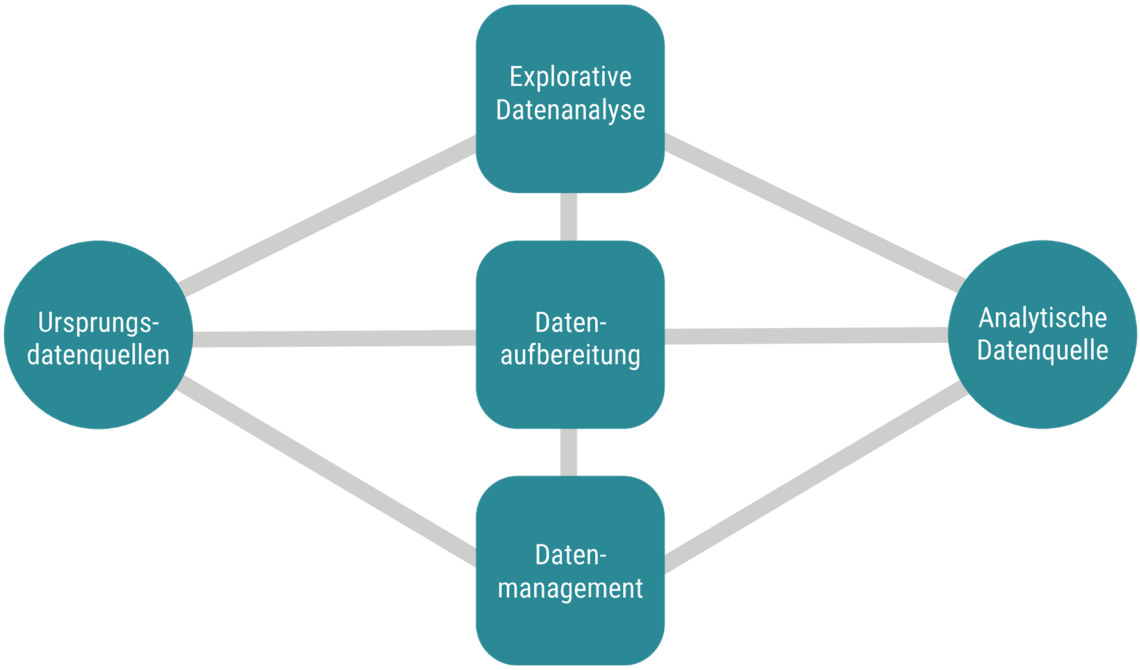


Abbildung 15: Kurzübersicht der Phase „Datenbereitstellung“

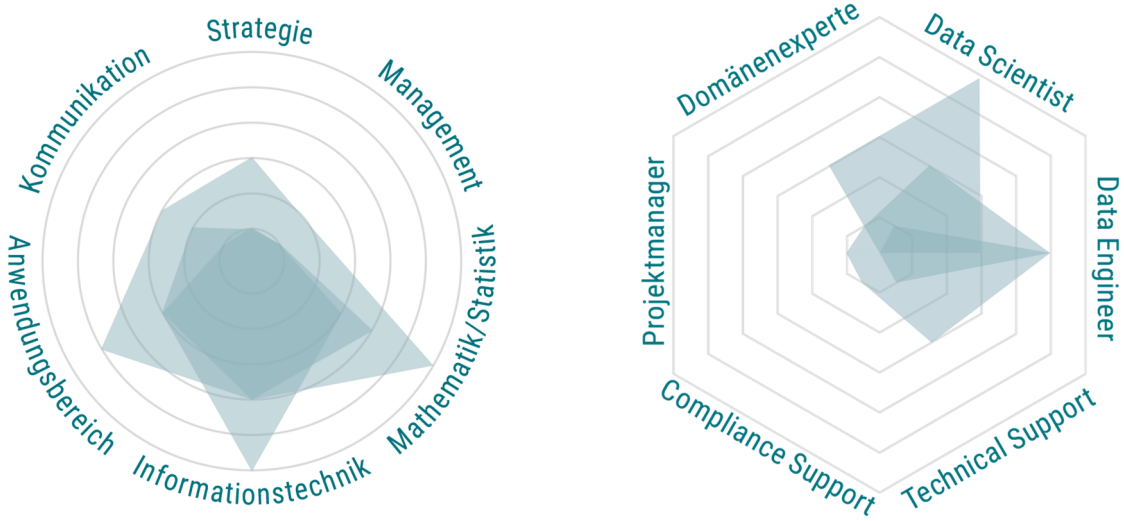
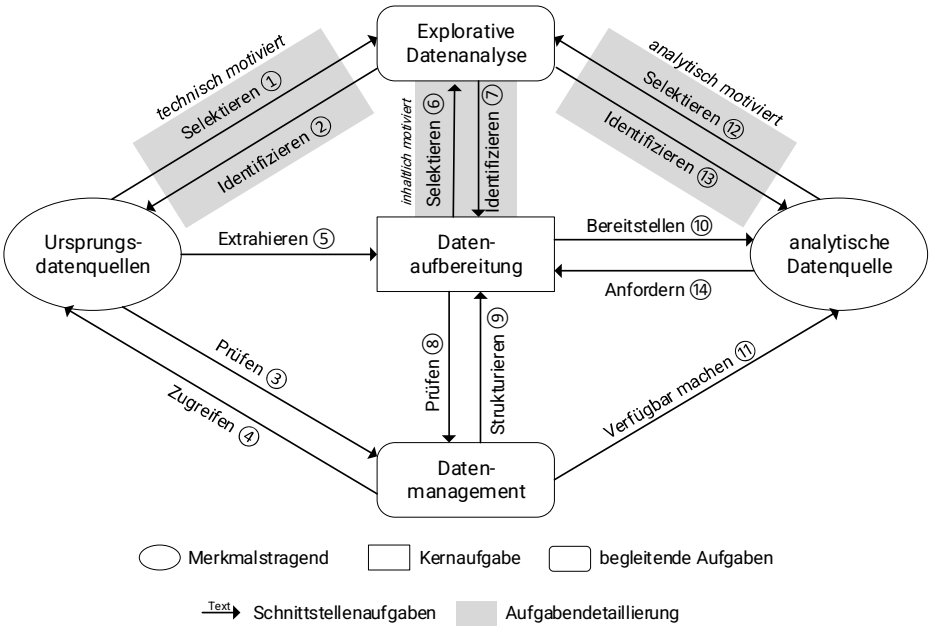


Abbildung 16: Kompetenz- und Rollenprofil der Phase „Datenbereitstellung“

Detaildarstellung der Phase Datenbereitstellung



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	Die explorative Datenanalyse identifiziert sinnvolle Aufbereitungsstrategien.
⑧	Prüfung der Anforderungen an das Datenmanagement im Data-Science-Projekt
⑨	Anforderungen an die Datenstrukturierung bzgl. der Datenspeicherungsarchitektur des aktuellen Projekts 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 für ein explizites Analysevorhaben zu bewerten.
⑬	Für ein explizites Analysevorhaben relevante Eigenschaften werden identifiziert.
⑭	Anpassung der Datenaufbereitung auf Basis der in ⑫ identifizierten Erkenntnisse

Abbildung 17: Detaildarstellung der Phase „Datenbereitstellung“

6.1 Merkmalstragender Bereich „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 7 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 8 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 7: Beschreibung der Merkmalskategorien im Bereich Ursprungsdatenquellen

Merkmalskategorie	**Beschreibung**
Beschaffungsaufwand	Die Verfügbarkeit von Daten kann große Auswirkungen darauf haben, welche Analysen 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 8: Häufig genannte Datenqualitätskriterien, aus Helfert et al. (2001)

Datenqualitätskriterien (vgl. 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	Verwendungsbereit-schaft
Vollständigkeit	Wahrheitsgehalt	Wahrscheinlichkeit	Wartungsfreundlichkeit
Wiederverwendbarkeit	Wirkungsdauer	Zeitadäquanz	Zeitbezug
Zeitoptimal	Zugänglichkeit	Zuverlässigkeit	

6.2 Kernaufgabe „Datenaufbereitung“

Grundsätzlich geht es bei der 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.

Die Verarbeitung großer Datenmengen erfordert den Einsatz leistungsfähiger Hard- und Software sowie teilweise innovativer Verfahren. Dies wird im Schlüsselbereich *IT-Infrastruktur* adressiert. Als mögliche Artefakte der Datenaufbereitung entstehen Skripte, die auch automatisierbar sein können, 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 Tabelle 9 werden von Teilnehmerinnen und Teilnehmern häufig genannte Teilaufgaben aufgeführt und beschrieben, die in Data-Science-Projekten durchgeführt werden müssen.

Tabelle 9: Häufig genannte Teilaufgaben der Aufgabe „Datenaufbereitung“

Teilaufgabe	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.

Teilaufgabe	Beschreibung
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 u. a. 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.

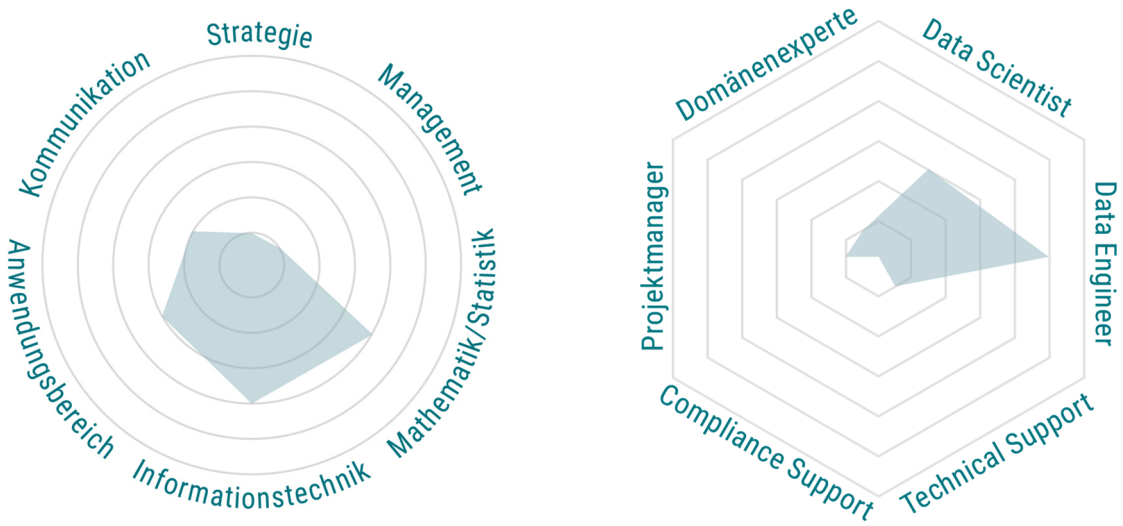


Abbildung 18: Kompetenz- und Rollenprofil der Aufgabe „Datenaufbereitung“

6.3 Begleitende Aufgabe „Datenmanagement“

Beim *Datenmanagement* wird der Fokus auf die Verfügbarmachung der benötigten Daten gelegt, ohne dabei bereits Anforderungen an eine IT-Infrastruktur zu formulieren. In Tabelle 10 werden von Teilnehmerinnen und Teilnehmern häufig genannte Teilaufgaben des *Datenmanagements* aufgeführt und beschrieben. Als Artefakt entsteht im Datenkatalog eine Erweiterung zur Nachvollziehbarkeit des Datenmanagements.

Tabelle 10: Häufig genannte Teilaufgaben der Aufgabe „Datenmanagement“

Teilaufgabe	Beschreibung
Datenarchivierung	Wenn Analyseverfahren 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 besteht die Notwendigkeit, sie vor unbefugtem Zugriff zu schützen oder sie ggf. ausschließlich anonymisiert oder pseudonymisiert unter Berücksichtigung verschiedener Zugriffsrollen bzw. Zugriffsrechte zu speichern.
Datensicherung aufbereiteter Daten	Es ist zu prüfen, ob die aufbereiteten Daten während der Durchführung des Data-Science-Projekts gesichert werden müssen oder ob sie durch erarbeitete 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, sind geeignete Prozesse und Infrastrukturen vorzusehen.
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.

