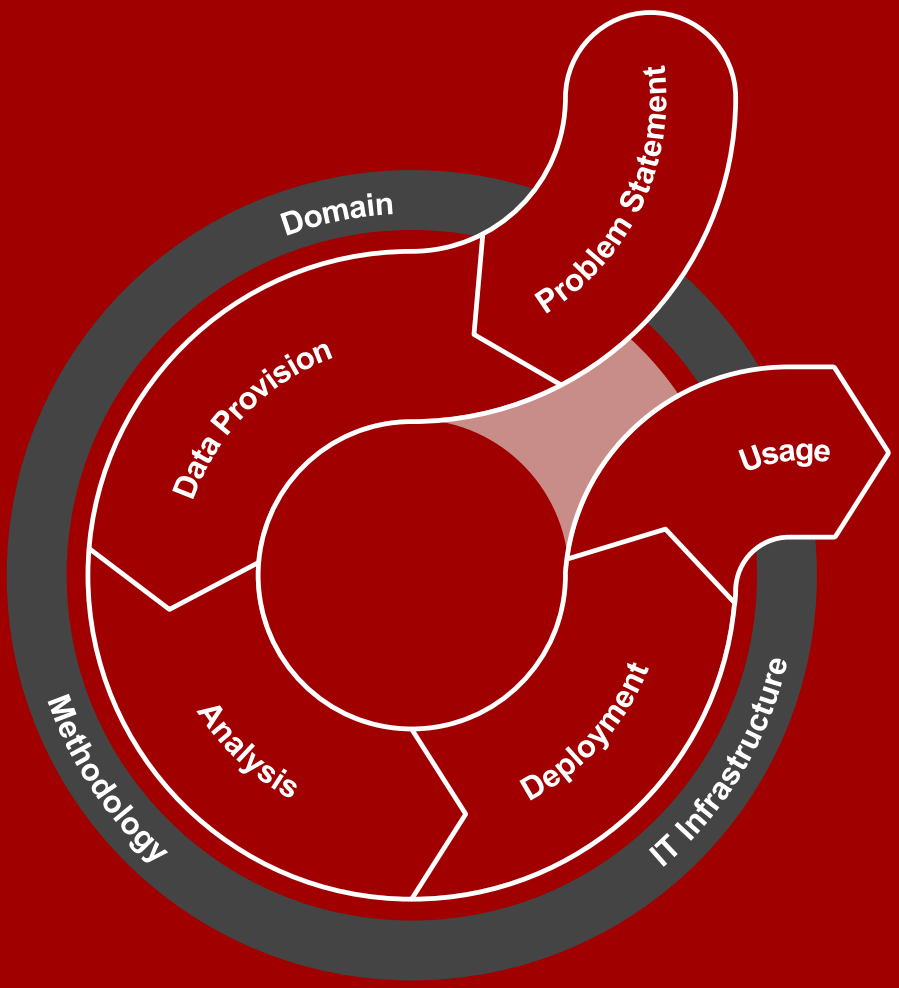
Michael Schulz ▪ Jens Kaufmann ▪ Stephan Kühnel ▪ Uwe Neuhaus

DASC-PM

Ein Vorgehensmodell für Data-Science- und KI-Projekte

v2.0

Heiko Rohde ▪ René Theuerkauf ▪ Carsten Lanquillon ▪ Stephan Daurer ▪ Jonas Dieckmann ▪ Stefan Sackmann ▪ Felix Welter ▪ Uwe Haneke ▪ Christian Beecks ▪ Martin Böhmer ▪ Bahne Christiansen ▪ André Drews ▪ Arne Ewald ▪ Viktor Harkov ▪ Franziska Herrmann ▪ Tim Hilbig ▪ Dirk Johannßen ▪ Lutz Kretschmann ▪ Bernhard Meussen ▪ Christian Pasold ▪ Stefan Rösl ▪ Klaus Schmerler ▪ Theo Schnelle de Lourenco ▪ Martin Schultz ▪ Michael Seifert ▪ Marcus Soll ▪ Robert Stahlbock ▪ Niklas Ullmann



Die vorliegende Version des Werks
DASC-PM - Ein Vorgehensmodell für Data-Science- und KI-Projekte – v2.0
basiert auf der Publikation

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.

DASC-PM v1.1 - Ein Vorgehensmodell für Data-Science-Projekte

NORDAKADEMIE gAG Hochschule der Wirtschaft, Elmshorn, 2022
ISBN: 978-3-9824465-0-9, https://doi.org/10.25673/85296.

die unter der Creative Commons (CC) Lizenz BY 4.0 veröffentlicht wurde
(https://creativecommons.org/licenses/by/4.0/)

Dieses Werk ist lizenziert unter einer Creative Commons
Namensnennung 4.0 International Lizenz.
https://creativecommons.org/licenses/by/4.0/

ISBN: 978-3-9824465-3-0 ISBN: 978-3-9824465-4-7 (eBook)

DOI: 10.25673/123274

Elmshorn 2026

info@dasc-pm.org

Herausgeber

NORDAKADEMIE gAG Hochschule der Wirtschaft
Köllner Chaussee 11
25337 Elmshorn

Förderhinweis

Die inhaltlichen Änderungen gegenüber der Version 1.1 wurden zu großen Teilen im Rahmen eines Workshops erarbeitet. Die Durchführung des Workshops sowie die Veröffentlichung dieses Dokumentes wurden durch eine Förderung der NORDAKADEMIE-Stiftung ermöglicht.

Die Aufbereitung der Ergebnisse im Themenfeld *Kompetenzen und Rollen* wurde durch das Projekt *Digital Learning Campus* unterstützt. Der Digital Learning Campus wird gefördert vom Land Schleswig-Holstein und der Europäischen Union. Weitere Informationen unter: dlc.sh

Modellgrafik: Mit freundlicher Unterstützung von Eike Behnke.

Inhaltsverzeichnis

Abbildungsverzeichnis

Vorwort zur Version 2.0

Mit der vorliegenden Version 2.0 von DASC-PM wird das ursprüngliche Vorgehensmodell inhaltlich und strukturell weiterentwickelt. Aufbauend auf den Erfahrungen aus der Anwendung der Versionen 1.0 und 1.1 wurde das Modell überarbeitet, um seine Verständlichkeit zu erhöhen und seine Nutzung in unterschiedlichen Projektkontexten zu erleichtern.

Seit der Veröffentlichung der ersten Version im Jahr 2020 hat sich das Umfeld datengetriebener Projekte weiter dynamisch entwickelt. Insbesondere Fortschritte im Bereich der Künstlichen Intelligenz haben neue Anforderungen an Methoden, Infrastruktur und Projektorganisation hervorgebracht. Gleichzeitig hat sich in zahlreichen praktischen Anwendungen gezeigt, welche Elemente des Modells sich bewährt haben und an welchen Stellen Anpassungen sinnvoll sind. Die vorliegende Version greift diese Erfahrungen auf und führt sie in einer überarbeiteten Darstellung zusammen.

Ein zentrales Ziel der Version 2.0 besteht darin, das Modell stärker ergebnisorientiert darzustellen. Frühere Fassungen des Dokuments enthielten an mehreren Stellen Hinweise auf den Entwicklungsprozess des Modells und die Beiträge einzelner Mitglieder der Arbeitsgruppe. In der vorliegenden Version liegt der Fokus stärker auf dem Modell selbst und den daraus abgeleiteten Strukturen und Aufgaben. In diesem Zusammenhang wurde auch die Struktur des Dokuments überarbeitet. Die in früheren Versionen dargestellten Schlüsselbereiche dienten ursprünglich der Strukturierung der Modellentwicklung und bildeten die Grundlage für die Ableitung der Projektphasen. In der praktischen Anwendung zeigte sich jedoch, dass ihre zusätzliche Darstellung neben den Phasen des Modells zu Unklarheiten führen kann. In Version 2.0 konzentriert sich die Darstellung daher stärker auf das eigentliche Vorgehensmodell und seine Phasen.

Die grundlegenden Phasen von DASC-PM bleiben dabei erhalten. Sie wurden jedoch inhaltlich überprüft und teilweise präzisiert, um ihre Anwendbarkeit in unterschiedlichen Arten daten- und analyseorientierter Projekte sicherzustellen. Dabei wurde insbesondere darauf geachtet, dass Aktivitäten im Kontext von Künstlicher Intelligenz in allen Phasen des Modells berücksichtigt werden können. Ziel dieser Anpassungen ist es, eine breite Anschlussfähigkeit des Modells zu gewährleisten, ohne es auf eine spezifische Klasse von KI-Projekten zu verengen.

Die Darstellung der im Projektkontext benötigten Kompetenzen wurde überarbeitet. Rückmeldungen aus Workshops und Arbeitsgruppen wurden zusammengeführt und in einer strukturierten Form ausgewertet. Um das Modell zudem langfristig stabil gegenüber sich wandelnden Berufsbezeichnungen zu halten, wird künftig stärker mit Rollentypen gearbeitet. Dabei wird zwischen Rollen im Projektkern, Kooperationsrollen sowie Rollen mit Steuerungs- und Einflussmöglichkeiten unterschieden. Konkrete Rollenbezeichnungen werden weiterhin beispielhaft genannt, stehen jedoch nicht mehr im Mittelpunkt der Modellbeschreibung.

Darüber hinaus wurden auch die Darstellung des Modells und einzelne Begriffe innerhalb der Modellabbildung überarbeitet. Künftig wird für die Modellgrafik eine englischsprachige Terminologie verwendet, um die internationale Anschlussfähigkeit des Modells zu erhöhen und seine Nutzung in internationalen Projekten zu erleichtern. Die deutsche Version wurde ferner in gendergerechter Sprache verfasst. Zur Vereinfachung des Sprachgebrauchs wird DASC-PM in diesem Dokument außerdem nur noch über sein Kürzel referenziert, nicht mehr über seine vollständige Bezeichnung (Data Science Process Model).

Wie bereits in den vorherigen Versionen versteht sich DASC-PM weiterhin als strukturierender Rahmen für daten- und analyseorientierte Projekte. Das Modell soll helfen, typische Aufgaben, Rollen und Zusammenhänge in solchen Projekten nachvollziehbar zu machen und eine gemeinsame Grundlage für die Kommunikation zwischen unterschiedlichen Beteiligten zu schaffen. Wir hoffen, dass die vorliegende Version dazu beiträgt, das Verständnis für die Struktur von Data-Science- und KI-Projekten weiter zu verbessern und Anwender:innen eine praktische Orientierung für ihre Projekte zu bieten.

Elmshorn, Flensburg, Halle (Saale) und Mönchengladbach im Juni 2026

DASC-PM-Kernteam

Teil A
Allgemeine Überlegungen und Gesamtmodell

1 Data Science und Künstliche Intelligenz

Der Begriff *Data Science* hat in den letzten Jahren sowohl in der Wissenschaft als auch in der Praxis erkennbar an Bedeutung gewonnen. In zahlreichen Organisationen werden datengetriebene Analyseverfahren eingesetzt, um Zusammenhänge in großen Datenbeständen zu erkennen, Prognosen zu erstellen oder Entscheidungsprozesse zu unterstützen. Gleichzeitig hat sich eine Vielzahl von Begriffen etabliert, die teilweise ähnliche oder überlappende Inhalte beschreiben.

Trotz der zunehmenden Verbreitung des Begriffs *Data Science* existiert bis heute keine allgemein akzeptierte und einheitliche Definition von diesem. In wissenschaftlichen Veröffentlichungen wird er in unterschiedlichen Disziplinen verwendet und stets aus der Perspektive der jeweiligen Fachgebiete interpretiert. In der praktischen Anwendung ist die Bandbreite der Begriffsverwendungen häufig noch größer. Daraus ergeben sich unterschiedliche Erwartungen an Inhalte, Methoden und Zielsetzungen von Data-Science-Aktivitäten.

Vor diesem Hintergrund verfolgt das vorliegende Dokument das Ziel, ein konsistentes Begriffsverständnis für den Kontext von Data-Science-Projekten zu schaffen. Dabei wird Data Science als ein interdisziplinäres Fachgebiet verstanden, das Methoden aus verschiedenen wissenschaftlichen Disziplinen miteinander verbindet und in unterschiedlichen Anwendungsfeldern eingesetzt werden kann.

Auf dieser Grundlage wird im Folgenden eine Definition von Data Science vorgestellt, die als Ausgangspunkt für die weiteren Ausführungen im Dokument dient. Die Merkmale dieser Definition sind in Abbildung 1 dargestellt und werden nachfolgend detaillierter betrachtet.

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.

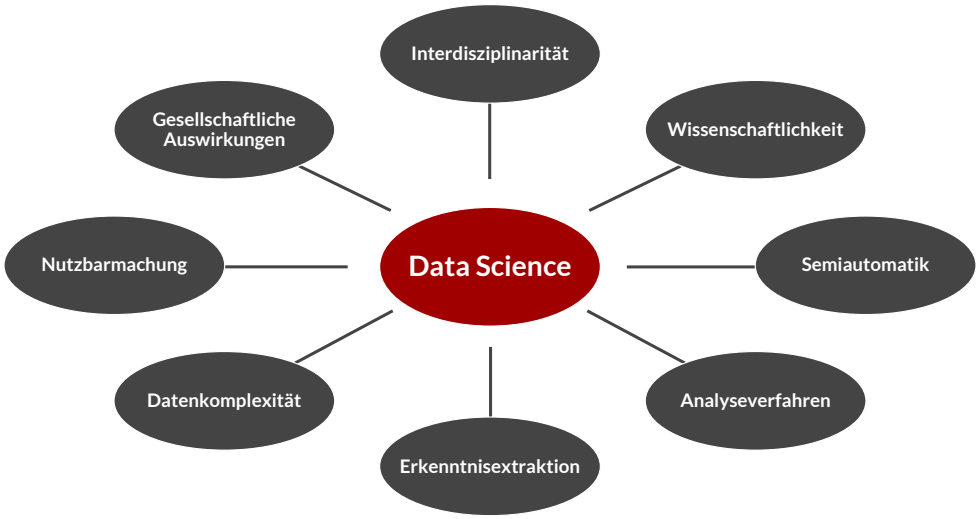


Abbildung 1: Merkmale der Data Science

1.1 Merkmale der Data Science

Interdisziplinarität

Die Interdisziplinarität stellt ein zentrales Merkmal der Data Science dar. Sie ergibt sich aus der häufig notwendigen Zusammenarbeit verschiedener wissenschaftlicher Disziplinen, die gemeinsam zur Analyse komplexer Daten beitragen. Besonders relevant sind hierbei unter anderem die Mathematik – insbesondere Statistik und Numerik – und die Informatik. Auch weitere Disziplinen, beispielsweise die Linguistik, leisten Beiträge zur methodischen und konzeptionellen Auseinandersetzung mit Daten. In all diesen Bereichen existieren bereits seit längerer Zeit Ansätze zur wissenschaftlichen Analyse und Interpretation von Daten (Provost & Fawcett, 2013).