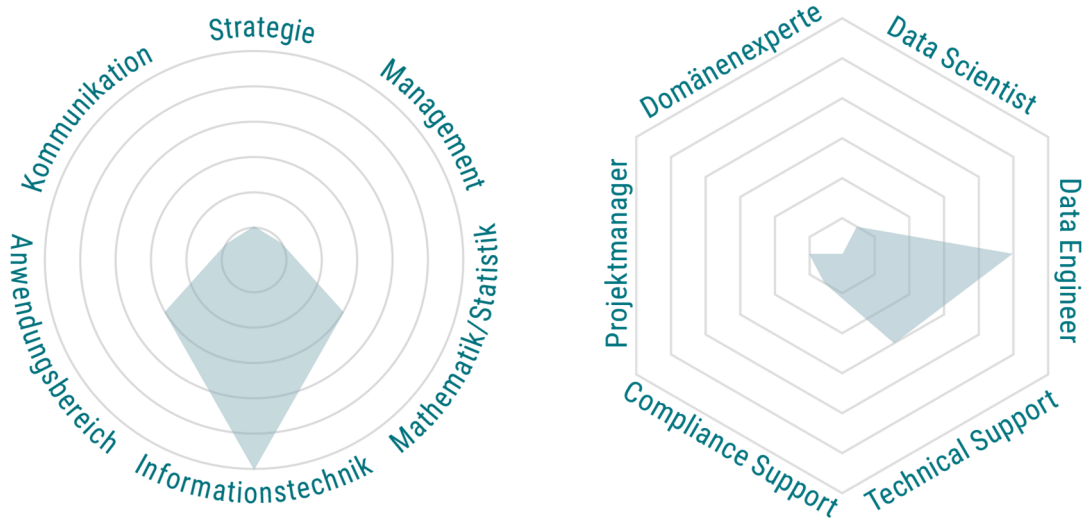


Abbildung 19: Kompetenz- und Rollenprofil der Aufgabe „Datenmanagement“

6.4 Begleitende Aufgabe „Explorative Datenanalyse“

Bei der *Explorativen 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 11 werden von Teilnehmerinnen und Teilnehmern häufig genannte Teilaufgaben der explorativen Datenanalyse aufgeführt und beschrieben.

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.

Tabelle 11: Häufig genannte Teilaufgaben der Aufgabe „Explorative Datenanalyse“

Teilaufgabe	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 ggf. 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 daher, möglichst zentrale, aussagekräftige Attribute zu identifizieren bzw. unerhebliche 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.

Teilaufgabe	Beschreibung
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.

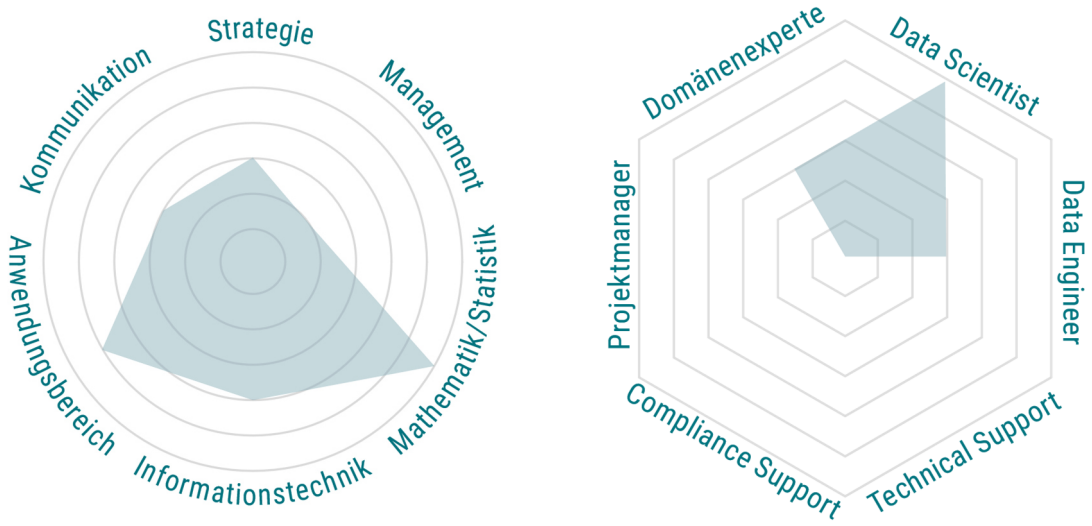


Abbildung 20: Kompetenz- und Rollenprofil der Aufgabe „Explorative Datenanalyse“

6.5 Merkmalstragender Bereich „Analytische Datenquelle“

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 Analyse

In einem Data-Science-Projekt können entweder bestehende Verfahren angewendet oder es müssen zunächst neue Verfahren entwickelt werden – die entsprechende Entscheidung ist eine eigene Herausforderung. Die Phase umfasst daher nicht nur die Analysedurchführung, sondern auch angrenzende Tätigkeiten. Das Artefakt der Phase ist ein Analyseergebnis, das eine methodische und fachliche Evaluation durchlaufen hat.

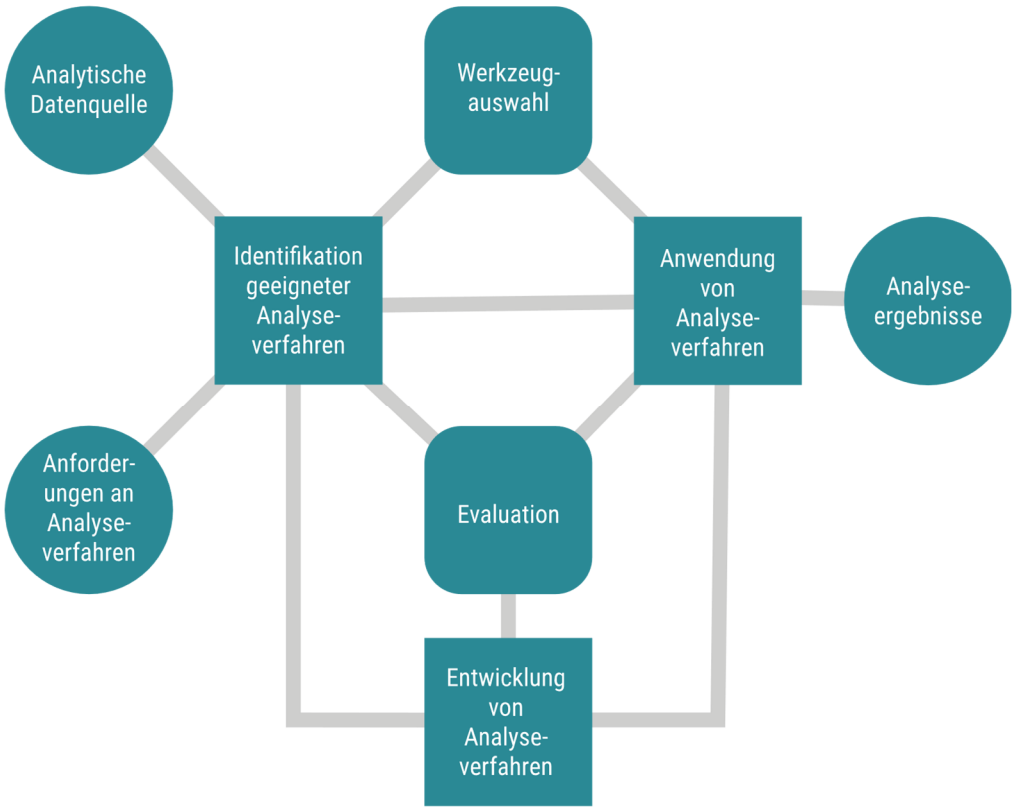


Abbildung 21: Kurzübersicht der Phase „Analyse“

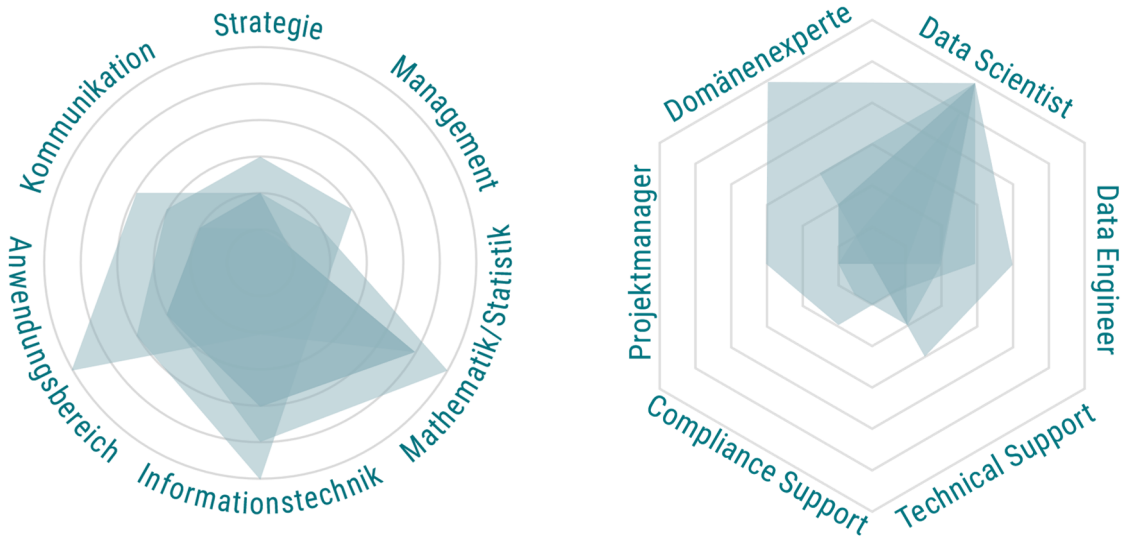
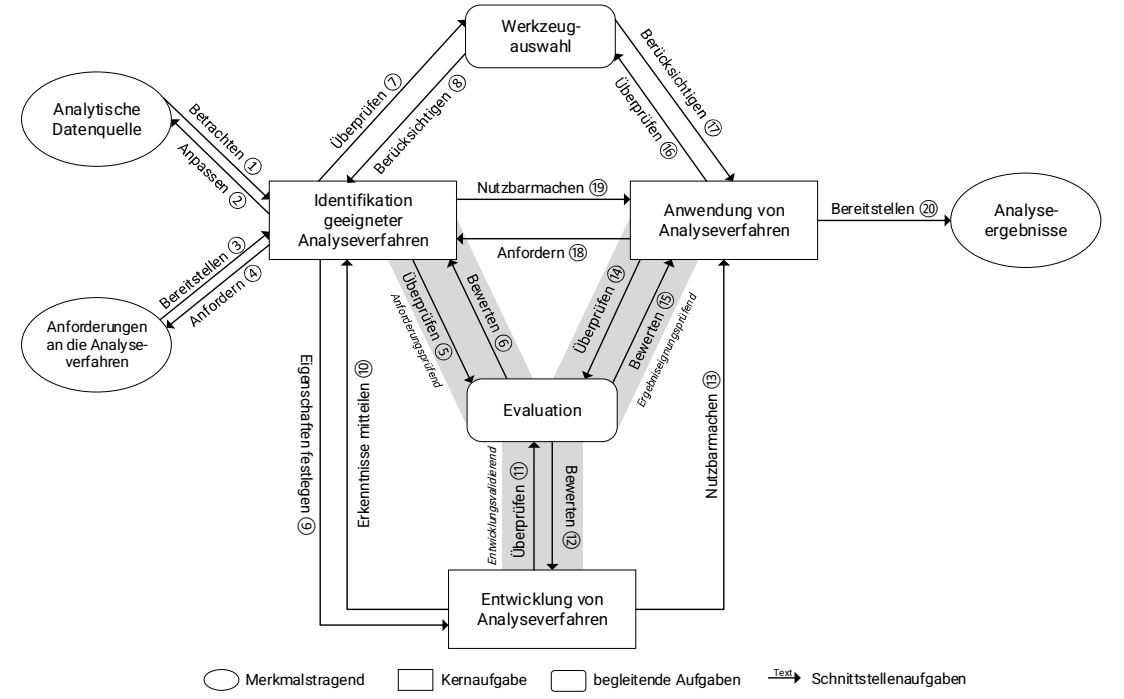


Abbildung 22: Kompetenz- und Rollenprofil der Phase „Analyse“

Detaildarstellung der Phase Analyse



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.

Abbildung 23: Detaildarstellung der Phase „Analyse“

7.1 Merkmalstragender Bereich „Analytische Datenquelle“

Die Identifikation geeigneter Analyseverfahren fußt auf den Merkmalen der vorliegenden analytischen Datenquelle (vgl. Abschnitt 6.5).

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.

Im Rahmen der Phase *Analyse* sind keine darüberhinausgehenden Besonderheiten zu erfassen.

7.2 Merkmalstragender Bereich „Anforderungen an 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 12 werden von Teilnehmerinnen und Teilnehmern häufig genannte Merkmale der Anforderungen an das Analyseverfahren aufgeführt und beschrieben.

Tabelle 12: Häufig genannte Merkmale des Bereichs „Anforderungen an 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 Kernaufgabe „Identifikation geeigneter Analyseverfahren“

Vor Beginn dieser Aufgabe sollte bereits klar sein, dass sich die gegebene Fragestellung tatsächlich mit Hilfe von Data Science beantworten 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, besteht die Möglichkeit, 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 13 werden von Teilnehmerinnen und Teilnehmern häufig genannte Teilaufgaben bei der Identifikation geeigneter Analyseverfahren aufgeführt und beschrieben.

Tabelle 13: Häufig genannte Teilaufgaben der Aufgabe „Identifikation geeigneter Analyseverfahren“

Teilaufgabe	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.