Das unter dem Begriff *Data Science* zusammengefasste Fachgebiet hat sich insbesondere dort entwickelt, wo die Methoden einzelner Disziplinen allein nicht mehr ausreichten, um aktuellen Herausforderungen im Umgang mit Daten zu begegnen. Dazu zählen etwa stark wachsende Datenmengen, häufig unstrukturierte Datenformate oder dynamisch entstehende Datenströme. Gleichzeitig haben technologische Entwicklungen – beispielsweise die zunehmende Verfügbarkeit digitaler Informationen, sinkende Kosten für Datenspeicherung sowie gestiegene Rechenkapazitäten – den Einsatz zunehmend komplexer Analyseverfahren ermöglicht (McAfee & Brynjolfsson, 2012; Donoho, 2017). Diese Entwicklungen haben wesentlich dazu beigetragen, dass sich Data Science als eigenständiges interdisziplinäres Arbeitsfeld etabliert hat.

Interdisziplinarität zeigt sich darüber hinaus auch in der Vielfalt der Anwendungsdomänen, in denen Data Science eingesetzt wird. Während viele Darstellungen das Fachgebiet traditionell vor allem im wirtschaftlichen Kontext verorten, beschränkt sich seine Anwendung nicht auf diesen Bereich. Vielmehr findet Data Science in zahlreichen wissenschaftlichen und praktischen Domänen Anwendung, etwa in der Biologie, der Medizin, der Physik oder der Astronomie. In diesen und vielen weiteren Feldern unterstützt sie die Analyse komplexer Datenbestände und trägt zur Gewinnung neuer Erkenntnisse bei.

Wissenschaftlichkeit

Der Begriff *Data Science* deutet bereits darauf hin, dass Datenanalysen nicht ausschließlich als technische Tätigkeit verstanden werden, sondern auf einem wissenschaftlich geprägten Vorgehen beruhen. Diese Wissenschaftlichkeit zeigt sich unter anderem darin, dass Data-Science-Aktivitäten häufig auf einen allgemeinen Erkenntnisgewinn abzielen. Neben der unmittelbaren Anwendung von Analyseergebnissen können daher auch weitergehende Fragestellungen im Fokus stehen, etwa die Untersuchung der Eignung bestimmter Verfahren für spezifische Problemstellungen, die Bewertung ihrer Aussagekraft im Kontext der zugrunde liegenden Datenbasis oder die Analyse der Komplexität unterschiedlicher Methoden.

Darüber hinaus zeichnet sich ein wissenschaftliches Vorgehen dadurch aus, dass die untersuchten Problemstellungen nicht trivial sind und die eingesetzten Verfahren nachvollziehbar, systematisch und reproduzierbar angewandt werden. Dazu gehört insbesondere, dass Analyseprozesse dokumentiert werden und ihre Ergebnisse grundsätzlich überprüfbar bleiben. Diese Aspekte tragen dazu bei, dass Erkenntnisse aus Datenanalysen nicht nur situativ nutzbar sind, sondern auch in einem breiteren fachlichen Kontext eingeordnet werden können.

Auch in unternehmerischen Anwendungen spielt dieser wissenschaftliche Hintergrund eine Rolle. Viele der heute verwendeten Analyseverfahren wurden ursprünglich im wissenschaftlichen Umfeld entwickelt und anschließend in praktische Anwendungen übertragen. Die Tiefe der wissenschaftlichen Auseinandersetzung kann dabei je nach Anwendungsfeld variieren. Während in Forschungsprojekten häufig neue Methoden entwickelt oder bestehende Verfahren detailliert untersucht werden, steht in vielen praktischen Anwendungen die sachgerechte Nutzung vorhandener Methoden im Vordergrund. In solchen Kontexten kann das Vorgehen eher den Charakter eines ingenieurmäßigen Umgangs mit etablierten Verfahren annehmen.

Analyseverfahren

Die explizite Nennung einzelner Algorithmen oder Algorithmengruppen in einer Definition der Data Science ist aufgrund der hohen Entwicklungsgeschwindigkeit des Fachgebiets nicht sinnvoll. Eine solche Festlegung würde das Fachgebiet implizit auf bestimmte Methoden beschränken und zukünftige Entwicklungen nur unzureichend berücksichtigen. Darüber hinaus ist der Begriff *Algorithmus* im gegebenen Zusammenhang nicht immer adäquat, da nicht alle Analysen ausschließlich auf algorithmischen Verfahren beruhen und viele der eingesetzten Methoden auch außerhalb der Data Science entwickelt wurden.

Aus diesem Grund wird in der Definition der allgemeinere Begriff *Analyseverfahren* verwendet. In Verbindung mit der Anwendung auf Daten beschreibt er den methodischen Kern der Data Science. Analyseverfahren können dabei sehr unterschiedliche Formen annehmen. Es können unter anderem hypothesenprüfende und hypothesenfreie Analysen mit deskriptivem, prädiktivem oder präskriptivem Ziel durchgeführt werden. Ziel solcher Analysen ist häufig das Aufdecken von Mustern, Trends und Zusammenhängen in Daten oder die Unterstützung von Optimierungsprozessen.

Welche Analyseverfahren in einem konkreten Projekt eingesetzt werden, hängt stark vom jeweiligen Anwendungsfall sowie von den verfügbaren Daten ab. In vielen Fällen werden bestehende Verfahren verwendet. Es kann jedoch auch erforderlich sein, Analyseverfahren weiterzuentwickeln oder neue Ansätze zu entwickeln, wenn für eine bestimmte Problemstellung keine geeigneten Methoden verfügbar sind.

Semiautomatik

Die Anwendung von Analyseverfahren erfolgt semiautomatisch und umfasst sowohl menschliche als auch maschinelle Arbeitsschritte. Neben der Tatsache, dass Verfahren in der Regel nicht vollständig automatisiert werden können, sind auch hybride Lernverfahren zu nennen, die speziell dafür entwickelt wurden, das Zusammenspiel von Expert:innenwissen und Analyseverfahren zu unterstützen (Olivotti et al., 2018). Häufig sind hierfür leistungsfähige Hard- und Softwareplattformen erforderlich, die zusammen eine komplexe Infrastruktur bilden. Zudem führen Entwicklungen auf dem Feld der generativen KI, insbesondere im Bereich großer Sprachmodelle, zu Werkzeugen, die bei einzelnen Analysetypen zu einer höheren Automatisierung in der Vorbereitung, Durchführung und Interpretation der Ergebnisse führen können.

Abhängig vom konkreten Szenario kann eine weitgehende Automatisierung angestrebt werden. Dafür sind jedoch in der Regel vorbereitende manuelle Arbeitsschritte notwendig. Auch bleibt der Erkenntnisgewinn aus Datenanalysen letztlich an die Beteiligung menschlicher Akteure gebunden.

Erkenntnisextraktion und Datenkomplexität

Ein Ziel der Data Science ist die Extraktion von Erkenntnissen aus meist komplexen Daten. Diese unterscheiden sich beispielsweise hinsichtlich ihrer Struktur, ihrer Qualität, ihrer Vollständigkeit, ihrer Größe und ihrer Dimensionalität. Es kann sich dabei sowohl um statische Daten als auch um Datenströme handeln. Zudem können Daten in vielfältigen und teilweise komplexen Beziehungen zueinanderstehen. Die Analyse großer und heterogener Datenbestände erforderte die Entwicklung neuer Verfahren, die häufig unter dem Begriff *Big Data* zusammengefasst wurden.

Bevor Analyseverfahren auf Daten angewendet werden können, müssen diese in der Regel zunächst aus Quellsystemen extrahiert, aufbereitet und in geeigneter Form bereitgestellt werden. Für diese Schritte werden häufig ebenfalls komplexe technische Infrastrukturen benötigt.

Nutzbarmachung

Data Science umfasst nicht nur die Extraktion von Erkenntnissen aus Daten, sondern auch deren Nutzbarmachung. Diese kann beispielsweise in der Bereitstellung von Analyseergebnissen für Domänenexpert:innen oder andere Anwender:innen bestehen. Ebenso ist eine Integration der Ergebnisse in bestehende Systeme oder eine automatisierte Anwendung von Analysemodellen auf neue Daten möglich. Verschiedene Autor:innen stellen bei Data-Science-Projekten die Schaffung eines ökonomischen Mehrwerts in den Vordergrund. In der vorliegenden Definition wird jedoch bewusst allgemeiner von Nutzbarmachung gesprochen, um neben wirtschaftlichen Zielsetzungen auch wissenschaftliche oder andere Formen der Nutzung einzuschließen.

Sowohl die semiautomatische Extraktion von Erkenntnissen als auch die Aufbereitung und Bereitstellung von Daten sowie deren spätere Nutzbarmachung erfordern in vielen Fällen eine geeignete IT-Infrastruktur. Diese umfasst Hard- und Softwarekomponenten, die an die jeweiligen Projektanforderungen angepasst werden müssen. Beispiele hierfür sind skalierbare Systemarchitekturen, die Verarbeitung verteilter Daten oder die Nutzung von Cloud-Infrastrukturen.

Gesellschaftliche Auswirkungen

Der Einsatz datengetriebener Analyseverfahren kann vielfältige gesellschaftliche Auswirkungen haben. Daher gehört es zur Data Science, sich mit den ethischen, rechtlichen und gesellschaftlichen Fragestellungen auseinanderzusetzen, die sich aus der Nutzung von Daten und Analyseverfahren ergeben. Dies betrifft sowohl die verwendeten Daten als auch die daraus abgeleiteten Analyseergebnisse.

Ein zentraler Aspekt ist der verantwortungsvolle Umgang mit Daten. Insbesondere beim Umgang mit rechtlich geschützten oder personenbezogenen Daten sind rechtliche Rahmenbedingungen zu berücksichtigen, etwa im Kontext des Datenschutzes oder der Datensicherheit. Darüber hinaus können auch Fragen der Transparenz und Nachvollziehbarkeit algorithmischer Entscheidungen eine wichtige Rolle spielen, insbesondere wenn Analysemodelle in Entscheidungsprozesse integriert werden.

Neben rechtlichen Anforderungen werden zunehmend auch ethische Fragestellungen diskutiert. Dazu gehören unter anderem mögliche Verzerrungen in Datenbeständen oder Analyseverfahren, die zu systematischen Benachteiligungen bestimmter Personengruppen führen können. Ebenso stellt sich die Frage, in welchem Umfang automatisierte Analysesysteme Entscheidungen vorbereiten, unterstützen oder treffen sollten und welche Rolle menschliche Verantwortung in solchen Kontexten einnimmt.

Vor diesem Hintergrund wird Data Science nicht ausschließlich als technische Disziplin verstanden. Vielmehr gehört auch die aktive Beteiligung am gesellschaftlichen Diskurs über Chancen, Risiken und Grenzen datengetriebener Analysen zu den Aufgaben dieses Fachgebiets. Ziel ist es, den Einsatz von Daten und Analyseverfahren so zu gestalten, dass ihre Potenziale genutzt werden können, ohne dabei gesellschaftliche oder rechtliche Rahmenbedingungen zu vernachlässigen.

1.2 Künstliche Intelligenz in der Verschränkung mit Data Science

In den letzten Jahren hat neben Data Science auch Künstliche Intelligenz (KI), im Englischen als Artificial Intelligence (AI) bezeichnet, stark an Bedeutung gewonnen. Die Begriffe werden in wissenschaftlichen Veröffentlichungen, in der öffentlichen Diskussion sowie in der praktischen Anwendung häufig in ähnlichen Zusammenhängen verwendet. Dabei zeigt sich, dass ihre Bedeutungen teilweise überlappen, je nach Kontext aber in Teilen auch unterschiedlich interpretiert werden.

Grundsätzlich beschreibt KI ein breites Forschungsfeld mit Wurzeln in der Informatik, das sich mit der Entwicklung von Systemen beschäftigt, die menschliche Fähigkeiten wie logisches Denken, Lernen, Planen und Kreativität imitieren können (in Anlehnung an Europ. Parl., 2026). Dazu gehören beispielsweise das Erkennen von Mustern, das Treffen von Entscheidungen oder die Verarbeitung natürlicher Sprache. Innerhalb dieses Forschungsfeldes haben sich diverse Teilbereiche entwickelt, zu denen unter anderem Machine Learning (ML) zählt. Dabei bezeichnet Machine Learning eine Klasse von Verfahren, bei denen Modelle in der Regel auf Grundlage vorhandener Daten trainiert werden. Deep Learning stellt eine spezielle Ausprägung solcher Verfahren dar, bei der vielschichtige neuronale Netze eingesetzt werden, etwa zur Verarbeitung von Bildern, Texten oder Sprache.

Die Beziehungen zwischen diesen Begriffen werden häufig in Form überlappender oder ineinander verschachtelter Konzepte dargestellt. Eine solche vereinfachte Einordnung zeigt beispielsweise Machine Learning als Teilbereich der Künstlichen Intelligenz, während Deep Learning wiederum eine spezielle Klasse von Machine-Learning-Verfahren darstellt. Abbildung 2 stellt dies exemplarisch dar.

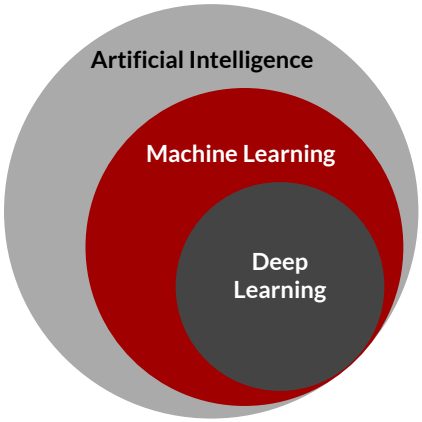


Abbildung 2: AI, ML und Deep Learning

Data Science weist in vielen Aspekten enge Berührungspunkte mit den beschriebenen Bereichen auf. Insbesondere Methoden des Maschinellen Lernens werden häufig eingesetzt, um aus Daten Modelle abzuleiten oder Prognosen zu erstellen. Gleichzeitig umfasst Data Science ein breites Spektrum weiterer Aktivitäten. Neben der eigentlichen Analyse gehören hierzu auch die Beschaffung und Aufbereitung von Daten, die Auswahl geeigneter Analyseverfahren, die Interpretation der Ergebnisse im Kontext einer Anwendungsdomäne sowie deren Nutzbarmachung. Weitere Bereiche der KI, z. B. Robotik, stehen wiederum nicht im Fokus von Data Science.

Vor diesem Hintergrund lassen sich Data Science und KI weder vollständig voneinander trennen noch eindeutig hierarchisch zueinander einordnen. In vielen praktischen Anwendungen überschneiden sich beide Felder. Abhängig von der Perspektive kann also von *Data-Science-Projekten* oder *KI-Projekten* oder *ML-Projekten* oder noch weiteren Varianten die Rede sein.

Zur Vereinfachung des Sprachgebrauchs und mit Rücksicht auf die kontinuierliche Weiterentwicklung des Modells werden in DASC-PM vorrangig die Begriffe Data Science und Data-Science-Projekte verwendet, um die im Fokus stehenden Aktivitäten zu beschreiben. DASC-PM lässt sich aber bezeichnungsunabhängig auf viele Varianten der oben genannten Projekttypen anwenden.

2 Vorgehensmodelle für Data-Science-Projekte

Die Durchführung von Data-Science-Projekten ist in der Regel mit einer Vielzahl unterschiedlicher Aufgaben verbunden. Diese reichen von der Identifikation geeigneter Fragestellungen über die Beschaffung und Aufbereitung von Daten bis hin zur Entwicklung, Bewertung und Nutzung von Analysemodellen. Hinzu kommen organisatorische, technische und fachliche Anforderungen, die im Verlauf eines Projekts berücksichtigt werden müssen. Vor diesem Hintergrund stellt sich die Frage, wie solche Projekte strukturiert geplant und durchgeführt werden können.

Vorgehensmodelle bieten hierfür einen Rahmen, der typische Aktivitäten eines Projekts beschreibt und deren Zusammenhänge verdeutlicht. Sie dienen dazu, komplexe Projektabläufe zu strukturieren, die Kommunikation zwischen verschiedenen Beteiligten zu unterstützen und ein gemeinsames Verständnis über Aufgaben, Ergebnisse und Verantwortlichkeiten zu schaffen. Gleichzeitig können sie dazu beitragen, die Nachvollziehbarkeit und Wiederholbarkeit von Analysetätigkeiten zu erhöhen.

Im Umfeld datengetriebener Analysen wurden bereits früh Vorgehensmodelle entwickelt, die eine systematische Struktur für entsprechende Projekte bereitstellen. Zu den bekanntesten Modellen zählen der *Knowledge-Discovery-in-Databases*-Prozess (KDD, Fayyad et al., 1996) sowie der *Cross Industry Standard Process for Data Mining* (CRISP-DM) (Wirth & Hipp, 2000). Beide Modelle beschreiben eine Abfolge typischer Aktivitäten, die von der Vorbereitung und Exploration der Daten über die Anwendung von Analyseverfahren bis hin zur Bewertung der Ergebnisse reichen. Zusätzlich können weitere Modelle betrachtet werden, die speziell für den Data-Science-Bereich entwickelt wurden.

Diese Modelle haben wesentlich dazu beigetragen, datenorientierte Analyseprojekte zu strukturieren und ein gemeinsames Verständnis zentraler Projektaktivitäten zu schaffen. Gleichzeitig zeigt sich in der praktischen Anwendung, dass Data-Science-Projekte häufig zusätzliche Aspekte berücksichtigen müssen. Dazu zählen beispielsweise die Integration von Analyseergebnissen in operative Systeme, die Zusammenarbeit interdisziplinärer Teams oder der Aufbau geeigneter technischer Infrastrukturen. Vor allem im Zusammenhang mit CRISP-DM sind Bestrebungen zu identifizieren, die dieses Modell an die Anforderungen von Data-Science-Projekten anpassen sollen, z. B. CRISP-ML(Q) (Studer et al., 2021).

DASC-PM greift diese Überlegungen im Kontext der aktuellen Entwicklungen auf und stellt ein Vorgehensmodell bereit, das speziell auf die Anforderungen moderner Data-Science- und KI-Projekte ausgerichtet ist. Ziel des Modells ist es, typische Phasen und Aktivitäten solcher Projekte strukturiert darzustellen und deren Zusammenhänge nachvollziehbar zu machen. Dabei soll das Modell sowohl als Orientierung für die Planung und Durchführung von Projekten dienen als auch als gemeinsame Grundlage für die Kommunikation zwischen den beteiligten Personen und Organisationseinheiten.

3 DASC-PM als Vorgehensmodell

Aufbauend auf den zuvor dargestellten Grundlagen wird im Folgenden DASC-PM als Vorgehensmodell für Data-Science-Projekte vorgestellt (vgl. Abbildung 3). Das Modell beschreibt typische Phasen und Aktivitäten, die bei der Planung, Durchführung und Nutzung datengetriebener Analysen auftreten können. Ziel ist es, eine strukturierte Darstellung dieser Aktivitäten zu bieten und deren Zusammenhänge nachvollziehbar zu machen.

Data-Science-Projekte zeichnen sich häufig durch eine hohe Komplexität aus. Unterschiedliche Kompetenzbereiche, technische Infrastrukturen sowie fachliche Anforderungen müssen zusammengeführt werden, um aus Daten belastbare Erkenntnisse zu gewinnen und diese anschließend nutzbar zu machen. DASC-PM bietet hierfür einen Rahmen, der typische Aufgabenbereiche strukturiert und die Kommunikation zwischen den beteiligten Personen erleichtern kann.

Das Modell stellt dabei keinen starren Prozess dar. Vielmehr beschreibt es eine strukturierte Sicht auf Aktivitäten, die in unterschiedlichen Projekten in variierender Form auftreten können. Je nach Anwendungsfall, organisatorischem Kontext oder technischer Infrastruktur können einzelne Schritte unterschiedlich ausgeprägt sein oder mehrfach durchlaufen werden. In der grafischen Abbildung des Modells ist diese Eigenschaft durch den hellroten Kreisteil dargestellt. Abschnitt 3.3 greift sie noch einmal separat auf.

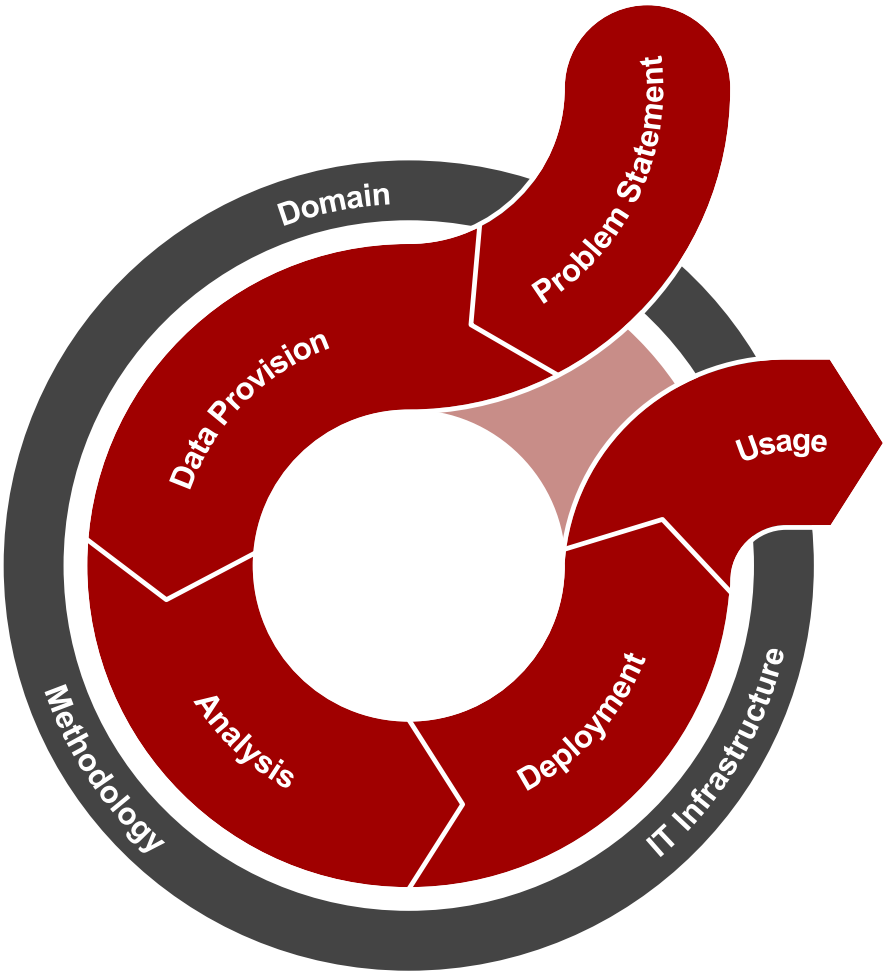


Abbildung 3: Data-Science-Vorgehensmodell DASC-PM

3.1 Phasen

Problem Statement

Das Problem Statement bildet den Ausgangspunkt eines Data-Science-Projekts. Innerhalb einer Domäne bestehende Fragestellungen oder Probleme können den Anstoß für die Entwicklung eines datengetriebenen Anwendungsfalls geben. In dieser Phase werden mögliche Use Cases identifiziert, bewertet und weiter konkretisiert. Ziel ist es, eine klar definierte Problemstellung zu formulieren und eine Projektskizze zu erstellen, die als Grundlage für die weitere Bearbeitung dient.

Data Provision

In der Phase der Data Provision werden alle Aktivitäten zusammengefasst, die mit der Beschaffung, Aufbereitung und Strukturierung von Daten verbunden sind. Dazu gehören beispielsweise die Identifikation relevanter Datenquellen, die Integration verschiedener Datenbestände sowie die Bereinigung und Transformation von Daten. Auch explorative Analysen können Teil dieser Phase sein, um ein erstes Verständnis der verfügbaren Daten zu gewinnen. Ergebnis dieser Phase ist eine Datenbasis, die für die anschließende Analyse geeignet ist.

Analysis

In der Analysephase werden geeignete Analyseverfahren ausgewählt und auf die verfügbaren Daten angewendet. Je nach Fragestellung können dabei bestehende Methoden eingesetzt oder neue Verfahren entwickelt werden. Neben der eigentlichen Anwendung von Analyseverfahren umfasst diese Phase auch Aktivitäten wie die Auswahl geeigneter Werkzeuge oder die Evaluation der Ergebnisse. Ziel ist es, aus den Daten belastbare Erkenntnisse abzuleiten und deren Aussagekraft zu bewerten.

Deployment

In dieser Phase werden die Ergebnisse der Analyse in eine Form überführt, die eine praktische Nutzung ermöglicht. Dies kann beispielsweise durch die Integration von Analysemodellen in bestehende Systeme, durch die Bereitstellung von Visualisierungen oder durch andere Formen der Ergebnisaufbereitung erfolgen. Abhängig vom jeweiligen Projekt können dabei sowohl technische als auch fachliche Aspekte berücksichtigt werden.

Usage

Die Nutzung der Analyseartefakte erfolgt häufig außerhalb des eigentlichen Data-Science-Projekts. Dennoch können sich aus der praktischen Anwendung wichtige Erkenntnisse ergeben, etwa im Hinblick auf die Qualität von Modellen oder mögliche Verbesserungen. Ein Monitoring der Nutzung kann dazu beitragen, Anpassungsbedarfe zu identifizieren oder neue Projektansätze abzuleiten.

3.2 Übergreifende Aspekte im Modell

Neben den beschriebenen Projektphasen existieren in DASC-PM mehrere Aspekte, die übergreifend berücksichtigt werden müssen und alle Phasen des Modells beeinflussen. Diese Aspekte bilden einen Rahmen für die Durchführung von Data-Science-Projekten und prägen die Ausgestaltung der einzelnen Aktivitäten.

Domain

Data-Science-Projekte werden stets innerhalb einer bestimmten Anwendungsdomäne durchgeführt. Neben dem Problem Statement als Ausgangspunkt eines Projekts ergeben sich in den einzelnen Phasen häufig domänenspezifische Anforderungen oder Rahmenbedingungen, die die Durchführung der jeweiligen Aufgaben beeinflussen. Fachliches Wissen über Prozesse, Datenquellen und Zusammenhänge innerhalb der betrachteten Domäne ist daher für viele Projektaktivitäten von zentraler Bedeutung. Die Domäne stellt somit einen durchgängigen Bezugsrahmen dar und muss in allen Phasen des Modells berücksichtigt werden.

Methodology

Der Aspekt der Wissenschaftlichkeit aus der Data-Science-Definition bedeutet im Kontext von Data-Science-Projekten nicht zwangsläufig ein vollständig formalisiertes oder ausschließlich forschungsorientiertes Vorgehen, wie es in wissenschaftlichen Forschungsprojekten häufig der Fall ist. Im praktischen Anwendungskontext bezieht es sich vor allem auf eine stringente und nachvollziehbare Methodik. Dazu gehören insbesondere eine strukturierte Vorgehensweise, eine angemessene Dokumentation sowie eine systematische Bewertung von Analyseergebnissen. Der konkrete Grad der erforderlichen Wissenschaftlichkeit hängt von den jeweiligen Projektzielen, den organisatorischen Rahmenbedingungen und den Besonderheiten der betrachteten Domäne ab.

IT Infrastructure

Die Durchführung von Data-Science-Projekten ist in hohem Maße von der zugrunde liegenden IT-Infrastruktur abhängig. Viele Projektaktivitäten – etwa die Speicherung, Verarbeitung und Analyse von Daten – setzen den Einsatz geeigneter Hard- und Softwarekomponenten voraus. Das konkrete Ausmaß der benötigten IT-Unterstützung kann jedoch je nach Projekt erheblich variieren. In vielen Organisationen sind bestimmte technische Plattformen oder Werkzeuge bereits vorgegeben. Dennoch sollten sowohl die Möglichkeiten als auch die Einschränkungen der vorhandenen Infrastruktur in allen Projektphasen berücksichtigt werden. Dies betrifft beispielsweise Aspekte wie verfügbare Rechenkapazitäten, Datenzugänge, Systemarchitekturen oder die Möglichkeit, bestehende Infrastrukturen bei Bedarf zu erweitern.

3.3 Iterationen und Abbruch bei der Modellnutzung

Die Darstellung des DASC-PM vermittelt zunächst den Eindruck eines seriellen Vorgehens. Diese Struktur dient in erster Linie der Übersichtlichkeit und der Beschreibung typischer Aufgabenbereiche innerhalb eines Data-Science-Projekts. In der praktischen Durchführung verlaufen solche Projekte jedoch nur selten strikt linear. Vielmehr ist es häufig erforderlich, einzelne Phasen mehrfach zu durchlaufen oder zu früheren Aktivitäten zurückzukehren.

Das Modell ist daher ausdrücklich als iteratives Vorgehen zu verstehen. Neue Erkenntnisse aus einer späteren Phase können dazu führen, dass vorherige Schritte angepasst oder erneut durchgeführt werden müssen. Beispielsweise kann die Nutzbarmachung und projektinterne Nutzung eines entwickelten Modells Hinweise darauf liefern, dass bestimmte Analyseverfahren weiter verbessert werden sollten.

Generell gilt, dass Iterationen nur diejenigen Phasen erneut durchlaufen müssen, die für die jeweilige Fragestellung relevant sind. Je nach Projektkontext kann es daher ausreichen, einzelne Aktivitäten in reduziertem Umfang zu wiederholen. Rücksprünge zwischen verschiedenen Phasen sind im Projektverlauf ebenso möglich, wenn das Projektteam dies für erforderlich hält. Darüber hinaus ist der Abbruch eines Data-Science-Projekts grundsätzlich in jeder Phase möglich. Auch wenn dadurch das im Problem Statement definierte Ziel nicht vollständig erreicht wird, bedeutet dies nicht zwangsläufig, dass das Projekt als gescheitert betrachtet werden muss. Erkenntnisse, die bis zu diesem Zeitpunkt gewonnen wurden, können dazu beitragen, das Verständnis des betrachteten Problems zu verbessern oder neue Fragestellungen zu identifizieren.

Durch diese iterative Struktur lässt sich DASC-PM in unterschiedlichen Projektmethoden einsetzen. Die Phasen des Modells beschreiben dabei Aufgabenbereiche, die innerhalb eines Projekts immer wieder adressiert werden können. Auf diese Weise kann das Modell sowohl in klassischen Projektstrukturen als auch in agilen Vorgehensweisen mit wiederholtem Durchlaufen einzelner Projektzyklen angewendet werden.

4 Kompetenzen und Rollen

Die Durchführung von Data-Science-Projekten erfordert eine Vielzahl unterschiedlicher Kompetenzen. Diese betreffen sowohl methodische und technische Aspekte der Datenanalyse als auch organisatorische, fachliche und kommunikative Fähigkeiten. In der öffentlichen Diskussion wird häufig der Eindruck vermittelt, dass diese Kompetenzen in der Rolle eines Data Scientist zusammenlaufen. Diese Vorstellung greift jedoch zu kurz, da die erfolgreiche Durchführung datengetriebener Projekte in der Regel das Zusammenspiel verschiedener Fachrichtungen und Rollen voraussetzt.