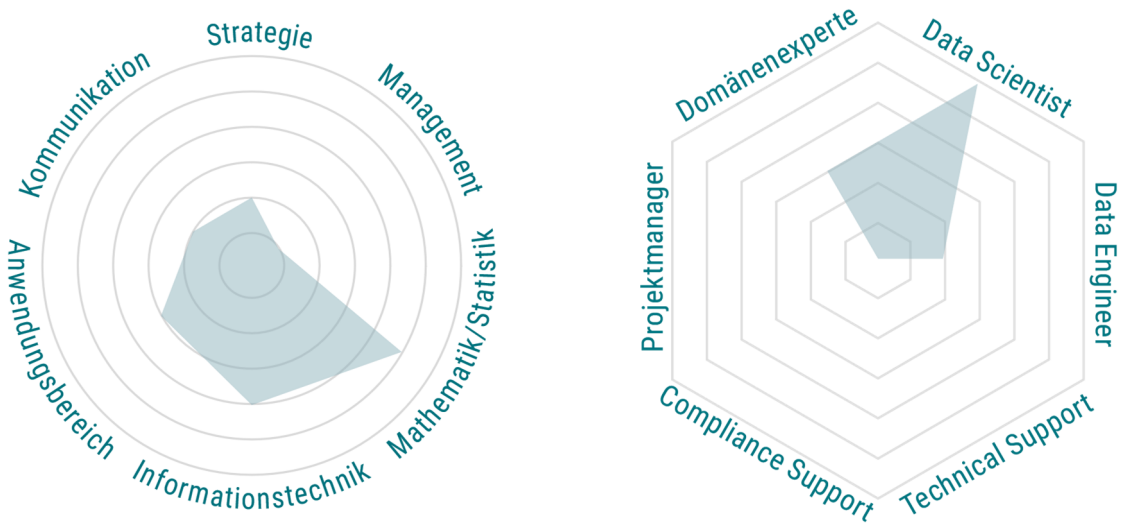


Abbildung 24: Kompetenz- und Rollenprofil der Aufgabe „Identifikation geeigneter Analyseverfahren“

7.4 Kernaufgabe „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 (siehe vorheriger Abschnitt) 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 14 (siehe nächste Seite) werden von Teilnehmerinnen und Teilnehmern häufig genannte Teilaufgaben bei der Anwendung bestehender Analyseverfahren aufgeführt und beschrieben.

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 Schnittstellendokumentation 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.

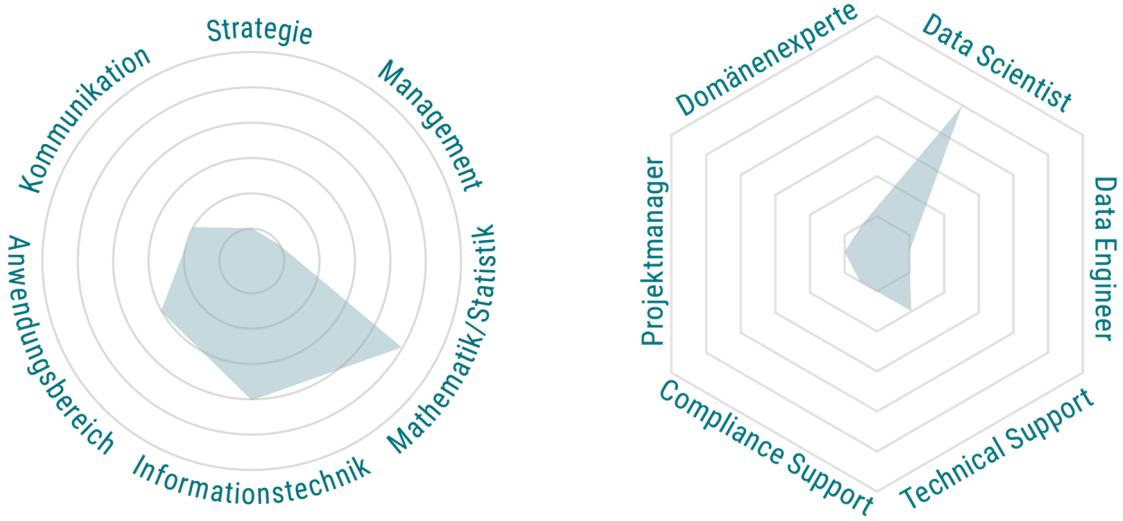


Abbildung 25: Kompetenz- und Rollenprofil der Aufgabe „Anwendung von Analyseverfahren“

Tabelle 14: Häufig genannte Teilaufgaben der Aufgabe „Anwendung von Analyseverfahren“

Teilaufgabe	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.

7.5 Begleitende 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 12), 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 15 (siehe nächste Seite) werden von Teilnehmerinnen und Teilnehmern häufig genannte Teilaufgaben der *Werkzeugauswahl* aufgeführt und beschrieben.

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

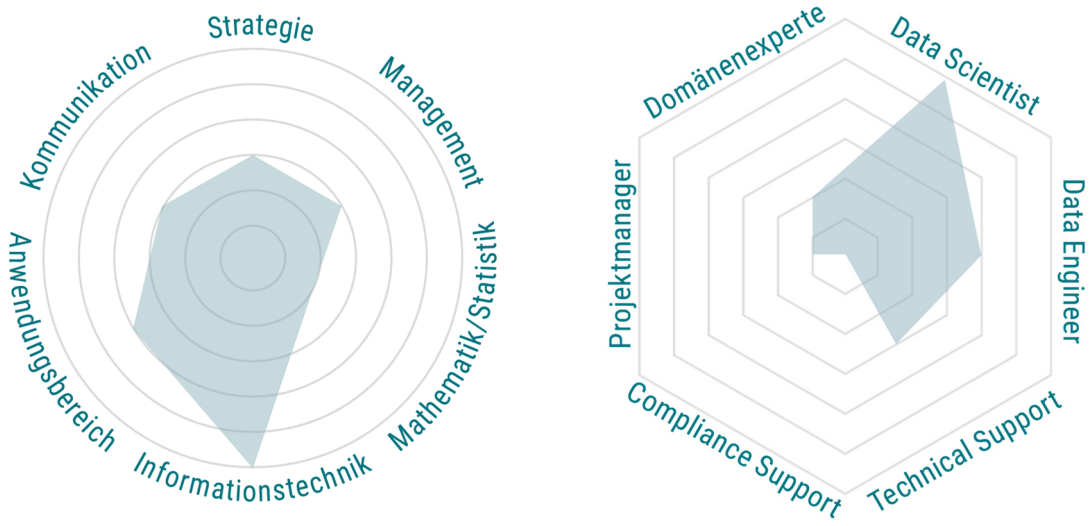


Abbildung 26: Kompetenz- und Rollenprofil der Aufgabe „Werkzeugauswahl“

Tabelle 15: Häufig genannte Teilaufgaben der Aufgabe „Werkzeugauswahl“

Teilaufgabe	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.

7.6 Kernaufgabe „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 16 werden von Teilnehmerinnen und Teilnehmern häufig genannte Teilaufgaben der Entwicklung neuer Analyseverfahren aufgeführt und beschrieben.

Tabelle 16: Häufig genannte Teilaufgaben der Aufgabe „Entwicklung von Analyseverfahren“

Teilaufgabe	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*

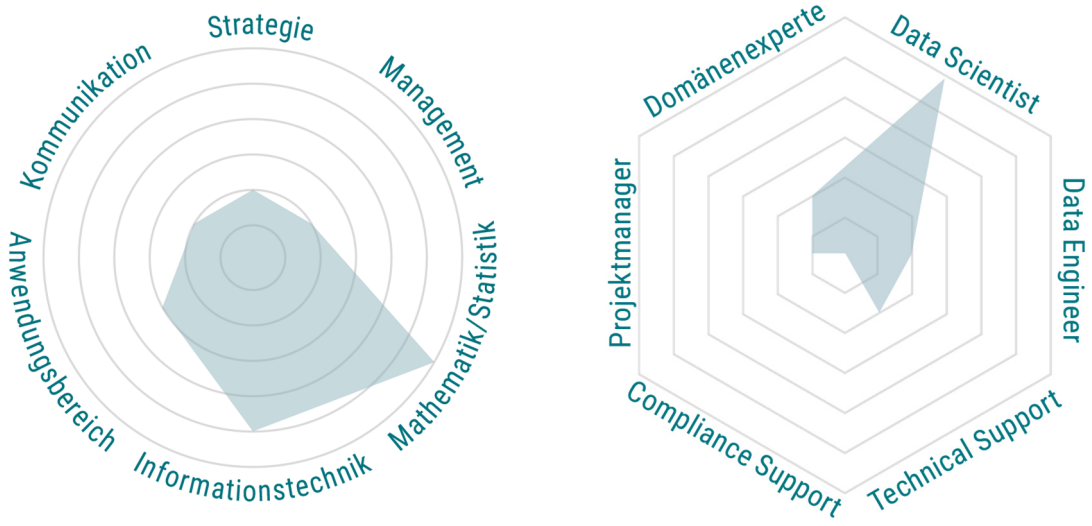


Abbildung 27: Kompetenz- und Rollenprofil der Aufgabe „Entwicklung von Analyseverfahren“

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 17 werden von Teilnehmerinnen und Teilnehmern häufig genannte Teilaufgaben der Evaluation aufgeführt und beschrieben.

Tabelle 17: Häufig genannte Teilaufgaben der Aufgabe „Evaluation“

Teilaufgabe	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 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. Zu beurteilen ist dann die Passung zwischen Verfahren und zu bearbeitender Problemstellung.

Teilaufgabe	Beschreibung
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 28 ist das Kompetenzprofil von Personen, die sich im Bereich *Evaluation* und den mit diesem direkt verbundenen Aufgaben spezialisieren, sowie beteiligte Rollen dargestellt. 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.

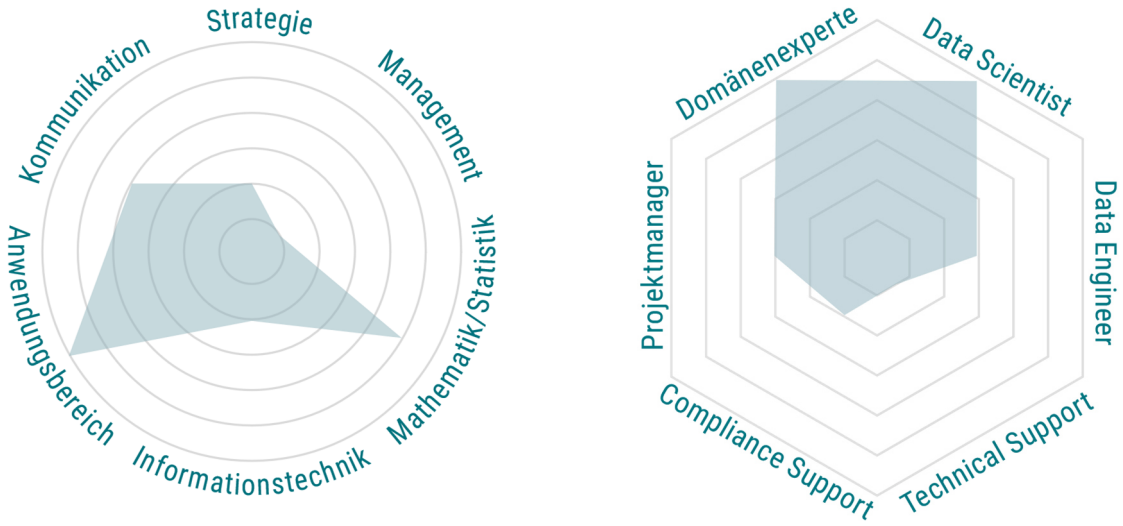


Abbildung 28: Kompetenz- und Rollenprofil der Aufgabe „Evaluation“

7.8 Merkmalstragender Bereich „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 18 werden von Teilnehmerinnen und Teilnehmern häufig genannte Merkmale der Analyseergebnisse aufgeführt und beschrieben.

Tabelle 18: 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? Ist zu erwarten, dass die Ergebnisse auch in Zukunft gültig sein werden und nicht nur im Ist-Zustand?
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 Ergebnisse auf weitere Daten übertragen?
Grenzen	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	Welche quantitativen Bewertungsmaße (Signifikanzniveau, Fehlerrate usw.) liegen vor?
Relevanz	Tragen die Ergebnisse zur Lösung der ursprünglichen Problemstellung bei oder beantworten sie eine andere Frage/haben sie weiteren 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 Nutzbarmachung

In der Phase der Nutzbarmachung wird eine anwendbare Form der Analyseergebnisse geschaffen. Projektspezifisch kann dies eine umfangreiche Betrachtung technischer, methodischer und fachlicher Aufgaben bedeuten oder pragmatisch gehandhabt werden. Die Analyseartefakte können sowohl Resultate als auch Modelle oder Verfahren selbst umfassen und werden den Adressaten in unterschiedlicher Form zur Verfügung gestellt.

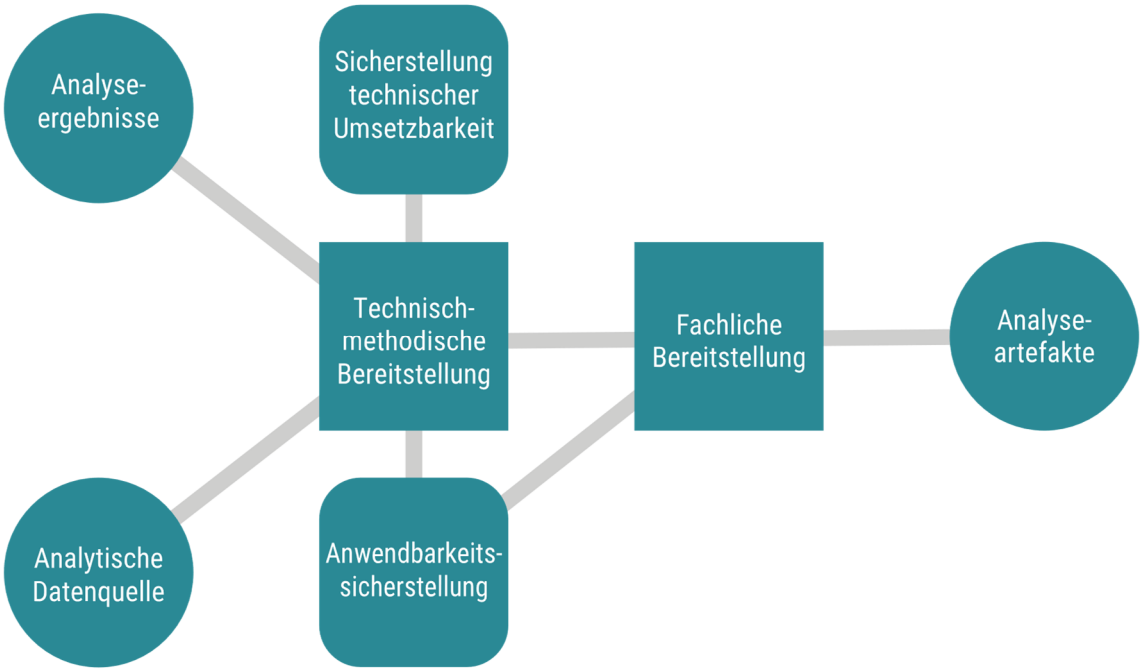


Abbildung 29: Kurzübersicht der Phase „Nutzbarmachung“

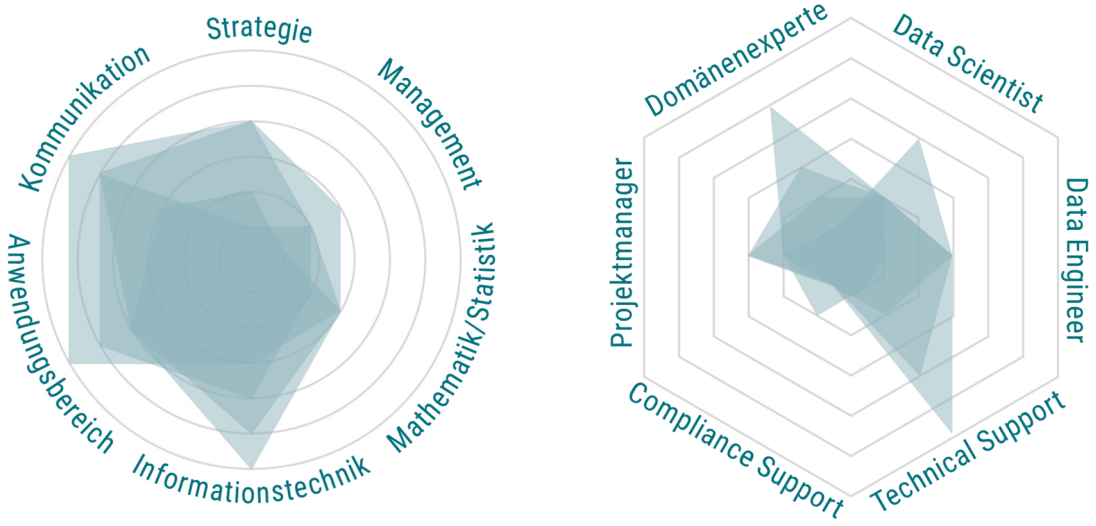
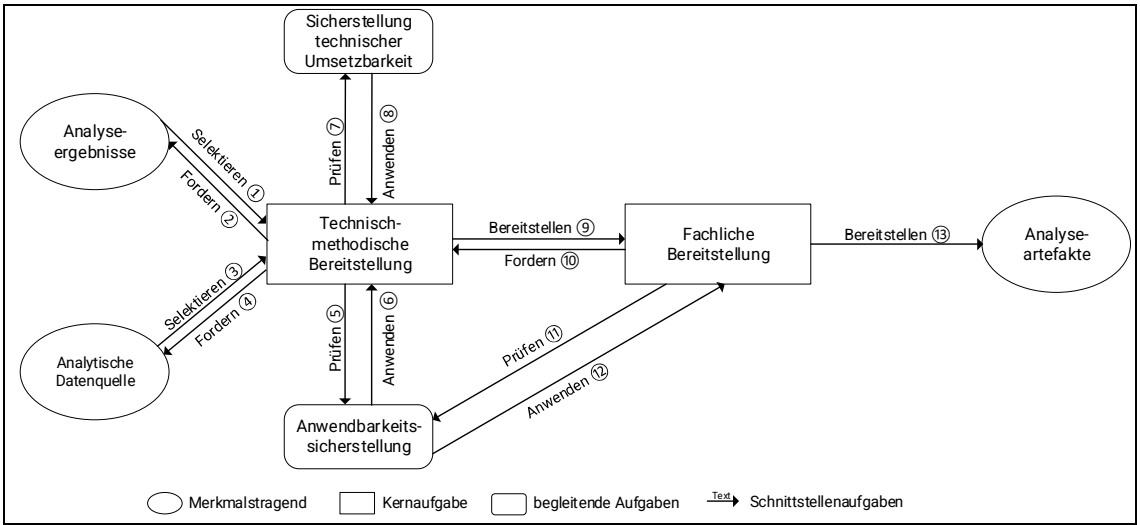


Abbildung 30: Kompetenz- und Rollenprofil der Phase „Nutzbarmachung“

Detaildarstellung der Phase Nutzbarmachung



① Auswahl der Analyseergebnisse für die technisch-methodische Bereitstellung

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.

Abbildung 31: Detaildarstellung der Phase „Nutzbarmachung“

8.1 Merkmalstragender Bereich „Analyseergebnisse“

Die Nutzbarmachung fußt auf den Merkmalen der in Abschnitt 7.8 beschriebenen Analyseergebnisse. Weitere Besonderheiten sind an dieser Stelle nicht zu erfassen.

8.2 Merkmalstragender Bereich „Analytische Datenquelle“

Für die Nutzbarmachung der Analyseergebnisse kann es nötig sein, erneut auf die analytische Datenquelle zuzugreifen (vgl. Abschnitt 6.5). Weitere Besonderheit sind in dieser Phase nicht zu erfassen.

8.3 Kernaufgabe „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*

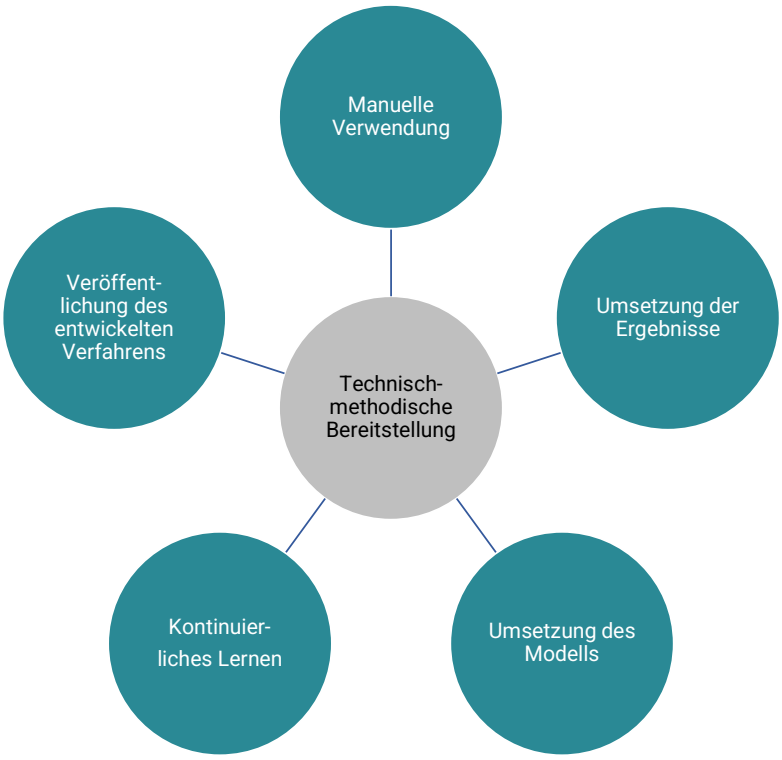


Abbildung 32: Formen der „Technisch-methodischen Bereitstellung“

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 einen Proof of Concept) relevant, ansonsten wird der Wert der Modelle i. d. R. dadurch geschöpft, dass sie kontinuierlich oder auf Anfrage in eine Produktivumgebung eingebettet werden.

In Tabelle 19 werden von Teilnehmerinnen und Teilnehmern häufig genannte Teilaufgaben der *technisch-methodischen Bereitstellung* aufgeführt und beschrieben.

Tabelle 19: Häufig genannte Teilaufgaben der Aufgabe „Technisch-methodische Bereitstellung“

Teilaufgabe	**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 sowie 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 Analysesystems muss überprüft werden, insbesondere, wenn es in die Produktivumgebung der Organisation integriert und an reale Datenquellen angeschlossen wurde.

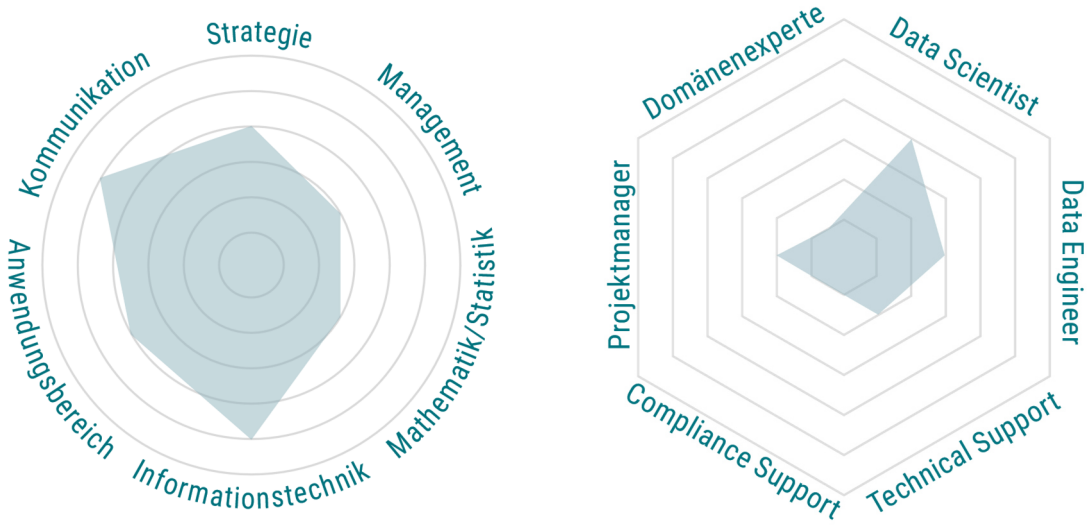


Abbildung 33: Kompetenz- und Rollenprofil der Aufgabe „Technisch-methodische Bereitstellung“

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 20 werden von Teilnehmerinnen und Teilnehmern häufig genannte Teilaufgaben der *Sicherstellung technischer Umsetzbarkeit* aufgeführt und beschrieben.

Tabelle 20: Häufig genannte Teilaufgaben der Aufgabe „Sicherstellung technischer Umsetzbarkeit“

Teilaufgabe	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?

Teilaufgabe	Beschreibung
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?

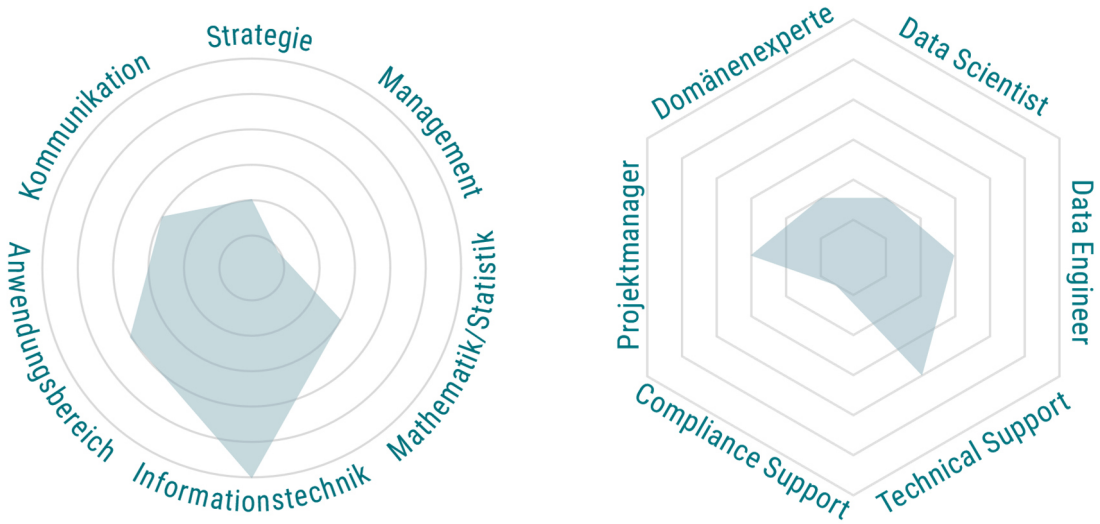


Abbildung 34: Kompetenz- und Rollenprofil der Aufgabe „Sicherstellung technischer Umsetzbarkeit“

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 21 werden von Teilnehmerinnen und Teilnehmern häufig genannte Teilaufgaben der Anwendbarkeitssicherstellung beschrieben.