In der Praxis zeigt sich daher, dass Data-Science-Projekte typischerweise von interdisziplinären Teams bearbeitet werden. Innerhalb solcher Teams werden unterschiedliche Kompetenzen zusammengeführt, die gemeinsam zur Lösung einer datenbezogenen Fragestellung beitragen.

Eine häufig verwendete Darstellung der grundlegenden Kompetenzbereiche in Data-Science-Projekten geht auf Conway (2010) zurück. In dieser Darstellung wird Data Science als Schnittmenge von mathematisch-statistischen, informationstechnischen sowie anwendungsspezifischen Kompetenzen beschrieben (vgl. Abbildung 4).

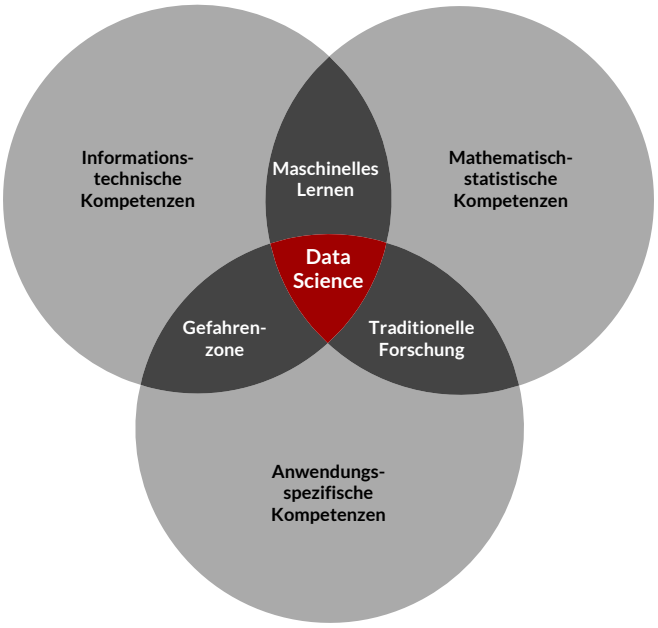


Abbildung 4: Kompetenzen von Data Scientists in Anlehnung an Conway (2010)

Die mathematisch-statistischen Kompetenzen umfassen insbesondere Kenntnisse und Fertigkeiten in Statistik, Mathematik und Modellierung. Diese Kompetenzen bilden die Grundlage für die Entwicklung und Anwendung von Analyseverfahren sowie für die Bewertung der Aussagekraft von Analyseergebnissen.

Die informationstechnischen Kompetenzen betreffen die Verarbeitung, Speicherung und Bereitstellung von Daten sowie die technische Umsetzung von Analyseverfahren. Dazu gehören beispielsweise Kenntnisse und Fertigkeiten in Programmiersprachen, Datenbanken oder verteilten Datenverarbeitungssystemen.

Die anwendungsspezifischen Kompetenzen schließlich beschreiben primär Kenntnisse und Fertigkeiten im jeweiligen Anwendungskontext eines Projekts. Hierzu zählen etwa fachliche Prozesse, relevante Fragestellungen oder die Interpretation von Analyseergebnissen im Kontext der jeweiligen Domäne.

Erst das Zusammenwirken dieser Kompetenzbereiche ermöglicht es, datengetriebene Fragestellungen angemessen zu bearbeiten. In vielen Projekten werden diese Kompetenzen daher durch unterschiedliche Personen oder Rollen eingebracht.

4.1 Kompetenzen in DASC-PM

Die in Abbildung 4 dargestellten Kompetenzen von Data Scientists bieten eine grobe Orientierung über die zentralen Themenfelder, die im Kontext datengetriebener Projekte zu berücksichtigen sind. In der praktischen Umsetzung von Projekten zeigen sich die tatsächlich benötigten Kenntnisse und Fertigkeiten jedoch häufig differenzierter und variieren je nach Anwendungsfall. Vor diesem Hintergrund ist es entscheidend, nicht nur alle relevanten Kompetenzbereiche möglichst vollständig zu identifizieren, sondern auch präzise zu formulieren, sodass sie als Grundlage für die gezielte Zusammensetzung von Data-Science-Teams dienen können. Insbesondere ermöglicht dies, unterschiedliche Rollen komplementär zu besetzen und die Zusammenarbeit im Team systematisch zu unterstützen (Neuhaus et al., 2024; Wicht et al., 2021).

Darüber hinaus reicht eine reine Auflistung von Kompetenzbereichen oft nicht aus. Um den konkreten oder gewünschten Kompetenzbedarf einer Projektphase (oder einer Aufgabe) mit den tatsächlich verfügbaren Kenntnissen und Fertigkeiten im Team abzugleichen, ist eine Differenzierung anhand von Kompetenzstufen erforderlich. Abbildung 5 veranschaulicht dies in Form eines radialen Diagramms, das neun Kompetenzbereiche mit entsprechenden Ausprägungsstufen kombiniert. Sowohl die einzelnen Kompetenzbereiche als auch die Stufen werden im Folgenden genauer beschrieben.

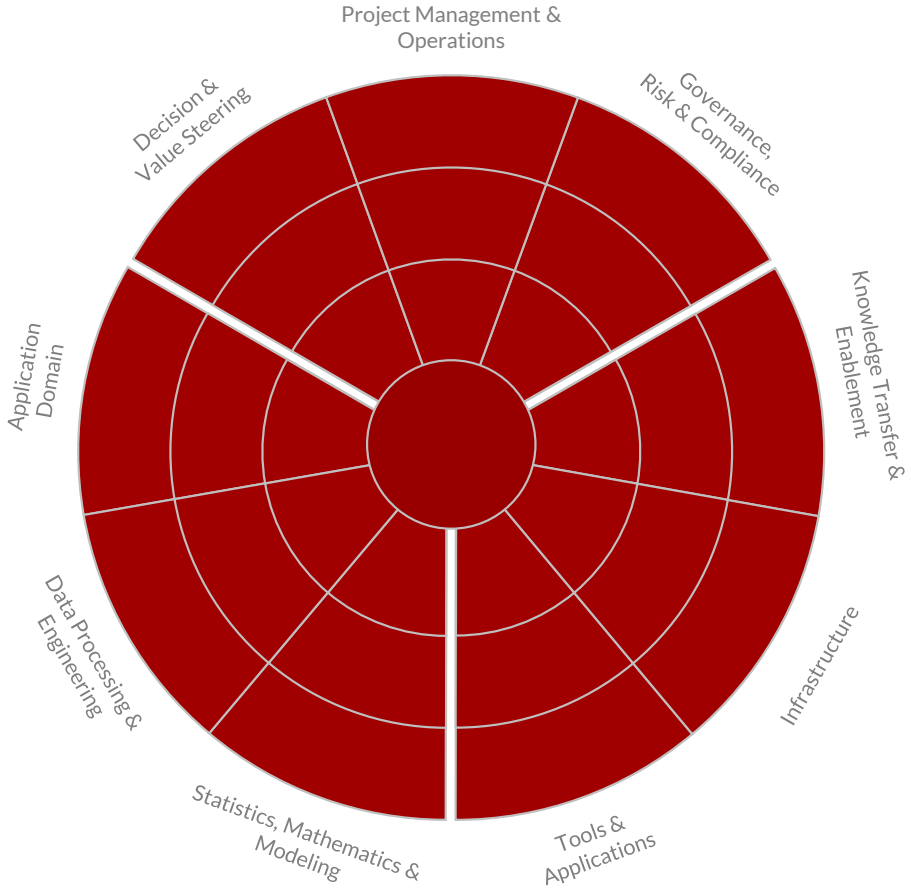


Abbildung 5: Notwendige Kompetenzen in einem Data-Science-Projekt

Project Management & Operations

Dieser Kompetenzbereich beschreibt die Kenntnisse und Fertigkeiten, datengetriebene Projekte strukturiert zu planen, zu koordinieren und operativ umzusetzen. Dazu gehören die Koordination beteiligter Akteur:innen, die Anleitung iterativer und agiler Vorgehensweisen, die Strukturierung von Arbeitsprozessen sowie der Umgang mit Projektartefakten. *Operations* bezeichnet das Vermögen, die Übertragung der im Rahmen des Projekts entwickelten Artefakte in den Live-Betrieb zu steuern.

Decision & Value Steering

Dieser Bereich umfasst die Kenntnisse und Fertigkeiten, datengetriebene Initiativen zu beschreiben, zu bewerten, zu priorisieren und an organisationalen Zielen auszurichten. Dazu gehören die Identifikation und die Auswahl von Use Cases, die Einschätzung ihres Wertbeitrags unter Berücksichtigung gegebener Rahmenbedingungen sowie die Begleitung von organisationalen Veränderungsprozessen.

Application Domain

Dieser Kompetenzbereich zielt auf das Verständnis fachlicher Kontexte und ihrer spezifischen Anforderungen ab. Dies umfasst Kenntnisse und Fertigkeiten wie die Erfassung sachlicher Zusammenhänge und betrieblicher Prozesse sowie die Identifikation zugehöriger Problemstellungen und potenzieller Lösungen.

Data Processing & Engineering

Dieser Kompetenzbereich beschreibt Kenntnisse und Fertigkeiten zur strukturierten Aufbereitung, Integration und Organisation von Daten. Dazu gehören Gestaltung und Implementierung von Datenarchitekturen, Datenmanagement sowie die Sicherstellung angemessener Datenqualität. Der Fokus liegt auf der Transformation von Rohdaten in eine nutzbare Form für Analyse und Modellierung. Die Kompetenzen dieses Bereichs beinhalten explizit nicht die Fertigkeit zur (technischen) Schaffung oder Anpassung von Infrastruktur.

Statistics, Mathematics & Modeling

Dieser Kompetenzbereich umfasst konzeptionelle und methodische Kenntnisse und Fertigkeiten im Kontext datengetriebener Analysen. Er verzahnt insbesondere fundierte mathematisch-statistische Kenntnisse gleichwertig mit der Beherrschung von Analyseverfahren und modernen Modellarchitekturen. Dies beinhaltet sowohl den Entwurf und die Evaluierung komplexer Analysemodelle als auch deren gezielte Anwendung und (Weiter-)Entwicklung.

Tools & Applications

Dieser Kompetenzbereich beschreibt die Kenntnisse und Fertigkeiten, datengetriebene Lösungen und Systeme umzusetzen und zu betreiben. Der Bereich umfasst die Auswahl und Nutzung von Software, Programmiersprachen, Frameworks und technischen Stacks sowie die Integration, Bereitstellung und Überwachung von Modellen in Anwendungen.

Infrastructure

Dieser Kompetenzbereich beschreibt die Kenntnisse und Fertigkeiten, eine geeignete technische Basisumgebung für datengetriebene Systeme zu konzipieren und bereitzustellen. Dazu gehören technische Infrastrukturen wie Cloud-Plattformen, verteilte Systeme oder Hochleistungsrechner sowie Aspekte wie Skalierbarkeit und Verfügbarkeit. Ziel ist es, stabile und leistungsfähige Rahmenbedingungen zu schaffen, in denen Datenverarbeitung und Anwendungen zuverlässig ausgeführt und langfristig betrieben werden können.

Knowledge Transfer & Enablement

Dieser Kompetenzbereich beschreibt die Kenntnisse und Fertigkeiten, Wissen zu vermitteln, adressat:innengerecht aufzubereiten und so die organisationale Nutzung zu ermöglichen. Dies umfasst nicht nur die Kommunikation/Präsentation analytischer Ergebnisse, sondern auch die Förderung von Zusammenarbeit zwischen unterschiedlichen Stakeholdern und die Befähigung von Nutzer:innen im Umgang mit Daten und KI.

Governance, Risk & Compliance

Dieser Kompetenzbereich beschreibt die Kenntnisse und Fertigkeiten, rechtliche, ethische und regulatorische Anforderungen im Umgang mit Daten und KI zu berücksichtigen und einzuhalten. Dies umfasst Themen wie Informationssicherheit, Datenschutz, Regelkonformität, Transparenz und Risikomanagement. Ziel ist es, datengetriebene Systeme verantwortungsvoll zu gestalten, potenzielle Auswirkungen zu erkennen und geltende Vorgaben einzuhalten.

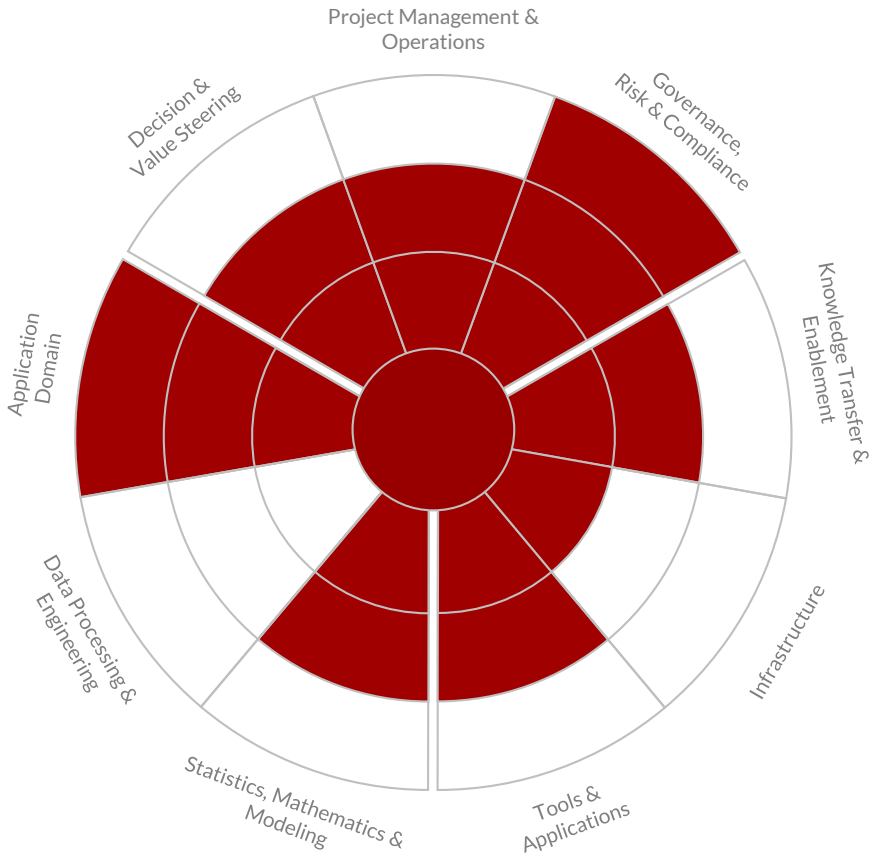


Abbildung 6: Notwendige Kompetenzen in einem Data-Science-Projekt in beispielhafter Ausprägung

Abbildung 6 zeigt die Kompetenzen in beispielhafter Ausprägung, wie sie etwa als Profil einer einzelnen Person zu finden wären. Dabei zeigt eine stärkere Befüllung von innen nach außen eine tiefergehende Kompetenz an, die mit vier Stufen beschrieben werden kann. Stufe 1 ist in diesem Modell bei Mitarbeiter:innen in Data-Science-Projekten prinzipiell immer als vorhanden anzunehmen, Stufe 4 beschreibt einen Expert:innenstatus. Im Detail können die einzelnen Stufen wie folgt charakterisiert werden:

Stufe 1: Basiskompetenzen

Auf dieser Stufe verfügen Personen über grundlegende Kenntnisse zentraler Konzepte (und deren Zusammenhänge). Sie sind fähig, klar strukturierten Anleitungen zu folgen und Routinen zuverlässig anzuwenden. Eigenständige Anpassungen oder Bewertungen erfolgen jedoch kaum.

Stufe 2: Aufbauende Kompetenzen

Personen auf dieser Stufe handeln zunehmend selbstständig und können ihr Vorgehen in begrenztem Umfang anpassen. Sie sind in der Lage, Ergebnisse und Erkenntnisse einzuordnen und diese im jeweiligen Anwendungskontext zu interpretieren.

Stufe 3: Fortgeschrittene Kompetenzen

Auf dieser Stufe sind Personen in der Lage, komplexere Situationen eigenständig zu beurteilen. Sie berücksichtigen dabei relevante Kontextfaktoren, Zielkonflikte sowie Unsicherheiten und treffen fundierte Entscheidungen. Darüber hinaus sind sie in der Lage, Ergebnisse und Erkenntnisse situationsangemessen zu bewerten.

Stufe 4: Kompetenzen auf Expert:innenniveau

Personen auf Expert:innenniveau sind in der Lage, Vorgehensweisen zu hinterfragen, eigenständig neue Lösungen für neue Probleme zu entwickeln sowie Mehrwert und Grenzen dieser Lösungen kritisch zu reflektieren. Sie können andere fundiert beraten und tragen aktiv zur Weiterentwicklung von Best Practices und Standards bei.

4.2 Rollen in DASC-PM

Die erfolgreiche Umsetzung von Data-Science-Projekten erfordert ein enges Zusammenspiel verschiedener Rollen, die gemeinsam technische, fachliche und zugleich rechtliche, strategische und prozessuale Herausforderungen adressieren. In vielen Darstellungen wird mit festen Rollenbezeichnungen gearbeitet, die allerdings je nach Organisation stark variieren und sich im Laufe der Zeit verändern können. DASC-PM beschreibt daher nicht spezifische Rollennamen, sondern drei übergeordnete Rollenbereiche (vgl. Abbildung 7).

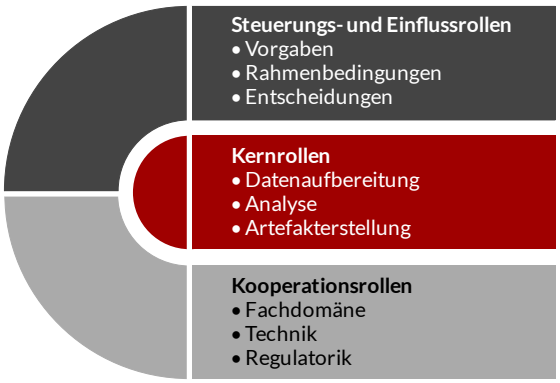


Abbildung 7: Rollenbereiche und Auszüge ihrer Aufgabenbereiche innerhalb eines Data-Science-Projektes

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

Kernrollen

Zu den Kernrollen gehören Funktionen, die die analytische und datenbezogene Umsetzung eines Data-Science-Projekts verantworten und die zentralen Artefakte entwickeln. Der Begriff *Data Scientist* wird in der Praxis unterschiedlich verwendet: teils als Oberbegriff für alle in einem Data-Science-Projekt tätigen Personen, teils in einem engeren Sinne für diejenigen, die sich auf die eigentliche Datenanalyse spezialisieren. In diesem Dokument wird der Begriff in dieser engeren Bedeutung verwendet:

> ***Im spezifischen Sinne versteht man unter Data Scientists Spezialist:innen für die Auswahl der Analysemethoden und -werkzeuge, die Durchführung von Analysen und die Interpretation der Ergebnisse.***

Die Rolle des Data Scientists kann in größeren und komplexeren Data-Science-Projekten in weitere Unterrollen aufgespalten werden. Dazu zählen unter anderem Data Analysts, die sich mit Datenaufbereitung, -analyse und -auswertung befassen und insbesondere in der explorativen Analyse und Prognoseerstellung tätig sind (häufig unter Einsatz traditioneller Analysemethoden). Methodenspezialist:innen hingegen beschäftigen sich mit der Erforschung und Weiterentwicklung von Data-Science-Methoden und sind stärker theorie- und forschungsorientiert. Machine Learning Engineers stellen eine software- und infrastrukturnahe Spezialisierung dar und sind für die technische Umsetzung sowie die produktive, skalierbare Bereitstellung analytischer Artefakte verantwortlich.

Eine ebenso zentrale Funktion im Kernteam übernehmen die *Data Engineers*, die die Beschaffung, Speicherung, Aufbereitung, Strukturierung und Bereitstellung der Daten verantworten. 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.

Kooperationsrollen

Zu den Kooperationsrollen zählen Funktionen, die an den Schnittstellen zwischen Fachdomäne, Technik und regulatorischen Anforderungen arbeiten und so die Integration und Anschlussfähigkeit der Lösungen sicherstellen. Eine zentrale Funktion innerhalb dieser Gruppe übernehmen die *Domänenexpert:innen*. Dies sind häufig die fachlichen Anwender:innen oder deren Vertreter:innen. Sie verfügen über spezifisches Wissen in Bezug auf die Anwendungsdomäne und besitzen ein inhaltliches Verständnis der Problemstellung bzw. des Anwendungsfalls. Ihr tiefes Domänenwissen ermöglicht es ihnen, die fachlich relevanten Schwerpunkte für Analyse und Modellierung festzulegen.

Innerhalb der Domänenexpert:innen kann es wieder Unterrollen geben. Im Business-Kontext häufig anzutreffen sind etwa die *Business Developer*, die den domänenspezifischen Use Case eines Projekts 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.

Ergänzend zu diesen fachlichen Schnittstellen treten unterschiedliche organisatorische und technische Funktionen hinzu. *Governance-, Risk- und Compliance-Expert:innen* unterstützen das Projekt, indem sie regulatorische Anforderungen interpretieren, deren Umsetzung begleiten und für eine verantwortungsvolle Anwendung sensibilisieren.

Plattform- und Infrastrukturfunktionen umfassen alle Aufgaben, die notwendig sind, um die technischen Voraussetzungen für ein Data-Science-Projekt zu schaffen. Typische Rollen in diesem Bereich sind etwa *IT Infrastructure Architect*, verantwortlich für den Entwurf einer geeigneten IT-Infrastruktur für das Projekt, und *IT-Techniker:innen/IT-Administrator:innen*, die die benötigte Hard- und Software bereitstellen und die zugrunde liegenden Systeme konfigurieren. Aber auch *Anwendungsentwickler:innen*, die sich mit der Implementierung von Anwendungssoftware/-werkzeugen zur produktiven Nutzung der Analyseergebnisse befassen, werden diesem Aufgabengebiet zugeordnet.

Solution Engineers und *Solution Architects* sind dafür verantwortlich, dass die entwickelten Lösungen in bestehende Systeme integrierbar sind und in eine tragfähige Gesamtarchitektur eingebettet werden. *Plattform Engineers* stellen Cloud- oder Datenplattformen bereit, auf denen Datenverarbeitung, Training und Modellbetrieb stattfinden können. Die Rolle des *Product Owners* stellt sicher, dass Anforderungen priorisiert und Projektergebnisse mit dem erwarteten Nutzen in Einklang gebracht werden. *Data Owner* treffen Entscheidungen über Datenqualität, Zugriffsrechte und die fachliche Auslegung zentraler Datenobjekte. *UI-/UX-Designer:innen* tragen dazu bei, dass Analyseartefakte in benutzerfreundliche, verständliche und im Arbeitsalltag gut nutzbare Systeme eingebettet werden.

Während andere Rollen primär Analyseergebnisse implementieren oder integrieren, stellt der *MLOps Engineer* sicher, dass diese langfristig zuverlässig, effizient und kontrollierbar betrieben werden können. Dazu gehören der Aufbau stabiler Betriebsumgebungen, die kontinuierliche Überwachung von Nützlichkeit, Kosten und Performance sowie die Automatisierung zentraler Abläufe wie Deployment, Retraining und Monitoring.

Steuerungs- und Einflussrollen

Zu den Rollen mit Steuerungs- und Einflussmöglichkeiten gehören Funktionen, die durch Entscheidungen, Vorgaben oder die Ausgestaltung rechtlicher, sicherheitsbezogener und organisatorischer Rahmenbedingungen den Verlauf eines Data-Science-Projekts maßgeblich prägen. *Projektmanager:innen* 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 in kleineren Projekten wird das Projektmanagement häufig von Personen übernommen, die zugleich als Data Scientists oder Data Engineers tätig sind. Das Projektmanagement kann aber auch von Personen ohne spezifisches Data-Science-Know-how übernommen werden, wenn ihnen entsprechende Expert:innen zur Seite stehen. Solche – auch als *Methodical Lead* oder *Technical Lead* bezeichneten – Expert:innen verfügen über vertieftes methodisches und technisches Know-how und unterstützen die fachlich-technische Ausrichtung des Projekts. Zusammen mit Domänenexpert:innen bestimmen sie den Scope der Analyse und Umsetzung.

Auf der strategischen Ebene übernehmen auf Data Science und KI spezialisierte *Manager:innen* Aufgaben der langfristigen Ausrichtung, identifizieren Potenziale, bewerten Risiken und entwickeln eine Data-Science-Strategie, die das Projektumfeld und die organisatorische Zielsetzung verbindet. *Sponsor:innen* und *Entscheider:innen* beeinflussen das Projekt durch Budgetverantwortung, Freigaben und die Festlegung zentraler Prioritäten. *Security-*, *Privacy- und Legal-Spezialist:innen* gestalten die rechtlichen, sicherheitsbezogenen und ethischen Rahmenbedingungen, innerhalb derer Data-Science-Projekte umgesetzt werden dürfen. Sie gewährleisten Datenschutz, Informationssicherheit sowie die Einhaltung gesetzlicher Vorgaben und wirken damit nicht nur als Kontrollinstanz, sondern auch als gestaltende Kraft im Projekt.

Teil B
DASC-PM im Detail

Hinweise zu Teil B

In den folgenden Kapiteln werden die einzelnen Phasen des DASC-PM detailliert betrachtet. Zur Strukturierung der einzelnen Phasen wird eine einheitliche Nomenklatur verwendet, die sich auf vier zentrale Begriffe konzentriert: *Kernaufgaben*, *Begleitende Aufgaben*, *Merkmalstragende Bereiche* und *Schnittstellenaufgaben*.

Kernaufgaben umfassen diejenigen Tätigkeiten, die unmittelbar zur Erreichung der Ziele der Phase beitragen. Ihre Durchführung ist für den Projekterfolg wesentlich. Sie bilden damit den fachlichen Schwerpunkt jeder Phase.

Begleitende Aufgaben unterstützen die Durchführung der Kernaufgaben. Sie beschreiben Tätigkeiten wie Qualitätssicherung, Risikomanagement oder vorbereitende Aktivitäten, ohne die die Kernaufgaben ihr Ziel nicht vollständig oder in gewünschter Qualität erreichen können.

Merkmalstragende Bereiche beschreiben die Bezugspunkte für die Aufgaben der jeweiligen Phase. Sie sind häufig als Objekte, Artefakte oder Ergebnisse beschrieben, können allerdings auch abstraktere Elemente wie beispielsweise Ereignisse beschreiben. Die Aufgaben der Phase greifen auf sie zurück, werden von ihnen beeinflusst oder produzieren sie als Ergebnis.

Kernaufgaben, Begleitende Aufgaben und Merkmalstragende Bereiche werden unter dem Begriff *Bereiche* zusammengefasst. Sie werden in den Kapiteldarstellungen dieses Dokumentteils vorgestellt und vertiefend erläutert. Die Ausführungen je Bereich sind dabei unterschiedlich umfangreich, so wie auch die Bereiche in den meisten Projekten nicht genau die gleiche Form von Detaillierung benötigen.

Schnittstellenaufgaben beziehen sich auf die Verknüpfung der zuvor genannten Elemente. Sie entstehen meist implizit und beschreiben beispielsweise Übergaben von Ergebnissen, Abstimmungen zwischen durchführenden Personen oder die Integration unterschiedlicher Elemente.

Jedes Kapitel beginnt mit einer einfachen sowie einer detaillierten Darstellung der Bereiche und Schnittstellenaufgaben der beschriebenen Phase. Die einfache Darstellung dient der Übersichtsgewinnung, die detaillierte Darstellung zeigt die einzelnen Schnittstellenaufgaben und möglichen Abläufe. Abbildung 8 zeigt die verwendete Nomenklatur in einer Beispielnotation für die Detaildarstellung, die in den einzelnen Kapiteln verwendet wird, um die Zusammenhänge darzustellen.

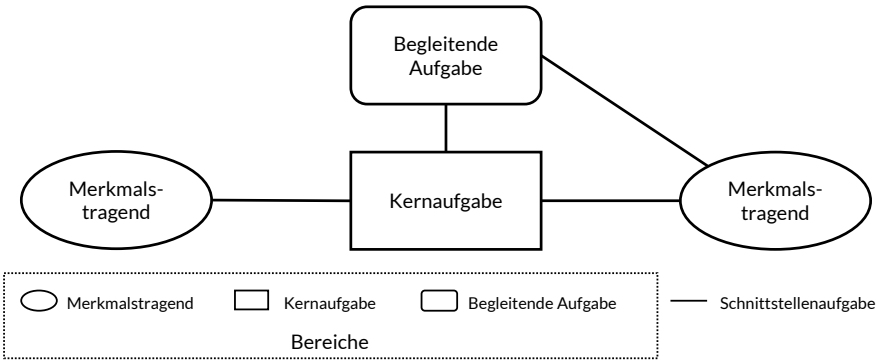


Abbildung 8: Verwendete Nomenklatur und Notation in den Phasen

Die einführende Übersicht je Phase wird ergänzt durch die Darstellung der notwendigen Kompetenzen. Diese Darstellung basiert auf den Ausführungen in Abschnitt 4.1 und insbesondere auf Abbildung 6. Die Aufbereitung ist dabei eine Aggregation der Einschätzungen der Arbeitsgruppe und soll 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. Insbesondere bei der Betrachtung von Kompetenzen im Team ist es auch denkbar, dass nicht vollständige, sondern teilweise erreichte Stufen der Teammitglieder markiert werden.

Teil B schließt mit einer Betrachtung der übergreifenden Aspekte Domain, Methodology und IT Infrastructure.

5 Problem Statement

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 *Problem Statement* wieder. Die typischen Kompetenzen fokussieren vor allem auf fachliche Expertise und projektübergreifende Steuerungsfähigkeiten.