Tabelle 21: Häufig genannte Teilaufgaben der Aufgabe „Anwendbarkeitssicherstellung“

Teilaufgabe	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.

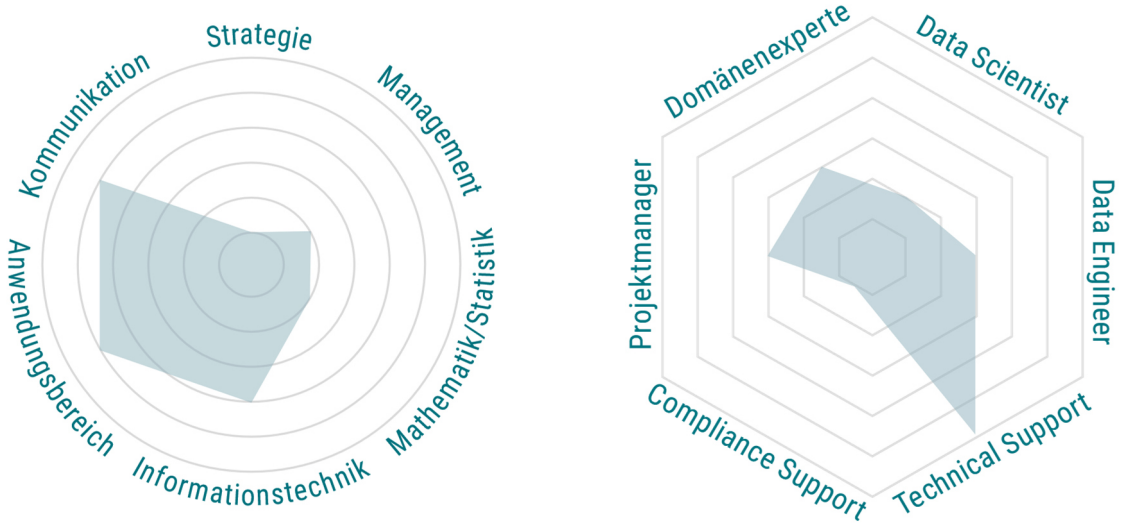


Abbildung 35: Kompetenz- und Rollenprofil der Aufgabe „Anwendbarkeitssicherstellung“

8.6 Kernaufgabe „Fachliche Bereitstellung“

Die Teilaufgaben der fachlichen Bereitstellung hängen stark von der Bereitstellungsform und der Domäne ab, in der das Projekt durchgeführt wird. Dargestellt werden daher ausschließlich allgemeingültige Aufgaben. In Tabelle 22 werden von Teilnehmerinnen und Teilnehmern häufig genannte Teilaufgaben der fachlichen Bereitstellung beschrieben.

Tabelle 22: Häufig genannte Teilaufgaben der Aufgabe „Fachliche Bereitstellung“

Teilaufgabe	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, sind ihre möglichen Auswirkungen unter anderem unter moralischen und wirtschaftlichen Gesichtspunkten einzuschätzen..
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.

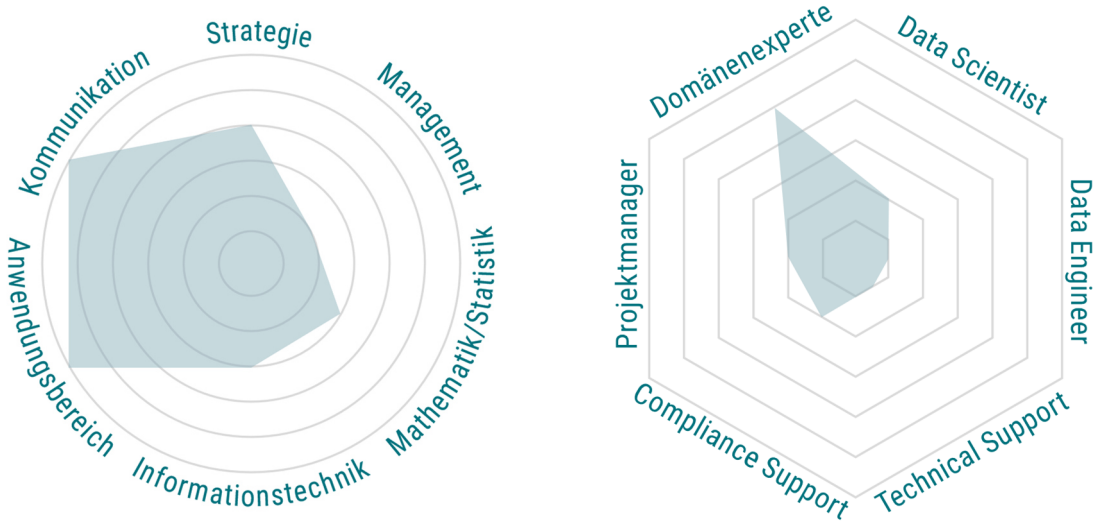


Abbildung 36: Kompetenz- und Rollenprofil der Aufgabe „Fachliche Bereitstellung“

8.7 Merkmalstragender Bereich „Analyseartefakte“

Die Merkmale der Analyseartefakte sind abhängig von der Form der Ergebnisbereitstellung (vgl. Kapitel 8.1). In Tabelle 23 werden von Teilnehmerinnen und Teilnehmern häufig genannte Merkmale von Analyseartefakten beschrieben.

Tabelle 23: 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 Nutzung

Die sich an die Projektdurchführung anschließende Verwendung von Analyseartefakten ist nicht als primärer Teil eines Data-Science-Projekts anzusehen. Ein Monitoring ist aber abhängig von der 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 (auch im Sinne iterativer Vorgehensweisen) zu erlangen.

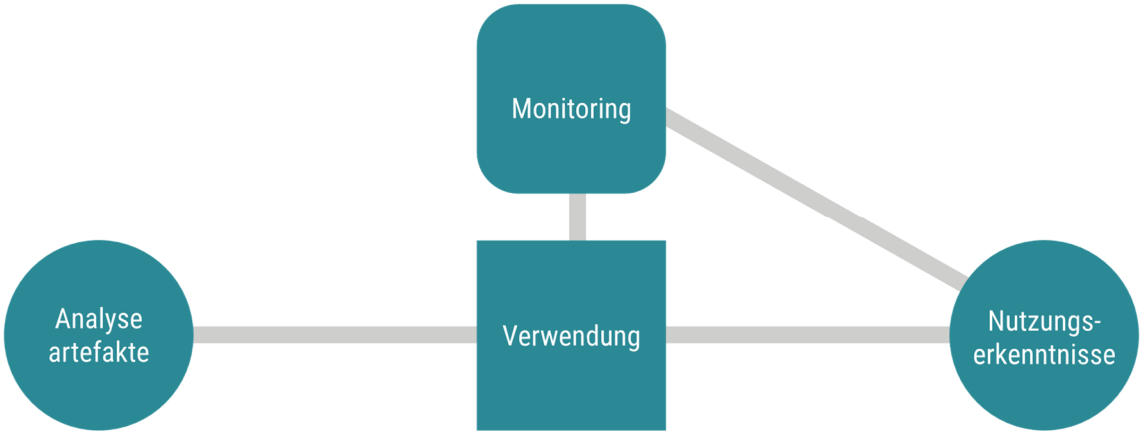


Abbildung 37: Kurzübersicht der Phase „Nutzung“

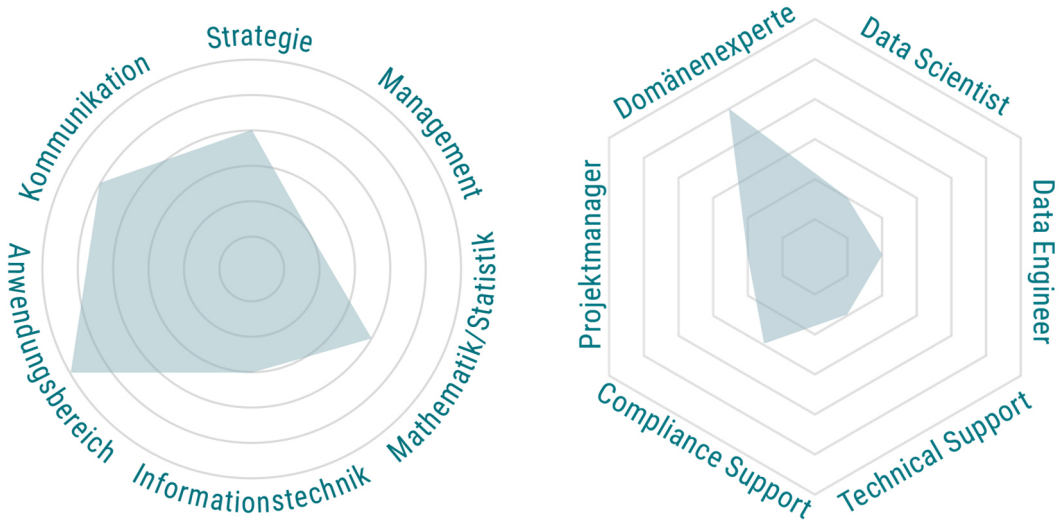
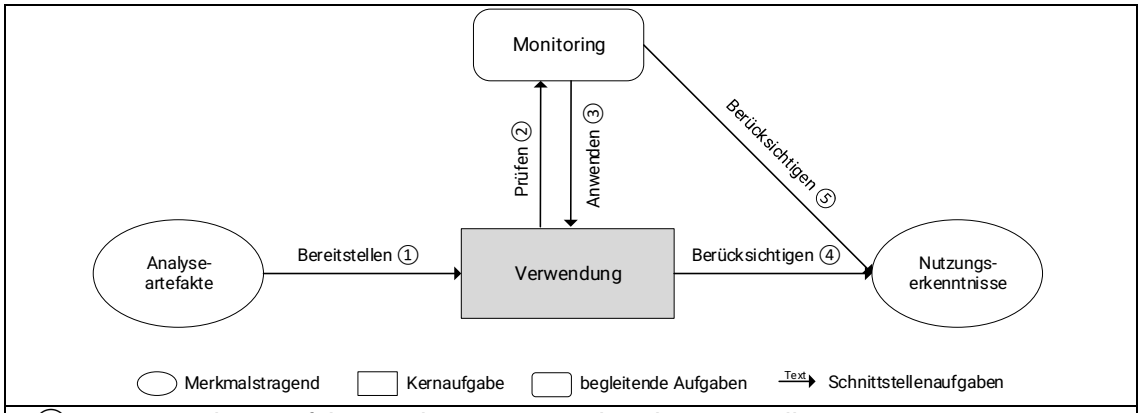


Abbildung 38: Kompetenz- und Rollenprofil der Phase „Nutzung“

Detaildarstellung der Phase Nutzung



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 39: Detaildarstellung der Phase „Nutzung“

9.1 Merkmalstragender Bereich „Analyseartefakte“

Die Nutzung der Analyseartefakte fußt auf deren Merkmalen (vgl. 8.7). Weitere Besonderheiten sind in dieser Phase nicht zu erfassen.

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 24 werden von Teilnehmerinnen und Teilnehmern häufig genannte Teilaufgaben beim Monitoring der Verwendung aufgeführt und beschrieben.

Tabelle 24: Häufig genannte Teilaufgaben der Aufgabe „Monitoring“

Teilaufgabe	**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, liegen keine spezifischen Grafiken für Kompetenzen und Rollen vor, sondern nur die allgemeinen Grafiken der Phasenübersicht.

9.3 Merkmalstragender Bereich „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 25 werden von Teilnehmerinnen und Teilnehmern häufig genannte Merkmale beschrieben, nach denen Nutzungserkenntnisse untergliedert werden können.

Tabelle 25: 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.