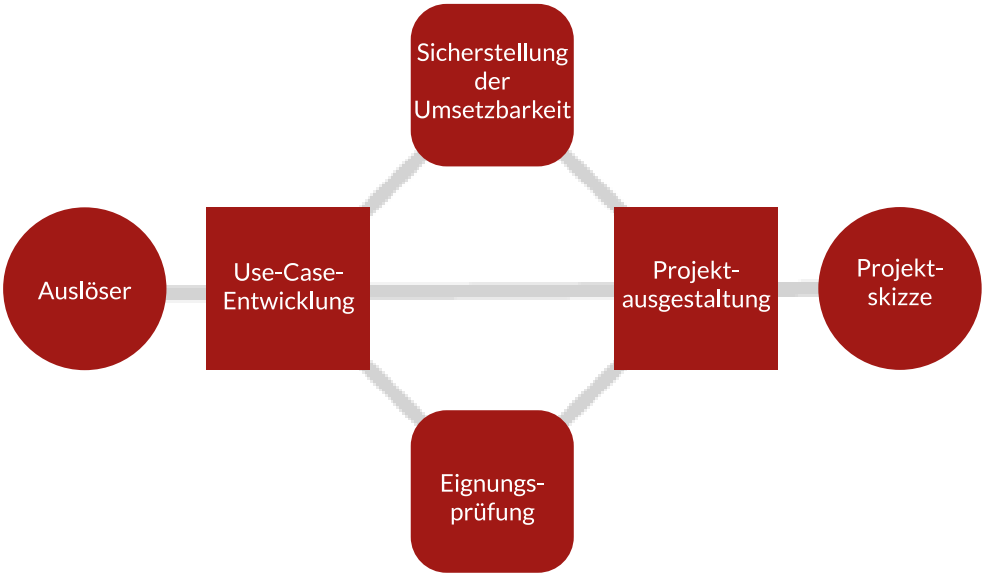


Abbildung 9: Kurzübersicht der Phase „Problem Statement“

Abbildung 10: Kompetenzprofil der Phase „Problem Statement“

Detaildarstellung der Phase Problem Statement

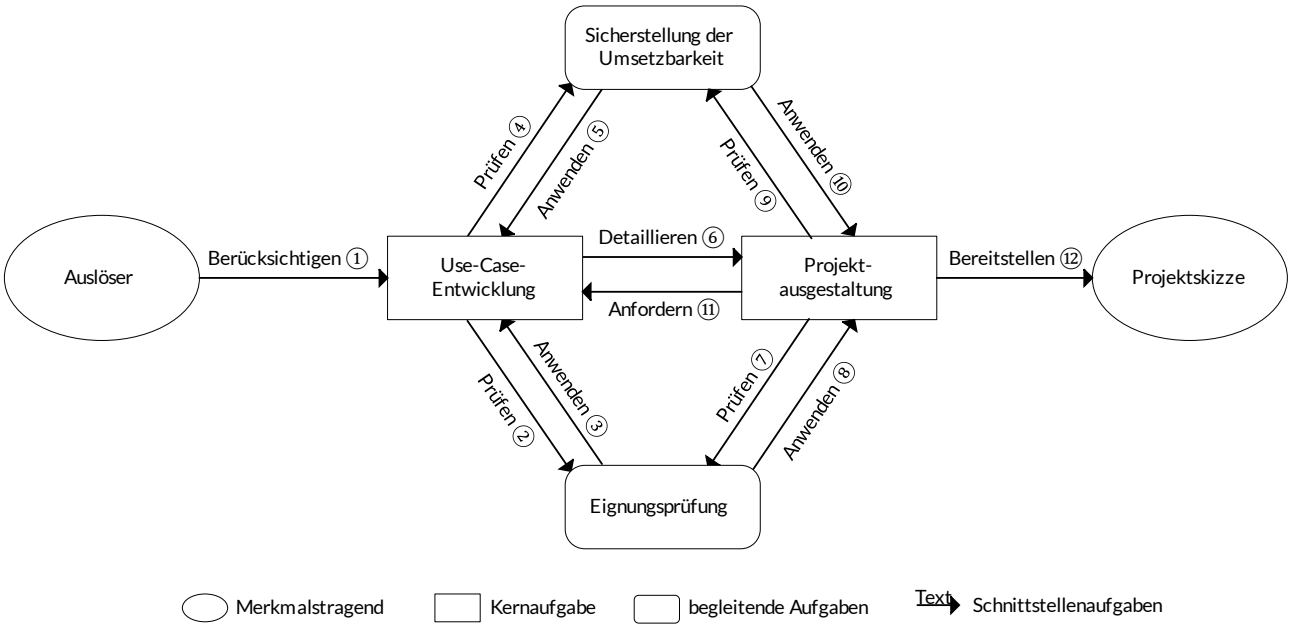


Abbildung 11: Detaildarstellung der Phase „Problem Statement“

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.

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.

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.

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.

5.2 Kernaufgabe „Use-Case-Entwicklung“

Eine häufig dargestellte Herausforderung bei der Beschäftigung mit der Disziplin der Data Science in Organisationen ist die Identifikation und Auswahl praktikabler Use Cases. Im Rahmen von DASC-PM gilt:

> *Ein **Use Case** ist ein klar definiertes, in sich abgeschlossenes (Teil-)Problem einer betrachteten Domäne mit einem ausformulierten Ziel und eindeutigem fachlichen Nutzen, zu dessen Lösung gegebenenfalls mehrere Aufgaben bearbeitet werden müssen.*
> *Mit einem **Projekt** können ein oder mehrere Use Cases behandelt werden.*

Auf die Frage nach den Personengruppen oder Abteilungen in einer Organisation, welche die Entwicklung von Problemen hin zu Use Cases vorantreiben sollten, gibt es keine eindeutige Antwort. Domänenexpert:innen und Data Scientists können und sollten ebenso wie andere relevante Gruppen beteiligt werden.

Teilaufgabe	Beschreibung
Aufbau eines Verständnisses für die Data-Science-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 Datenspezialist:innen dagegen erkennen ggf. nicht die relevantesten Use Cases einer Organisation. Die Kommunikation und Abstimmung zwischen diesen beiden Personengruppen sind daher von hoher Relevanz.

Problem Statement

Bei der Auswahl von geeigneten Use Cases ist es wichtig, eine vertrauensvolle Basis zwischen den beteiligten Personengruppen zu schaffen. Domänenexpert:innen 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 Ansprechpartner:innen aufzubauen und Interviews bzw. Workshops vorzubereiten. Die gewählten Methoden müssen dabei an das gegebene Umfeld der Organisation angepasst werden.

Häufig genannte Formate zur Identifikation geeigneter Use Cases sind Interviews und Workshops.

Interviews eignen sich dabei eher für kleinere Bereiche mit wenigen Beteiligten. Die Durchführung sollte dabei von erfahrenen Interviewer:innen auf Augenhöhe erfolgen.

Workshops eignen sich auch für größere Gruppen und Bereiche. Workshops mit vielen Teilnehmer:innen werden insbesondere zur Ideensammlung eingesetzt. Die Konkretisierung und Priorisierung von Use Cases kann anschließend in kleineren Workshops durchgeführt werden. Um alle Facetten beleuchten zu können, sollte das Feld der Teilnehmer:innen möglichst heterogen sein, also verschiedene Bereiche (bspw. Domänenexpert:innen, Data Scientists, Entscheider:innen) 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 Fundierung 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:innen 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 Teilnehmer:innen fördern.

Zur Identifikation und Auswahl geeigneter Data Science Use Cases haben sich allgemeine Methoden wie Fokusgruppen, Fish Bowls, Design Thinking oder Hackathons bewährt. 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).

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 erfahrene:n, neutrale:n Moderator:in (tritt aber selbst nicht als Ideengeber:in auf) ▪ Gute Strukturierung zur Erreichung der Ergebnisse notwendig (z. B. durch Orientierung an Leitfragen), da sonst Gefahr der Ziellosigkeit besteht ▪ Teilnehmer:innenzahl 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:innen 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 Transkript, 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:innen je nach Thema, Perspektive und Kompetenzen möglich ▪ Geringerer Druck auf die Teilnehmer:innen, 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).
Methode: Design Thinking	
▪ Fokus auf die Perspektive der Endanwender:innen 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)

Ergänzend können folgende Best Practices zur Auswahl geeigneter Data Science Use Cases genutzt werden:

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:innen 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:innen 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 IT-Infrastruktur usw.) zu belegen.

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	▪ 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 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, Benutzungsfreundlichkeit) ▪ 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 grundsätzlich erfüllt werden können. Die eigentliche Prüfung, ob die notwendigen technischen, fachlichen und organisatorischen Voraussetzungen vorliegen, findet im Rahmen der *Sicherstellung der Umsetzbarkeit* (vgl. Abschnitt 5.4) statt.

Wird der betrachtete Use Case für die Umsetzung in einem Projekt ausgewählt, erfolgt eine detailliertere Prüfung in der *Projektausgestaltung* (vgl. Abschnitt 5.5). Hierbei können auch Priorisierungen der Use Cases vorgenommen werden.

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.

5.4 Begleitende Aufgabe „Sicherstellung der Umsetzbarkeit“

In diesem Schritt ist zu prüfen, welche Projektideen sich konkret 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.

Während die *Eignungsprüfung* (Abschnitt 5.3) bewertet, ob ein Use Case grundsätzlich für Data Science geeignet ist, fokussiert die *Sicherstellung der Umsetzbarkeit* darauf, unter welchen konkreten technischen, organisatorischen und fachlichen Bedingungen dieser tatsächlich realisiert werden kann.

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:innen 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.

Wird der betrachtete Use Case für die Umsetzung in einem Projekt ausgewählt, erfolgt eine detailliertere Prüfung in der *Projektausgestaltung* (vgl. Abschnitt 5.5).

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.

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 die entsprechende 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.

5.6 Merkmalstragender Bereich „Projektskizze“

Im Gegensatz zu anderen Disziplinen sind 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 angestrebte 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 Data Provision

Die Phase *Data Provision* 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.

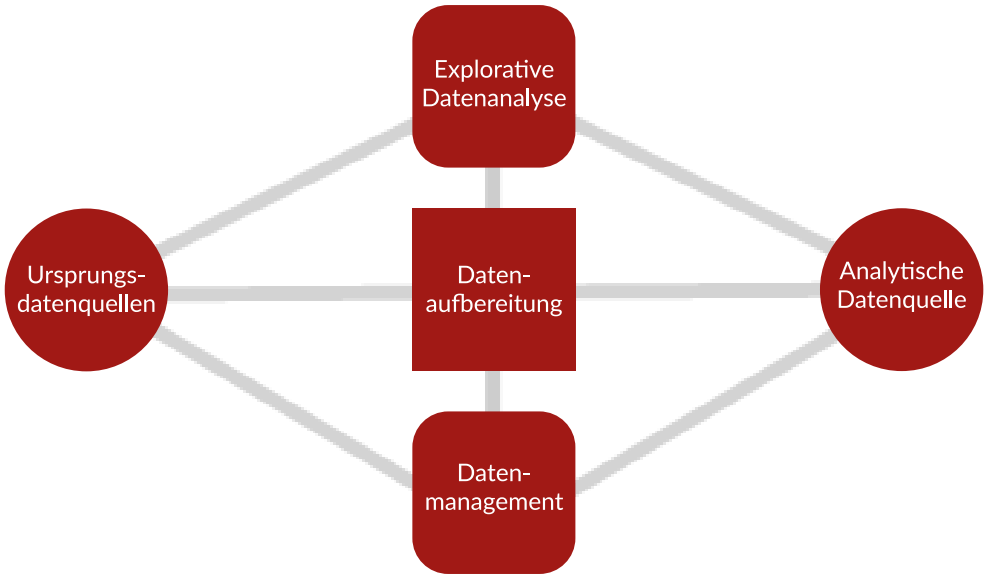


Abbildung 12: Kurzübersicht der Phase „Data Provision“

Abbildung 13: Kompetenzprofil der Phase „Data Provision“

Detaildarstellung der Phase Data Provision

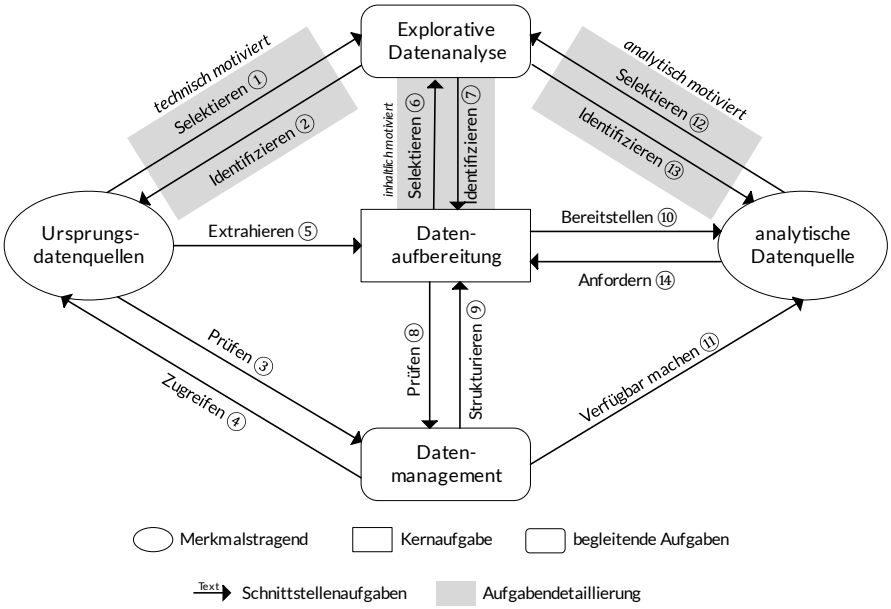


Abbildung 14: Detaildarstellung der Phase „Data Provision“

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

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 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.

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.

Eine ausführlichere Betrachtung der möglichen Merkmale von Datenquellen ist z. B. bei Jayawardene et al. (2013) zu finden. Die Aufzählung der Datenqualitätskriterien erhebt dabei, trotz ihres Umfangs, keinen Anspruch auf Vollständigkeit. Die Relevanz der einzelnen Merkmale ist projektindividuell zu bewerten.

Data quality criteria (Jayawardene et al., 2013)			
Ability to Represent Null Values	Access Security	Accessibility	Accessibility and Clarity
Accessibility Timeliness	Accessible	Accuracy to Reality	Accuracy to Surrogate Source
Accuracy	Accuracy / Validity	Allowing Access to Relevant Metadata	Applicability
Appropriate Amount of Data	Appropriateness	Authority	Availability
Believability	Business Rule Validity	Clarity	Coherence
Cohesiveness	Comparability	Complete	Completeness
Complexity	Comprehensiveness	Concise Representation	Conciseness
Concurrency of Redundant or Distributed Data	Conformance	Conforming to Metadata	Conformity
Consistency	Consistency and Synchronization	Convenience	Correct Interpretation
Correctness	Credibility	Currency	Currency / Timeliness
Data Coverage	Data Decay	Data Integrity Fundamentals	Data Specifications
Definition Conformance	Derivation Validity	Document Standardization	Duplication/ Non-duplication
Ease of Understanding	Ease of Use and Maintainability	Enterprise Agreement of Usage	Equivalence of Redundant or Distributed Data
Fact Completeness	Flexibility	Flexibly Presented	Format Precision
Informativeness/Redun-dancy	Integrity	Interactivity	Interpretability
Maintainability	Mapped Completely	Mapped Consistently	Mapped Meaningfully
Mapped Unambiguously	Naturalness	Null Values	Objectivity
Perception Relevance and Trust	Perceptions	Phenomena Mapped Correctly	Portability
Precision	Precision/Completeness	Presentation Clarity	Presentation Media Appropriateness
Presentation Objectivity	Presentation Quality	Presentation Standardization	Presentation Utility
Properties Mapped Correctly	Provenance	Record Existence	Referential Integrity
Relevance / Aboutness	Relevance / Relevancy	Reliability	Representation Con-sistency
Representation of Null Values	Reputation	Secure	Security
Semantic Consistency	Semantic Definition	Signage Accuracy and Clarity	Source Quality and Security Warranties or Certifications
Speed	Structural Consistency	Structured Valued Standardization	Suitably Presented
Timeliness	Timeliness and Availability	Timeliness and Punctuality	Timely
Traceability	Transactability	Type-sufficient	Ubiquity
Unambiguity	Understandable	Understood	Uniqueness / Unique
Usability	Valid	Validity	Value Added
Value Completeness	Value Existence	Value Validity	Verifiability
Volatility			

6.2 Kernaufgabe „Datenaufbereitung“

Ziel der *Datenaufbereitung* ist es, die verfügbaren Rohdaten in eine konsistente, qualitativ geeignete und strukturierte Form zu überführen, sodass sie effizient und verlässlich für nachgelagerte Analyseverfahren genutzt werden können. Ein wesentliches Teilziel der Aufgabe ist daher die Erhöhung der Datenqualität.