Teil C
Übergreifende Schlüsselbereiche

10 Domäne

Neben den in Kapitel 5 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 für 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.*

11 Wissenschaftlichkeit

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 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 die Wissenschaftlichkeit 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 sind nachvollziehbar festzulegen. Eine Übersicht zu vorhandenen 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 geeignete 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“

- *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.*

12 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.*

Teil D
Schlussbemerkungen und Anhang

Schlussbemerkungen

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 Vorgehensmodell 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 Vorgehensmodell 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 allen Anwendern des Modells, eine kritische Betrachtung der eigenen Vorgehensweise durchzuführen oder sich auf Basis der dargestellten Möglichkeiten die für ihr Unternehmen relevantesten Merkmale herauszusuchen. Dabei kann eine bewusste Selektion durchgeführt werden, die ein mühsames Zusammensuchen aus diversen Quellen erspart und 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. Der (begriffliche) Siegeszug der Data Science begann 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 relevanten Skills nahezu unmöglich, einen Tausendsassa oder die „eierlegende Wollmilchsau“ 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.

Das vorliegende Vorgehensmodell 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-Science-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.

Als praktisch wertvolle Ergänzung liefert DASC-PM mit Version 1.1 auch einen umfangreichen Anhang, der insbesondere die Phase des Projektauftrags und dessen ganzheitliche Betrachtung unterstützt.

Literatur

Alekozai E. M., Kaufmann J., Kühnel S., Neuhaus U., Schulz M. (2021). Data-Science-Projekte mit dem Vorgehensmodell „DASC-PM“ durchführen: Kompetenzen, Rollen und Abläufe. In: Barton T., Müller C. (eds) Data Science anwenden. Angewandte Wirtschaftsinformatik. Springer Vieweg, Wiesbaden.

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.

Dorard, L. (2015). Machine Learning Canvas, https://www.ownml.co/machine-learning-canvas, aufgerufen am 20.12.2021.

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.

Kerzel, U. (2021). Enterprise AI Canvas Integrating Artificial Intelligence into Business. In: *Applied Artificial Intelligence*, 35 (1), 1-12.

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

Microsoft (2021): What is the Team Data Science Process?, https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview, aufgerufen am 06.02.2022.

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, aufgerufen am 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.

Schmarzo, B. (2015). Big Data MBA: Driving Business Strategies with Data Science. Wiley.

Schulz, M., Neuhaus, U., Kaufmann, J., Badura, D., Kuehnel, S., Badwitz, W., Dann, D., Kloker, S., Alekozai, E. M., & Lanquillon, C. (2020). Introducing DASC-PM: A Data Science Process Model. In: *Australasian Conference on Information Systems (ACIS) Proceedings*.

Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., & Müller, K. R. (2021). Towards CRISP-ML (Q): A Machine Learning Process Model with Quality Assurance Methodology. Machine Learning and Knowledge Extraction, 3(2), 392-413.

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.

Verzeichnis der Autor:innen

Als Autorinnen und Autoren der Version 1.1 werden alle aktiv an der Bearbeitung dieser Version Beteiligten geführt, die dieser Nennung zugestimmt haben.

Sie bedanken sich bei allen Mitwirkenden der Version 1.0 für die Arbeit an den bisherigen Ausarbeitungen.

Prof. Dr. Michael Schulz, NORDAKADEMIE Hochschule der Wirtschaft

Dipl.-Inform. Uwe Neuhaus, NORDAKADEMIE Hochschule der Wirtschaft

Prof. Dr. Jens Kaufmann, Hochschule Niederrhein

Dr. Stephan Kühnel, Martin-Luther-Universität Halle-Wittenberg

Dr. Emal M. Alekozai, Robert Bosch GmbH

Heiko Rohde (M.Sc.), valantic

Sayed Hoseini (M.Sc.), Hochschule Niederrhein

René Theuerkauf (M.Sc.), Martin-Luther-Universität Halle-Wittenberg

Daniel Badura, valantic

Prof. Dr. Ulrich Kerzel, IU Internationale Hochschule

Prof. Dr. Carsten Lanquillon, Hochschule Heilbronn

Prof. Dr. Stephan Daurer, DHBW Ravensburg

Prof. Dr. Maik Günther, IU Internationale Hochschule

Dr. Lukas Huber, FH Kufstein Tirol

Lukas-Walter Thiée, Universität Lüneburg

Philipp zur Heiden (M.Sc.), Universität Paderborn

Dr. Jens Passlick, VHV Gruppe

Jonas Dieckmann (B.Sc.), Philips

Dr. Florian Schwade, Universität Koblenz

Dr. Tobias Seyffarth, Martin-Luther-Universität Halle-Wittenberg

Wolfgang Badewitz, FZI Forschungszentrum Informatik

Dr. Raphael Rissler, SAP SE

Prof. Dr. Stefan Sackmann, Martin-Luther-Universität Halle-Wittenberg

Prof. Dr.-Ing. Philipp Gölzer, TH Nürnberg

Felix Welter, Universität Hamburg

Jochen Röth (M.Sc.), Shopfloor Management Systems GmbH

Julian Seidelmann (M.Sc.), Hapag-Lloyd

Prof. Dr. Uwe Haneke, Hochschule Karlsruhe

Anhang

Die folgenden Seiten stellen einen Fragebogen dar, der im Rahmen der Erarbeitung von DASC-PM v1.1 erstellt wurde. Er unterstützt bei der Anwendung des Modells in der Phase des Projektauftrags dabei, die Eckpunkte des Projekts, seine Ziele und mögliche Fallstricke zu erkennen und zu verdeutlichen. Für jede Frage sind Antwortmöglichkeiten vorgegeben und/oder der Hinweis auf eine Freitextbearbeitung aufgenommen.

Den Abschluss des Dokuments bildet das Poster zu DASC-PM v1.1. Die entsprechende Seite kann auch in hoher Qualität für großformatige Ausdrucke zur Verfügung gestellt werden. Auch den Fragebogen stellen wir gerne für den komfortableren Einsatz als Datei für ihr Tabellenkalkulationsprogramm zur Verfügung, gleiches gilt für einen Präsentationsfoliensatz oder einzelne Grafiken. Bitte wenden Sie sich in allen Fällen an info@dasc-pm.org.

Schlüsselbereich	Unterthema	Frage		MC-Items	Freitext (ggf. hier dargestellt: Beispielantworten)
Hinweise zu Nutzung des Fragebogens **(72 Fragen, je Schlüsselbereich/Unterthema nach durchschnittlich erwartetem Einfluss auf Projekterfolg gerankt)**				Die in dieser Spalte dargestellten Werte sind zeilenweise als Multiple-Choice-Antworten anzusehen. Ein "_____" signalisiert, dass bei Auswahl zusätzlich eine Spezifikation anzugeben ist. Ausgegraute Felder sind nicht zu berücksichtigen.	Ein weißes Feld in dieser Spalte signalisiert die Möglichkeit einer Freitextantwort, ggf. in Ergänzung zu Multiple-Choice-Antworten. Sofern die Felder Text beinhalten, ist dieser ausschließlich als Beispielantwort anzusehen, um die Intention der Frage zu verdeutlichen und kann im Anwendungsfall gelöscht werden. Ausgegraute Felder sind nicht zu berücksichtigen.
Domäne	**Problemstellung und Ziele**	*Welche Problemstellung soll mit dem Projekt gelöst werden?*			Wir möchten herausfinden, welche unserer Kunden im nächsten Jahr voraussichtlich ihren Vertrag kündigen werden. Wir möchten herausfinden, was die Conversions-Wahrscheinlichkeit für einen Besuch im Webshop ist und was die Haupteinflussfaktoren sind.
Domäne	**Problemstellung und Ziele**	*Welche Ziele werden mit dem Projekt verfolgt?*		Neue Erkenntnisse über das Fachthema gewinnen Kompetenzerhöhung Neuartige Problemstellung lösen Neue Geschäftsfelder/Zielgruppen erschließen Kostenersparnis Zeitersparnis Reduktion des Arbeitsaufwands	
Domäne	**Problemstellung und Ziele**	*Welche Ergebnisse werden erwartet?*		Manuelle Verwendung der Ergebnisse (z. B. Seminar, Workshop) Umsetzung der Ergebnisse (z. B. Aufbereitung in Form eines einmaligen Berichts) Umsetzung des Modells (Anwendung des trainierten Modells auf neue, unbekannte Daten) Kontinuierliches Lernen (selbsttätige Anpassung des Modells durch wiederholte Anwendung auf unbekannte Daten) Veröffentlichung des entwickelten Verfahrens (ggf. nur organisationsintern)	
Domäne	**Problemstellung und Ziele**	*Wie wird der Erfolg gemessen?*		Relevante KPI: _____ Relevante technische Metrik: _____ Vergleich zu einem Baseline-Modell Vergleich zum Vorzustand	Bei technischen Metriken Beispiele nennen
Domäne	**Problemstellung und Ziele**	*Worin besteht die Motivation, die Problemstellung durch ein Data-Science-Projekt zu adressieren?*		Klassische Data-Science-Problemstellung, z.B. Segmentierung, Klassifizierung, Regression Komplexität des Themas / nicht-offensichtliche Zusammenhänge Gute Vorerfahrung mit Data-Science-Ansätzen Umfangreiche / geeignete Datenbasis Misserfolg vorheriger Methoden Neugierde, ob Data Science neue Erkenntnisse liefert	
Domäne	**Problemstellung und Ziele**	*Welche projektnahen Ziele sollen definitiv nicht verfolgt werden?*			Ablösung des Standardberichtwesens Vollständige Automatisierung der Entscheidungsprozesse

Schlüsselbereich	Unterthema	Frage		MC-Items	Freitext (ggf. hier dargestellt: Beispielantworten)
Domäne	**Beteiligte und Stakeholder**	*Welche Organisationseinheiten sind konkret involviert?*		Management Fachabteilung: _____ IT Data-Science-Team Externe: _____	
Domäne	**Beteiligte und Stakeholder**	*Welche Organisationseinheiten sind konkret verantwortlich?*		Management Fachabteilung: _____ IT Data-Science-Team Externe: _____	
Domäne	**Beteiligte und Stakeholder**	*Wer hat das Projekt beauftragt?*		Management Fachabteilung: _____ IT Data-Science-Team Externe: _____	
Domäne	**Beteiligte und Stakeholder**	*Welche Anspruchsgruppe dienen zusätzlich zu den Projektbeteiligten als Input-Geber zu fachlichen Aspekten?*			Kunden Rechtsabteilung
Domäne	**Beteiligte und Stakeholder**	*Wer unterstützt / fördert das Projekt?*		Management Fachabteilung: _____ IT Data-Science-Team Externe: _____ (z. B. Wissenschaft / Politik)	
Domäne	**Beteiligte und Stakeholder**	*Gibt es mögliche "Störer" für das Projekt?*		Abteilungen / Organisationseinheiten: _____ Einzelpersonen: _____ Externe: _____	
Domäne	**Beteiligte und Stakeholder**	*Welches sind die Aufgabenfelder eines externen Dienstleisters?*		Projektmanagement IT-Infrastruktur Datenaufbereitung Datenanalyse / Modellerstellung Betrieb / Weiterentwicklung	

Schlüsselbereich	Unterthema	Frage		MC-Items	Freitext (ggf. hier dargestellt: Beispielantworten)
Domäne	**Projektorganisation**	*Welche Projektmanagementmethode ist vorgesehen?*		Agiles Projektvorgehen Wasserfall-Modell Continuous Integration DevOps-Ansatz Keine Methode / Mischung verschiedener Methoden	
Domäne	**Projektorganisation**	*Welche Rollen sind am Projekt beteiligt?*		Data Scientist Data Engineer Domänenexpterte: _____ Projektmanager Technischer Support Compliance Support	
Domäne	**Projektorganisation**	*Wie sieht die Organisationsform des Projektes aus?*			Virtuelles Team Hierarchische Teamstruktur
Domäne	**Ressourcen**	*Welche zeitlichen Rahmenbedingungen existieren während der Projektdurchführung bis zur Vorlage des Ergebnisses?*		kurzfristige Deadline: _____Wochen langfristige Deadline Entwicklung in agilen Sprints _____ Wochen	
Domäne	**Ressourcen**	*Welche Kompetenzen haben die Projektmitglieder?*		Mathematik / Statistik Informationstechnik (z.B.Programmierkenntnisse, Datenbanken) Anwendungsbereich: _____ Kommunikation Strategie Management	
Domäne	**Ressourcen**	*Welche finanziellen Rahmenbedingungen gibt es?*		Personen: _____ IT-Infrastruktur: _____ Extern: _____	
Domäne	**Ressourcen**	*Wieviel Vorlaufzeit existiert bis das Projekt beginnen muss?*		Keine Eher wenig: _____Wochen Eher ausreichend / viel: _____Wochen Flexibler Starttermin	
Domäne	**Vorerfahrungen**	*Welche Lösungsansätze bestehen schon?*			Deskriptive Ansätze Modelle Automatisierung Berichtswesen
Domäne	**Vorerfahrungen**	*Welche Erfahrungen wurden durch vorherige ähnliche Projekte gesammelt?*		Keine Teilweise übertragbare Erfahrungen Exakt übertragbare Erfahrungen Positive: _____ Negative: _____	
Domäne	**Vorerfahrungen**	*Wo lagen bei vergangenen Projekten die Schwierigkeiten?*		Komplexität des Themas Datenbasis Projektorganisation Projektumfeld Personalstruktur Qualität der Modelle Interpretation der Ergebnisse	
Domäne	**Vorerfahrungen**	*Welche Organisationseinheiten haben Vorerfahrung mit Data Science?*		Management Fachabteilung: _____ IT	