Die Verarbeitung großer Datenmengen erfordert den Einsatz leistungsfähiger Hard- und Software sowie teilweise innovativer Verfahren.

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.

Teilaufgabe	Beschreibung
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.
Datenanonymisierung	Werden innerhalb von Data-Science-Projekten vertrauliche Daten (z. B. personenbezogene Daten) benötigt, müssen diese ggf. zunächst anonymisiert oder pseudonymisiert werden.
Datenbereinigung	Identifizierte Fehler oder auch fehlende Werte können ggf. manuell oder auch automatisiert bereinigt werden. Wenn dies nicht möglich ist, ist eine Datenfilterung oder Dimensionsreduzierung zu prüfen.
Datenfilterung	Nicht benötigte oder auch fehlerhafte Daten sollten aus der Datenbasis entfernt werden.
Datenintegration	Daten aus verschiedenen Quellen müssen zusammengeführt und vereinheitlicht werden.
Datenstrukturierung	Abhängig von den anzuwendenden Analyseverfahren müssen unstrukturierte Daten zuvor strukturiert werden. Dafür können bspw. Methoden des Natural Language Processing oder der Bilderkennung genutzt werden.
Datentransformation	Transformationen sind durchzuführen, um Daten für die Analyse vorzubereiten. Dies beinhaltet sowohl den bei der explorativen Datenanalyse identifizierten Transformationsbedarf als auch durch das Datenmanagement getriebene Transformationen aus eher technischer Sicht.
Dimensionsreduzierung	Irrelevante oder redundante Merkmale sollten aus der Datenbasis entfernt werden.
Erstellung von Datenaufbereitungsplänen	Vor der Datenaufbereitung sind basierend auf dem Datenbedarf Aufbereitungspläne zu erstellen.
Formatanpassung	Quellformate sind i. d. R. nicht primär für die Anwendung von Analyseverfahren definiert worden. Deshalb ist hier häufig eine Überführung in ein geeignetes Format nötig.
Merkmalserzeugung	Aus den bestehenden Daten können zusätzliche bzw. alternative Merkmale abgeleitet werden.
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 aufgrund 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.

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.

Im Unterschied zu den weiteren Bereichen der Datenbereitstellung, die sich auf die Identifikation, Bewertung und Aufbereitung konkreter Datenbestände konzentrieren, umfasst das Datenmanagement die übergreifende Organisation, Steuerung und Sicherstellung der nachhaltigen Nutzung von Daten über den gesamten Projektverlauf hinweg.

Als Artefakt entsteht im Datenkatalog eine Erweiterung zur Nachvollziehbarkeit des Datenmanagements.

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, insbesondere ihrer Sensibilität und Personenbezogenheit sind datenschutzrechtliche Anforderungen zu prüfen und einzuhalten. Dazu gehören beispielsweise Pflichten zur Anonymisierung und Pseudonymisierung. Rollen- und Rechtekonzepte sind datenschutzkonform auszugestalten.
Datensicherheit	Insbesondere der Aspekt der Vertraulichkeit ist durch geeignete technische und organisatorische Maßnahmen zu berücksichtigen. Je nach Schutzbedarf der Daten sind daher z. B. Zugriffskontrollen, Verschlüsselungen, Protokollierung und Wiederherstellungsverfahren in geeigneter Form einzusetzen.
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.

6.4 Begleitende Aufgabe „Explorative Datenanalyse“

Ziel der *Explorativen Datenanalyse* ist es, ein strukturbezogenes und inhaltliches Verständnis der Daten zu gewinnen, Auffälligkeiten zu identifizieren und Hypothesen für die weitere Analyse abzuleiten. Im Unterschied zur Analysephase stehen dabei nicht die Anwendung und Bewertung konkreter Analyseverfahren im Vordergrund, sondern die vorbereitende Durchdringung der Daten als Grundlage für deren gezielte Nutzung.

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.

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-, Faktoren- 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 notwendige 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.

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 es 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.
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 zugrunde liegenden 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.

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 in einem frühen Stadium eines Data-Science-Projekts nicht immer eindeutig möglich, eine Einschätzung durch Domänenexpert:innen 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 Analysis

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 *Analysis* 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.

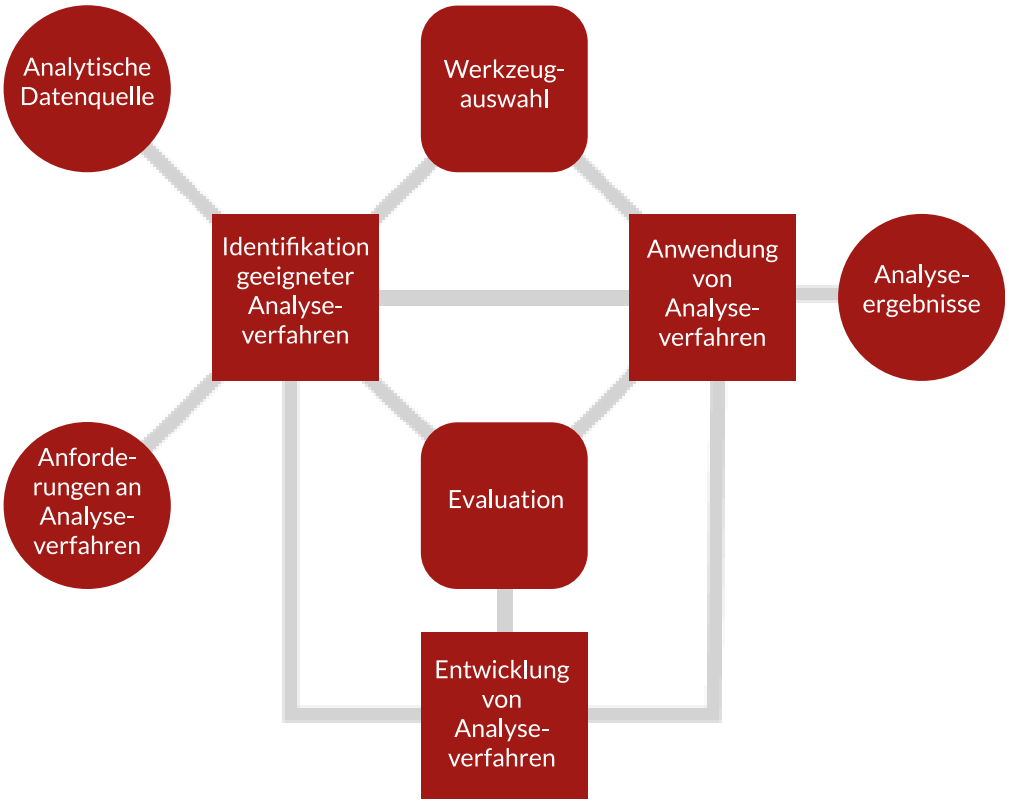


Abbildung 15: Kurzübersicht der Phase „Analysis“

Abbildung 16: Kompetenzprofil der Phase „Analysis“

Detaildarstellung der Phase Analysis

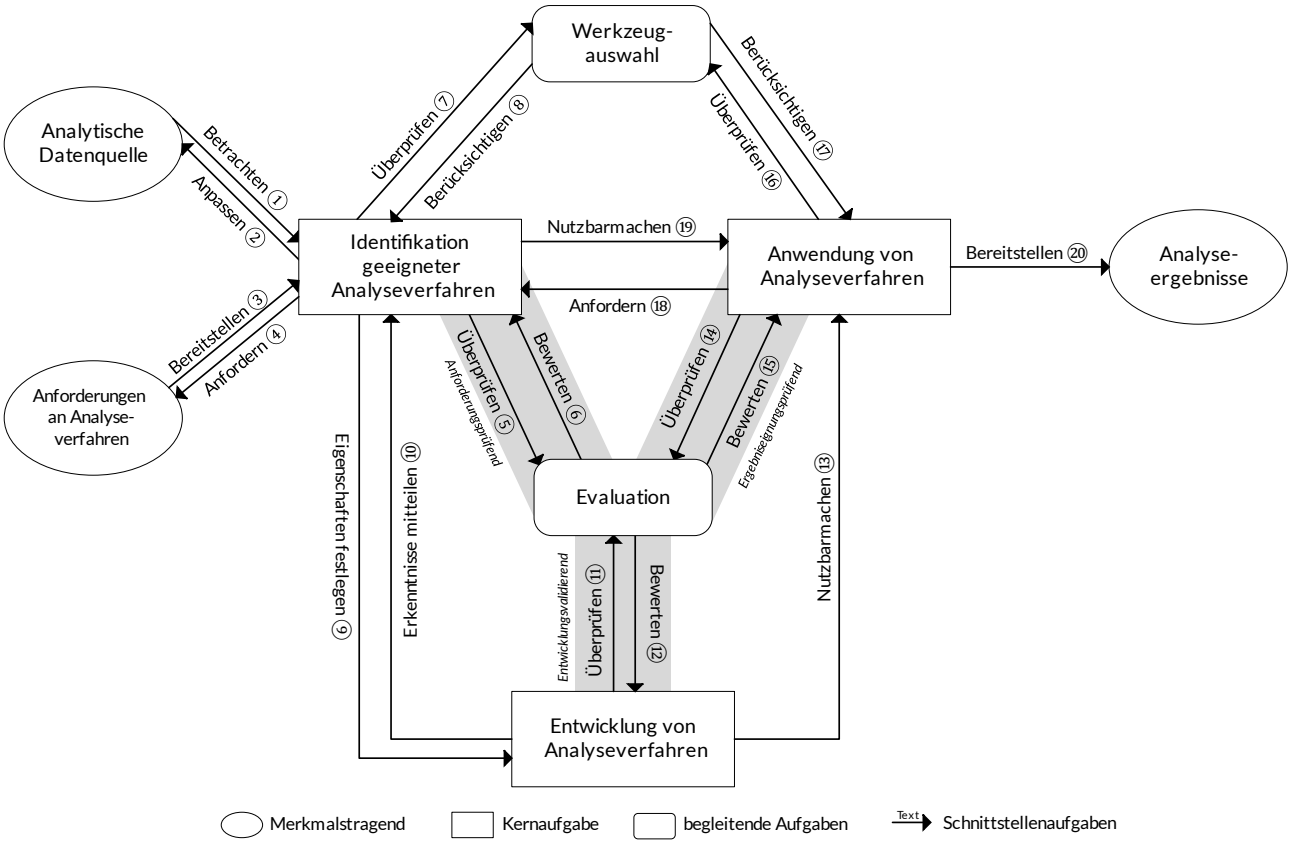


Abbildung 17: Detaildarstellung der Phase „Analysis“

①	Die analytische Datenquelle wird durch Bearbeitung der in der Phase *Datenbereitstellung* 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 das Deployment bereitgestellt werden.

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 *Analysis* 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.

Sie bilden damit eine zentrale Grundlage für die systematische Auswahl, Bewertung und ggf. Weiterentwicklung geeigneter Analyseverfahren im weiteren Projektverlauf.

Merkmal	Beschreibung
Anforderungsabdeckung	Nicht immer können die gewählten Analyseverfahren alle Anwendungsanforderungen vollständig erfüllen. Wünschenswert ist dennoch ein möglichst hoher Abdeckungsgrad.
Effizienz	Das Verfahren muss mit der 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 eindeutig festgestellt worden 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 geeigneter Analyseverfahren* stellt häufig eine große Herausforderung dar. Obwohl eine sehr große Anzahl von Analyseverfahren existiert, besteht die Möglichkeit, dass keines für die Problemstellung geeignet ist. In diesem Fall ist zu prüfen, ob ausgewählte Projektrahmenbedingungen geändert werden können, ob die Entwicklung eines neuen Analyseverfahrens denkbar ist oder ob das Projekt nötigenfalls abgebrochen werden muss.

Bei dieser Aufgabe stehen daher 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.

Als Artefakt dieser Aufgabe entsteht eine Liste von Analyseverfahren, in der auch Begründungen enthalten sind, aus welchem Grund das jeweilige Verfahren für die Fragestellung geeignet ist. 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 Aufgabe 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.

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.

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 die anzuwendenden Verfahren die jeweiligen Aufgaben in geeigneter Form erfüllen. Dies spielt bereits bei der Identifikation (siehe vorheriger Abschnitt) eine zentrale Rolle, kann jedoch erst in der tatsächlichen Anwendung auf die zu analysierenden Daten abschließend überprüft werden. Ziel ist es, das bestmögliche Analyseergebnis zu erzielen. Im Detail hängt dies vom gewählten Verfahren sowie von den spezifischen Anforderungen der jeweiligen Domäne ab. So ist bei einigen Verfahren abzuwägen, ob ein möglichst genaues Ergebnis angestrebt wird oder ein Modell, das auf möglichst viele Szenarien anwendbar ist.

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 Domänenexpert:innen gut verständlich aufbereitet werden, inkl. Hinweise, 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.

Teilaufgabe	Beschreibung
Aufsetzen einer Entwicklungsumgebung	Besonders wenn mehrere Anwender:innen 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 Hard- als auch auf Software. Somit überschneidet sich dieser Bereich auch teilweise mit dem übergreifenden Aspekt der *IT Infrastructure* (vgl. Abschnitt 10.3), der jedoch normalerweise nicht zur Kernaufgabe von Data Scientists gehört und sehr viel weitläufiger gefasst ist.

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.

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

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 Integration 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, sofern dies im Rahmen des Projekts realisierbar ist, bestehende Verfahren angepasst bzw. zusammengeführt werden.

Alternativ können vollständig neue Lösungen entwickelt werden. Dabei ist festzulegen, ob das Verfahren möglichst vielseitig anwendbar sein 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- und Budgetbeschränkungen sind zu berücksichtigen.

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 Benutzungshandbuch
- Anwendungsbeispiele
- Ein Lessons-learned-Dokument
- Schwächen und Stärken des Verfahrens
- Potenzielle Weiterentwicklungsmöglichkeiten

7.7 Begleitende Aufgabe „Evaluation“

Die *Evaluation* ist eine vielfältige Aufgabe, da sie an drei Stellen ausgeführt wird.

- Bei der Auswahl potenziell für die Aufgabenstellung geeigneter Analyseverfahren,
- bei der Entwicklung neuer Analyseverfahren und
- 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.

Im Rahmen der Verfahrensauswahl müssen insbesondere die Gegenüberstellung von Vor- und Nachteilen der betrachteten Analyseverfahren sowie eine Beschreibung geeigneter Anwendungsfälle dokumentiert werden.

Bei der Ergebnisevaluation sind 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) relevant.

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.

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.
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).