Schlüsselbereich	Unterthema	Frage		MC-Items	Freitext (ggf. hier dargestellt: Beispielantworten)
Daten		Welche Daten sollen Verwendung finden (Fachlichkeit / Typ)?			Stammdaten Sensordaten Transaktionsdaten
Daten		Welche Daten stehen grundsätzlich zur Verfügung (Fachlichkeit / Typ)?			Stammdaten Sensordaten Transaktionsdaten
Daten		Wie ist die Datenqualität (Vollständigkeit, Fehlerfreiheit usw.)?		Hoch: _____ Mittel: _____ Gering: _____	
Daten		Welche Datenquellen sind für das Projekt relevant?		Operative Datenquellen (z. B. ERP-System, CRM-System) Analytische Datenquellen (z. B. Data Warehouse, Data Lake) Streaming-Datenquellen (z. B. Sensordaten) Externe Datenquellen: _____	
Daten		Wer ist der Data Owner?		Fachabteilung: _____ IT Data-Science-Team	
Daten		Wer übernimmt die Datenbereitstellung und -aufbereitung?		Fachabteilung: _____ IT Data-Science-Team Externe: _____	
Daten		Sind die benötigten Datenquellen zugreifbar?		Ja Nein, Folgendes ist zu tun: _____	
Daten		Wie groß ist die Bedeutung von Datenschutz und Datensicherheit?		Hoch: _____ Mittel: _____ Gering: _____	
Daten		Müssen Daten (vollständig) neu erhoben werden?		Ja, vollständig Ja, überwiegend Ja, teilweise Ja, wenig Nein Noch zu prüfen	
Daten		Wie hoch ist der Ressourcen-Anteil der Datenbereitstellung und -aufbereitung?		Hoch: _____ Mittel: _____ Gering: _____	
Daten		Was ist die Struktur der Daten?		Strukturierte Daten: _____ Seminstrukturierte Daten: _____ Unstrukturierte Daten: _____	
Daten		Wie erfolgt der Zugriff auf die Daten?			Datenbank-API CSV-Dateien Crawler

Schlüsselbereich	Unterthema	Frage		MC-Items	Freitext (ggf. hier dargestellt: Beispielantworten)
Analyseverfahren		Besteht Klarheit darüber, ob die Problemstellung mit Data-Science-Analysen beantwortet werden kann / soll?		Nein, es besteht noch keine Klarheit darüber Nein, Prüfung der Eignung von DS-Methoden ist Ziel des Projekts Ja, weil: _____	
Analyseverfahren		Besteht bereits Klarheit bezüglich des Typs des benötigten Analyseverfahrens (Klassifikation, Regression, Clustering, Ausreißererkennung usw.)?		Nein, am Problemverständnis wird noch gearbeitet Nein, eine Abbildung auf einen Aufgabentypen ist noch nicht erfolgt Ja, es besteht Klarheit und zwar (Nennung Aufgabentyp): _____	
Analyseverfahren		Wird davon ausgegangen, dass ein etabliertes Analyseverfahren eingesetzt werden kann oder muss ein neues entwickelt werden?		Etablierte Verfahren wurden noch nicht evaluiert Etablierte Verfahren werden gerade evaluiert Mit etablierten Verfahren konnte bislang keine zufriedenstellende Lösung gefunden werden	
Analyseverfahren		Welche besonderen Anforderungen werden an das Analyseverfahren gestellt?		Keine Laufzeit Skalierbarkeit Robustheit Datenverfügbarkeit Erklärbarkeit	

Schlüsselbereich	Unterthema	Frage		MC-Items	Freitext (ggf. hier dargestellt: Beispielantworten)
Nutzbarmachung		*Sollen die Analysemodelle auch auf zukünftige Daten angewendet werden?*		Ja, das ist geplant, für (z.B. rollierende Prognosen) : _____ Nein, weil (z.B. nur einmaliger Erkenntnisgewinn): _____	
Nutzbarmachung		*Falls die Analysemodelle dauerhaft genutzt werden sollen: Ist eine Stand-alone-Anwendung geplant oder sollen die Modelle in bestehende operatives Systeme integriert werden?*		Es ist eine Stand-alone-Anwendung geplant Integration in bestehende Systeme wird angestrebt	
Nutzbarmachung		*Ist davon auszugehen, dass sich die Analysemodelle kontinuierlich an neue Daten anpassen müssen?*		Ja, wichtig Ja, prinzipiell Nein, weil… Daten stabil Datenquellen stabil _____ Derzeit unbekannt	
Nutzbarmachung		*Wie soll das Projektergebnis den Stakeholdern vermittelt werden (Bericht, Workshops, Seminare usw.)?*		Zusammenfassung / Projekt-Bericht (Schriftlich, Präsentationsdokument, etc.) Digitaler Bericht (i.S.e. Reportings) Abschlusspräsentation Live-Demonstration (PoC, MVP, Final) Workshop Schulung	
Nutzbarmachung		*Was ist notwendig, damit die späteren Nutzer die Analysemodelle effizient und korrekt nutzen können?*			Handbuch Schulung technische Dokumentation
Nutzbarmachung		*Welche Betriebskonzept wird angestrebt?*			Ein Betrieb wird von einem speziellen MLOps-Team übernommen. Die Sicherstellung der fachlichen Weiterentwicklung soll in erster Linie durch Folgeprojekte erreicht werden.
Nutzbarmachung		*Wer ist später für die Pflege der Daten und Analysemodelle verantwortlich und dient als inhaltlicher Ansprechpartner?*		Data Scientist (auch: Team, etc.) Data Engineer Fachbereich Anwender IT Externe	
Nutzbarmachung		*Wer ist für die inhaltliche Weiterentwicklung der Analysemodelle verantwortlich?*		Data Scientist (auch: Team, etc.) Data Engineer Fachbereich Anwender IT Externe	
Nutzbarmachung		*Wer übernimmt die Integration der Analysemodelle in die operative IT-Infrastruktur der Organisation?*		Data Scientist (auch: Team, etc.) Data Engineer Fachbereich Anwender IT Externe	
Nutzbarmachung		*Wer übernimmt später den technischen Betrieb der Analyseanwendung?*		Data Scientist (auch: Team, etc.) Data Engineer Fachbereich Anwender IT Externe	

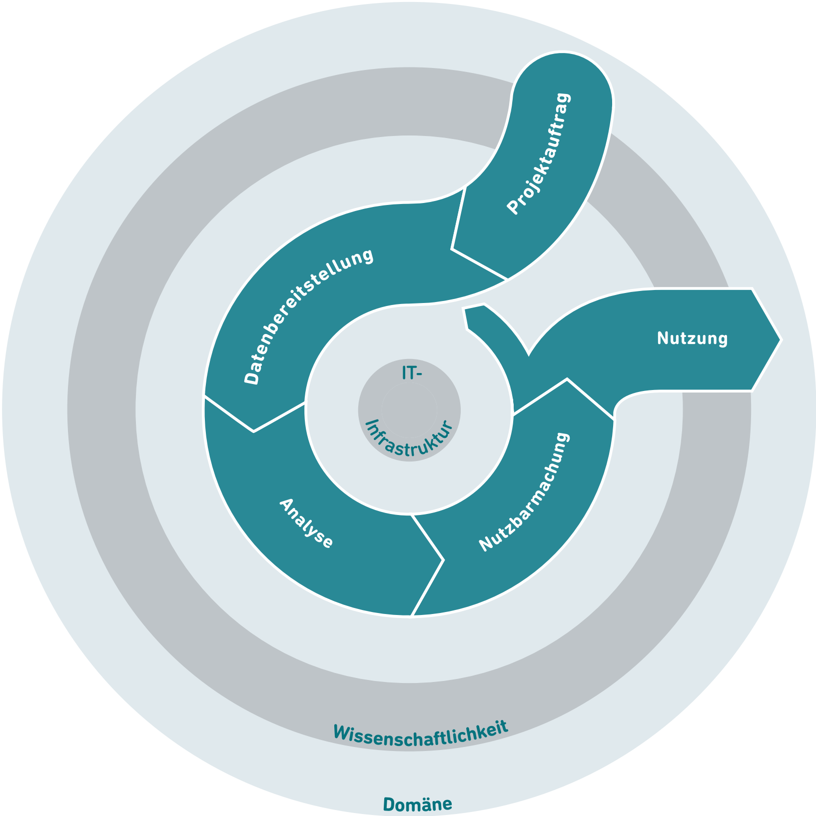
Schlüsselbereich	Unterthema	Frage		MC-Items	Freitext (ggf. hier dargestellt: Beispielantworten)
Nutzung		*Wie sollen die Ergebnisse genutzt werden?*		Für konkrete Verbesserungen (z.B. Kosteneinsparungen) Für Forschung und Entwicklung Als Basis für weiterführende Analysen Steht noch nicht fest	
Nutzung		*Welche Zielgruppe ist hauptsächlich an dem Ergebnissen interessiert?*		Management Fachabteilung: _____ Data-Science-Team Externe: _____	
Nutzung		*Wer sind die Nutzer der Ergebnisse?*		Management Fachabteilung: _____ Data-Science-Team Externe: _____	
Nutzung		*Wie wird die Aufrechterhaltung der Ergebnisqualität gewährleistet?*			Festlegung von Schwellwerten als Messkriterien Regelmäßige inhaltliche Auseinandersetzung mit den Ergebnissen Tracking von Änderungen in Verteilungen der Daten
Nutzung		*Wird das Nutzungspotential durch Compliance-Anforderungen eingeschränkt?*		Ja, folgendermaßen: _____ Nein	

Schlüsselbereich	Unterthema	Frage		MC-Items	Freitext (ggf. hier dargestellt: Beispielantworten)
IT-Infrastruktur		*Welche Art von Software wird für die Durchführung des Projekts voraussichtlich benötigt?*			Datenbanksoftware Analyssoftware Visualisierungssoftware
IT-Infrastruktur		*Welche projektrelevante Software wird im Unternehmen bereits eingesetzt?*			Datenbanksoftware Analyssoftware Visualisierungssoftware
IT-Infrastruktur		*Ist für das Projekt eine enge Verzahnung mit der IT-Infrastruktur notwendig?*		Ja Nein Abhängig von: _____	
IT-Infrastruktur		*Wie hoch sind die erwarteten Anforderungen an die Leistungsfähigkeit der Hard- und Software?*		Hoch: _____ Mittel: _____ Gering: _____	
IT-Infrastruktur		*Welche Vorgaben zur Anschaffung / Verwendung bestimmter Software-Produkte für das Projekt gibt es?*		Open Source erwünscht Software-as-a-Service-Lösung Datenschutz / Compliance: _____ Software mit offiziellem Support / Schulungen	
IT-Infrastruktur		*Welche Hardware steht zur Durchführung des Projekts zur Verfügung?*		Laptops / Desktop-PCs Serverumgebung umfangreiche Ressourcen (z.B. auch in der Cloud)	
IT-Infrastruktur		*Was muss an zusätzlicher Hardware bereitgestellt werden?*			Rechenkapazität Speicherkapazität

Schlüsselbereich	Unterthema	Frage		MC-Items	Freitext (ggf. hier dargestellt: Beispielantworten)
Wissenschaftlichkeit		*Welches Vorgehensmodell wird verwendet?*			CRISP-DM DASC-PM KDD
Wissenschaftlichkeit		*Welche Methode wird zur Ergebnisevaluation verwendet?*			Vergleich mit Ist-Zustand Vergleich mit Baselines Expertenbefragung
Wissenschaftlichkeit		*Ist geplant, die im Projekt verwendeten Daten auch anderen zur Verfügung zu stellen?*		Ja, Folgenden: _____ Ja, teilweise: _____ Nein	
Wissenschaftlichkeit		*Werden voraussichtlich repräsentative Ergebnisse erzeugt, d. h. können die Ergebnisse verallgemeinert und über den eigenen Anwendungskontext hinaus verwendet werden?*		Ja Nein	
Wissenschaftlichkeit		*Welches Forschungsparadigma liegt dem Projekt zugrunde?*			Empirisch quantitativ Empirisch qualitativ Design orientiert
Wissenschaftlichkeit		*Wie wird der State-of-the-Art der wissenschaftlichen Literatur berücksichtigt?*		Umfangreiche Recherche Verwendung der Standardliteratur Gar nicht, weil: _____	
Wissenschaftlichkeit		*Wird das gewählte Vorgehen detailliert dokumentiert?*		Ja, der gesamte Prozess inklusive aller Zwischeschritte kann von Dritten dadurch nachvollzogen werden Ja, aber nur das Ergebnis kann von Dritten dadurch nachvollzogen werden Nein, eine detaillierte Dokumentation ist nicht nötig	
Wissenschaftlichkeit		*Wie werden die Erkenntnisse des Projekts später voraussichtlich mit anderen geteilt?*		Weitergabe an ausgewählte Dritte Freie Veröffentlichung der Ergebnisse	
Wissenschaftlichkeit		*Welche Evaluationskriterien (Gütemaße) werden herangezogen?*			
Wissenschaftlichkeit		*Leistet das Projekt einen eigenen Forschungsbeitrag?*		Ja, Forschungsbeitrag: _____ Nein, es wird ein Standardansatz repliziert	

dasc°pm v1.1

Das Data Science Process Modell (DASC-PM) ist ein Vorgehensmodell für Data-Science-Projekte. Es beschreibt die projektrelevanten Schlüsselbereiche und zu durchlaufenden Phasen, erläutert die typischen Aufgaben innerhalb der Phasen und stellt die beteiligten Projektrollen und benötigten Kompetenzen dar.



Domäne

Ein breites Hintergrundwissen in der Domäne ist an vielen Stellen des Data-Science-Prozesses relevant, z. B. 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.

Wissenschaftlichkeit

Wissenschaftlichkeit formuliert keinen Anspruch auf ein vollständig formalisiertes, akademisches, forschungsorientiertes Vorgehen. Während dies im Kontext von Forschungsprojekten so sein kann, bezieht sich der Aspekt der Wissenschaftlichkeit im Business-Kontext vor allem auf eine saubere Methodik, wie sie eben typischerweise als Eigenschaft wissenschaftlichen Arbeitens erwartet wird. Der definierte Projektauftrag ist in jeder einzelnen Projektphase entsprechend methodisch 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.

IT-Infrastruktur

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. Auch wenn die Nutzung spezifischer Hard- und Software häufig bereits organisationsintern festgelegt ist, sollten, wenn auch nicht die Auswahl, so doch zumindest die limitierenden und befähigenden Merkmale der IT-Infrastruktur in sämtlichen Projektphasen berücksichtigt werden.

Rollen

Kernrolle „Data Scientist"
Data Scientists sind Spezialisten für den Analysebereich eines Data-Science-Projekts.

Kernrolle „Data Engineer"
Data Engineers kümmern sich um die Beschaffung, Speicherung, Aufbereitung, Strukturierung und Weitergabe von Daten.

Kernrolle „Domänenexperte"
Domänenexperten sind Fachanwender oder Vertreter der Fachanwender.

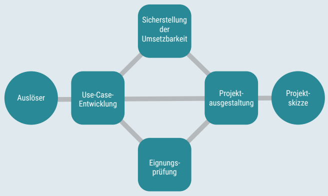
Kernrolle „Projektmanager"
Projektmanager planen, steuern und koordinieren den Gesamtablauf eines Data-Science-Projekts.

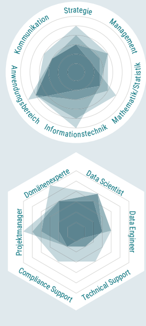
Ergänzende Rolle „Technischer Support"
Der technische Support umfasst alle Aufgaben, die erledigt werden müssen, um die technischen Voraussetzungen für die Durchführung des Data-Science-Projekts zu schaffen.

Ergänzende Rolle „Compliance-Support"
Der Compliance-Support ist für die Einhaltung gesetzlicher Vorgaben, die Kompatibilität mit organisationsinternen Regelwerken und das korrekte Verhalten der Projektmitarbeiter verantwortlich.

Projektauftrag

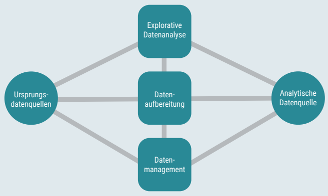
Innerhalb einer Domäne bestehende Probleme lösen eine Use-Case-Entwicklung aus. Die vielversprechendsten Use Cases werden anschließend zu einer Data-Science-Projektskizze ausgestaltet. Alle zugehörigen Aufgaben finden sich in der Phase des Projektauftrags. Durch die frühe, relativ umfassende Betrachtung des Projekts sind hier häufig auch umfassende Fähigkeiten in fast allen Kompetenzbereichen erforderlich.

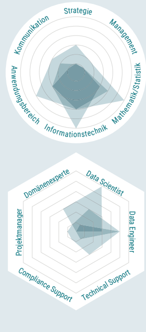




Datenbereitstellung

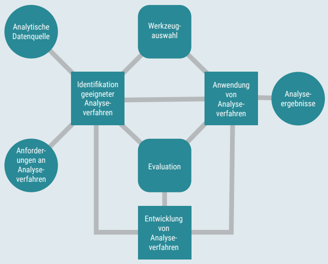
Innerhalb der Phase der Datenbereitstellung werden die Aktivitäten zusammengefasst, die dem Schlüsselbereich Daten zuzuordnen sind, weshalb der verwendete Begriff weit gefasst ist. Die Phase beinhaltet die Datenaufbereitung (von der Erfassung bis zur Speicherung), das Datenmanagement und eine explorative Analyse. Als Ergebnis dieser Phase entsteht eine für die weitere Analyse geeignete Datenquelle.

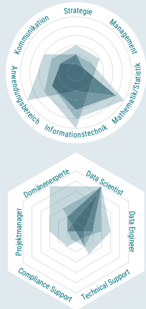




Analyse

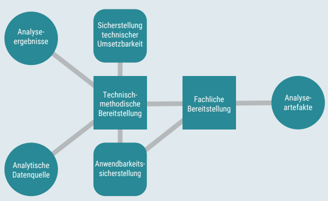
In einem Data-Science-Projekt können entweder bestehende Verfahren angewendet oder es müssen zunächst neue Verfahren entwickelt werden – die entsprechende Entscheidung ist eine eigene Herausforderung. Die Phase umfasst daher nicht nur die Analysedurchführung, sondern auch angrenzende Tätigkeiten. Das Artefakt der Phase ist ein Analyseergebnis, das eine methodische und fachliche Evaluation durchlaufen hat.

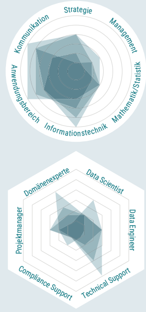




Nutzbarmachung

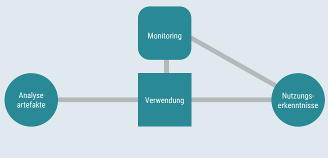
In der Phase der Nutzbarmachung wird eine anwendbare Form der Analyseergebnisse geschaffen. Projektspezifisch kann dies eine umfangreiche Betrachtung technischer, methodischer und fachlicher Aufgaben bedeuten oder pragmatisch gehandhabt werden. Die Analyseartefakte können sowohl Resultate als auch Modelle oder Verfahren selbst umfassen und werden den Adressaten in unterschiedlicher Form zur Verfügung gestellt.

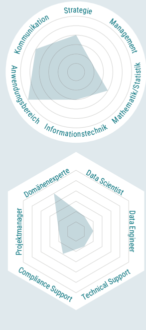




Nutzung

Die sich an die Projektdurchführung anschließende Verwendung von Analyseartefakten ist nicht als primärer Teil eines Data-Science-Projekts anzusehen. Ein Monitoring ist aber abhängig von der 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 (auch im Sinne iterativer Vorgehensweisen) zu erlangen.





Merkmalstragender Bereich
Kernaufgabe
Begleitende Aufgabe
Schnittstellenaufgabe

Kompetenzen
Rollen

Herausgeber:
NORDAKADEMIE gAG Hochschule der Wirtschaft
Köllner Chaussee 11, 25337 Elmshorn

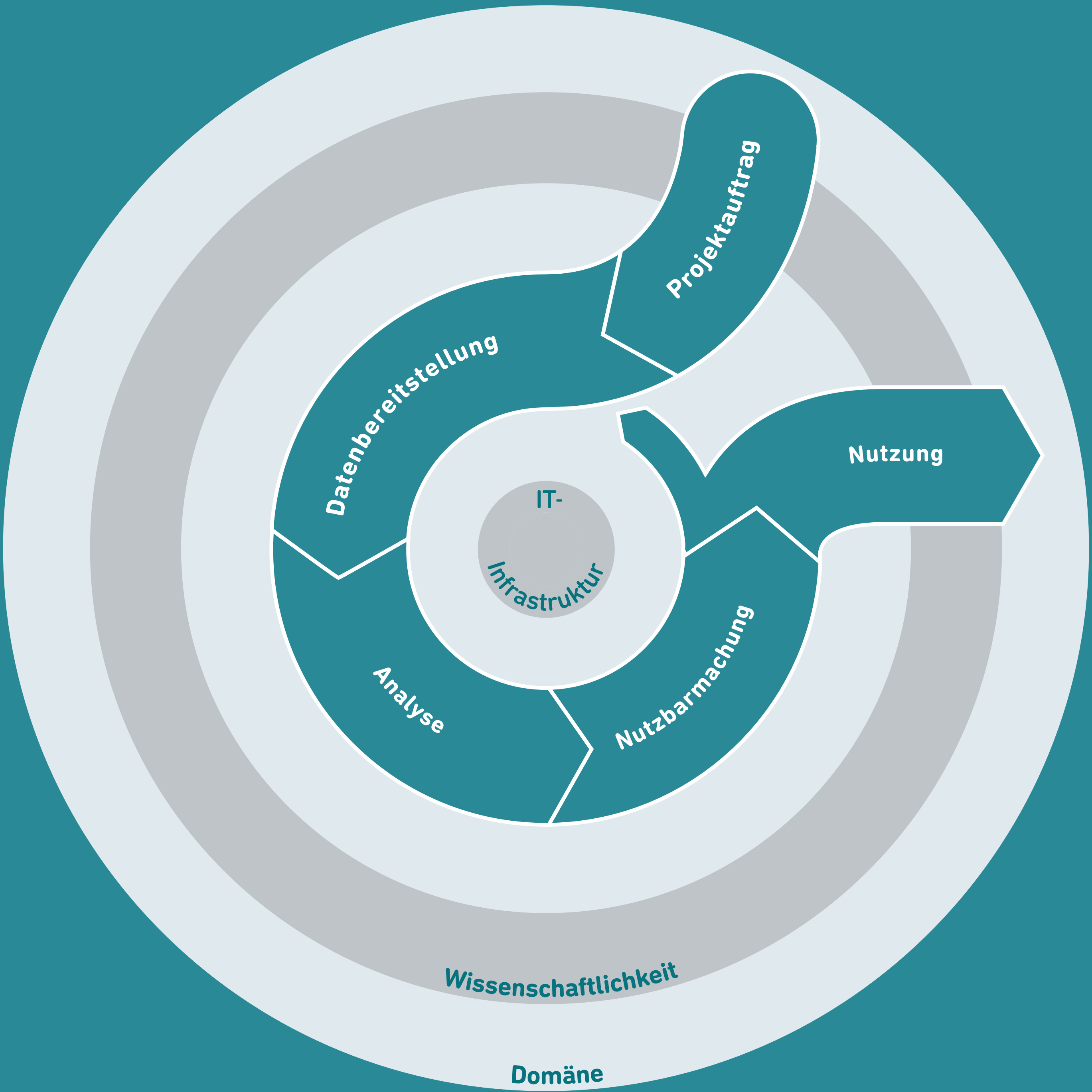
Gefördert durch die NORDAKADEMIE-Stiftung

Autoren:
Schulz, M., Neuhaus, U., Kaufmann, J., Kühnel, S., Alekozai, E.M., Rohde, H., Hoseini, S., Theuerkauf, R., Badura, D., Kerzel, U., Lanquillon, C., Daurer, S., Günther, M., Huber, L., Thiée, L.-W., zur Heiden, P., Passlick, J., Dieckmann, J., Schwade, F., Seyffarth, T., Badewitz, W., Rissler, R., Sackmann, S., Gölzer, P., Welter, F., Röth, J., Seidelmann, J., Haneke, U. (2022): DASC-PM V1.1 — Ein Vorgehensmodell für Data-Science-Projekte.

Grafik: Fritjof Wild

dasc°pm v1.1

Das Data Science Process Modell (DASC-PM) ist ein Vorgehensmodell für Data-Science-Projekte. Es beschreibt die projektrelevanten Schlüsselbereiche und zu durchlaufenden Phasen, erläutert die typischen Aufgaben innerhalb der Phasen und stellt die beteiligten Projektrollen und benötigten Kompetenzen dar.



Herausgeber:
NORDAKADEMIE gAG Hochschule der Wirtschaft
Köllner Chaussee 11, 25337 Elmshorn

Gefördert durch die NORDAKADEMIE-Stiftung

Autoren:
Schulz, M., Neuhaus, U., Kaufmann, J., Kühnel, S., Alekozai, E.M., Rohde, H., Hoseini, S., Theuerkauf, R., Badura, D., Kerzel, U., Lanquillon, C., Daurer, S., Günther, M., Huber, L., Thiée, L.-W., zur Heiden, P., Passlick, J., Dieckmann, J., Schwade, F., Seyffarth, T., Badewitz, W., Rissler, R., Sackmann, S., Gölzer, P., Welter, F., Röth, J., Seidelmann, J., Haneke, U. (2022): DASC-PM V1.1 — Ein Vorgehensmodell für Data-Science-Projekte.

Grafik: Fritjof Wild

asc°pm v1.1

ISBN: 978-3-9824465-0-9