7.8 Merkmalstragender Bereich „Analyseergebnisse“

Die Ergebnisse des Analyseprozesses können – abhängig von Fragestellung, Zielsetzung, eingesetzten Methoden und verfügbarer Datenbasis – sehr unterschiedliche Formen annehmen. Die Bandbreite reicht dabei von deskriptiven und diagnostischen Auswertungen über prognostische und präskriptive Modelle bis hin zu (teil-)automatisierten und adaptiven Systemen.

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 Deployment

In der Phase *Deployment* 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 18: Kurzübersicht der Phase „Deployment“

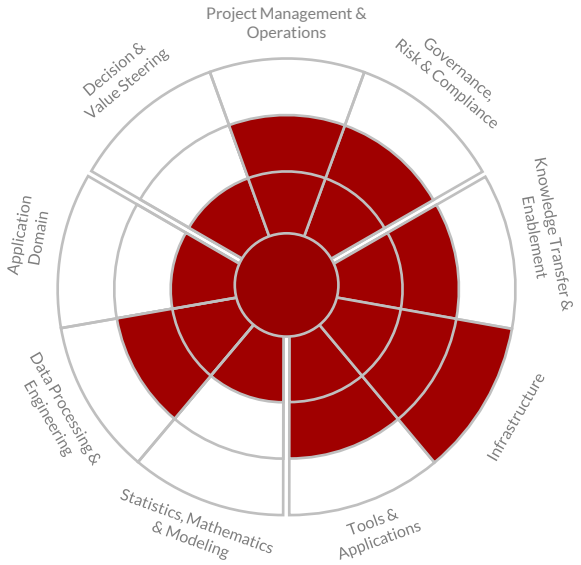


Abbildung 19: Kompetenzprofil der Phase „Deployment“

Detaildarstellung der Phase Deployment

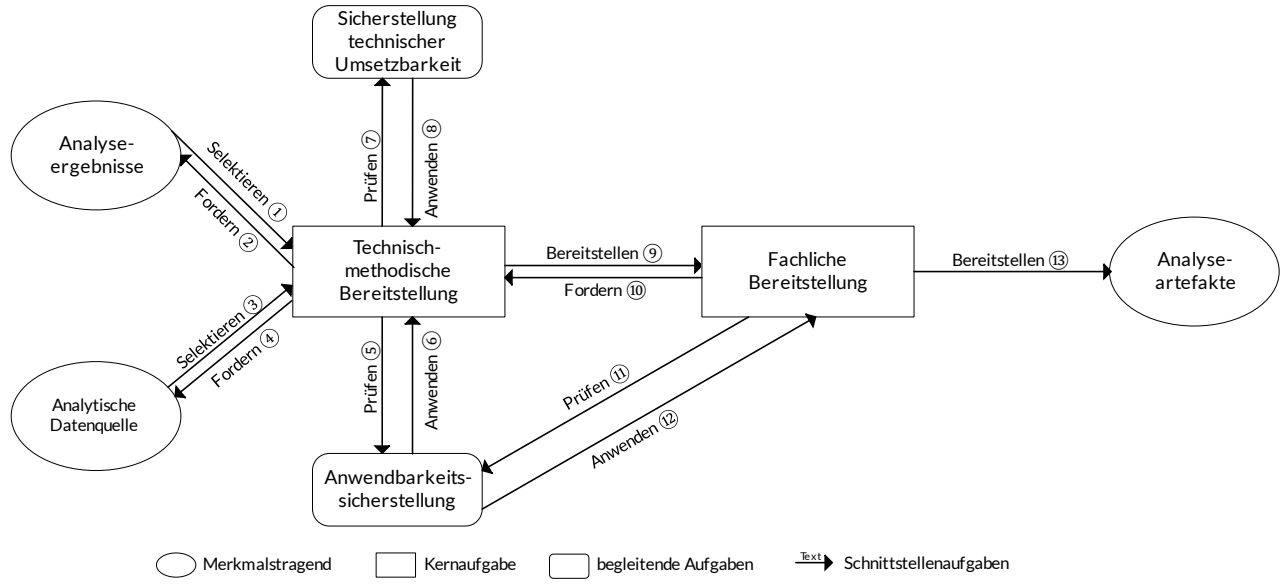


Abbildung 20: Detaildarstellung der Phase „Deployment“

①	Die Analyseergebnisse werden für die technisch-methodische Bereitstellung ausgewählt.
②	Sollten die Analyseergebnisse für die angedachte technisch-methodische Bereitstellung nicht geeignet sein, können Änderungen gefordert werden.
③	Wenn für die technisch-methodische Bereitstellung der Zugriff auf Daten benötigt wird, sind diese zu selektieren. Werden neue Daten verwendet, ist sicherzustellen, dass sie die gleichen Eigenschaften besitzen wie die Daten, mit denen das Modell entwickelt wurde.
④	Sollten die selektierten Daten für die angedachte technisch-methodische Bereitstellung nicht geeignet sein, können Änderungen gefordert werden.
⑤	Bei der technisch-methodischen Bereitstellung muss die Anwendbarkeit durch die Zielgruppe der Analyse geprüft werden.
⑥	Identifizierte Möglichkeiten zur Sicherstellung der Anwendbarkeit sind bei der technisch-methodischen Bereitstellung zu berücksichtigen.
⑦	Bei der technisch-methodischen Bereitstellung müssen technische Anforderungen geprüft werden.
⑧	Identifizierte technische Umsetzungsmöglichkeiten sind bei der technisch-methodischen Bereitstellung zu berücksichtigen.
⑨	Die Ergebnisse der technisch-methodischen Bereitstellung bilden die Grundlage zur fachlichen Bereitstellung der aufbereiteten Analyseergebnisse für die Anwendungszielgruppe. 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 wird die Anwendbarkeit durch die Zielgruppe der Analyse geprüft.
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.

8.1 Merkmalstragender Bereich „Analyseergebnisse“

Das Deployment 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 das Deployment der Analyseergebnisse kann es nötig sein, erneut auf die analytische Datenquelle zuzugreifen (vgl. Abschnitt 6.5). Weitere Besonderheiten 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. Je nach Projekt ist auch die Auswahl mehrerer Implementierungsmöglichkeiten denkbar.

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 nutzen 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.

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.

Ergebnis der *technisch-methodischen Bereitstellung* sind somit implementierte und einsatzfähige Lösungen, in denen die Analyseergebnisse in geeigneter Form bereitgestellt werden, sodass sie in bestehenden Systemen genutzt oder in operative Prozesse integriert werden können.

Teilaufgabe	Beschreibung
Adressat:innengerechte Aufbereitung der Ergebnisse	Die Ergebnisse müssen technisch-methodisch aufbereitet werden, damit die Anwender:innen sie interpretieren können.
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.

8.4 Begleitende Aufgabe „Sicherstellung technischer Umsetzbarkeit“

Die begleitende Aufgabe der *Sicherstellung technischer Umsetzbarkeit* fokussiert – über die reine Bereitstellung hinaus – auf die nachhaltige Betriebsfähigkeit der Analyseanwendung. Sie umfasst die Gewährleistung eines stabilen und wirtschaftlichen Betriebs, einschließlich (technischer) Bedienbarkeit, Wartbarkeit sowie der Möglichkeit, technische Anpassungen effizient umzusetzen.

Teilaufgabe	Beschreibung
Automatisierung	Wie weit können die Auswertung der Daten und die Integration der Ergebnisse automatisiert werden? In welchen Zeitintervallen werden die Analysen wiederholt?
Berücksichtigung von Laufzeiten	Wie rechenaufwendig ist der Algorithmus? Skaliert er z. B. gut mit der Datenmenge?
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?
Betrieb und Support sicherstellen	Wer ist für den Produktivbetrieb der Analyseanwendung zuständig? Wer kann bei technischen/methodischen Fragen und Problemen unterstützen?
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 Softwarelizenzen	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?
Umgang mit angebundenen Datenquellen	Wie kann auf Änderungen bei den Datenquellen (Formate, Qualität, Rechte usw.) reagiert werden? Wer ist zuständig? Wie ist der Informationsfluss?
Zugriffskonzept erstellen	Ist es möglich, den Zugriff auf Analyseergebnisse auf berechtigte Anwender:innengruppen einzuschränken? Wurden Vorkehrungen getroffen, um die Sicherheit aller Daten zu gewährleisten?

8.5 Begleitende Aufgabe „Anwendbarkeitssicherstellung“

Ziel der *Anwendbarkeitssicherstellung* ist es, die Analyseergebnisse so aufzubereiten und bereitzustellen, dass sie von der vorgesehenen Zielgruppe verstanden und genutzt werden können. Dabei ist insbesondere zu berücksichtigen, in welcher Form Ergebnisse interpretiert, weiterverarbeitet oder in Entscheidungsprozesse eingebunden werden.

Die Ausgestaltung erfolgt im engen Zusammenspiel von methodischer Expertise und Domänenwissen, um sowohl die fachliche Relevanz als auch die praktische Nutzbarkeit der Ergebnisse sicherzustellen.

Teilaufgabe	Beschreibung
Adressat:innen identifizieren	Um eine Anwendbarkeit sicherzustellen, müssen die Adressat:innen der Analyse bekannt sein.
UI/UX-Design festlegen	Die Oberfläche sollte für alle Benutzer:innengruppen 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:innen 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 Anwender:innendokumentationen 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.

8.6 Kernaufgabe „Fachliche Bereitstellung“

Die *fachliche Bereitstellung* umfasst die zielgerichtete Aufbereitung und Einbettung der Analyseergebnisse in den jeweiligen Anwendungskontext. Ziel ist es, die Ergebnisse so bereitzustellen, dass sie in fachlichen Prozessen genutzt und in bestehende Entscheidungs- oder Arbeitsabläufe integriert werden können.

Die konkreten Ausprägungen hängen dabei stark von der gewählten Bereitstellungsform sowie den Anforderungen der jeweiligen Domäne ab. Entsprechend werden im Folgenden allgemeingültige Aufgaben beschrieben, die eine fachlich sinnvolle Nutzung der Ergebnisse unterstützen.

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:innen festlegen	Es muss fachliche Ansprechpartner:innen 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 sicherzustellen.
Internes Kostenverrechnungsmodell	Für den Betrieb der Analyseartefakte sind Personal- und IT-Kosten zu ermitteln und ggf. auf die Anwender:innen zu verteilen.
Schulung durchführen	Die im Zuge der Anwendbarkeitssicherstellung konzipierten Schulungen sind in geeigneter Form durchzuführen (Präsenzschulungen, Online-Schulungen, Webinare etc.).
Benutzer:innenhandbuch erstellen	Die im Zuge der Anwendbarkeitssicherstellung konzipierte Benutzer:innendokumentation ist zu erstellen.
Problembehandlung festlegen	Es müssen Prüfmechanismen und Verhaltensweisen für den Fall festgelegt werden, dass das Analyseartefakt keine sinnvollen Ergebnisse (mehr) liefert.

8.7 Merkmalstragender Bereich „Analyseartefakte“

Die Merkmale der *Analyseartefakte* sind abhängig von der Form der Ergebnisbereitstellung (vgl. Abschnitt 8.1).

Merkmal	Beschreibung
Benutzungsdokumentation	Den Nutzer:innen des Analysesystems muss ein Benutzungshandbuch zur Verfügung gestellt werden, in dem die vorhandenen Berichte, Dashboards, Datenbanken etc. inklusive ihrer Zugriffsrechte beschrieben sind. Ferner sind fachliche Ansprechpartner:innen 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:innen 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:innen 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 Usage

Die 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.

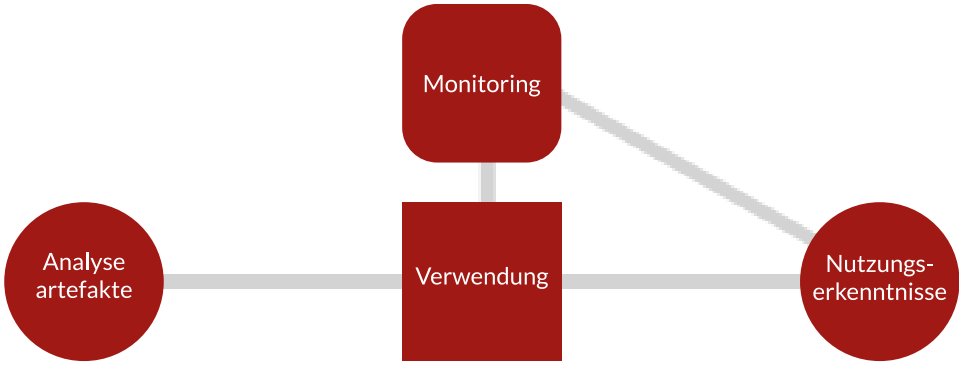


Abbildung 21: Kurzübersicht der Phase „Usage“

Abbildung 22: Kompetenzprofil der Phase „Usage“

Detaildarstellung der Phase Usage

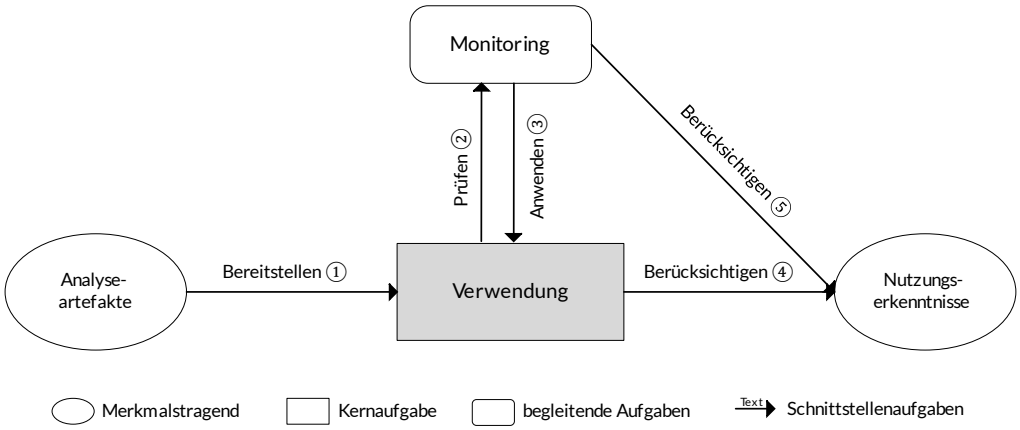


Abbildung 23: Detaildarstellung der Phase „Usage“

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.

9.1 Merkmalstragender Bereich „Analyseartefakte“

Die Nutzung der *Analyseartefakte* fußt auf deren Merkmalen (vgl. Abschnitt 8.7). Weitere Besonderheiten sind hierbei 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.

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

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 Domänenexpert:innen 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.

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 aufgrund veränderter Gegebenheiten nötig werden oder weil sich die erarbeitete Lösung im produktiven Einsatz nicht bewährt hat.

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

10 Übergreifende Aspekte

Neben den Phasen des DASC-PM existieren Aspekte, die in allen Phasen eines Data-Science-Projekts zu berücksichtigen sind und dessen Ausgestaltung maßgeblich beeinflussen. Hierzu zählen insbesondere die Domäne, die Wissenschaftlichkeit sowie die IT-Infrastruktur.

Diese übergreifenden Aspekte wirken nicht isoliert, sondern prägen die einzelnen Phasen in unterschiedlicher Weise – von der Formulierung der Problemstellung über die Verarbeitung und Analyse von Daten bis hin zur Bereitstellung und Nutzung der Ergebnisse. Ihre Berücksichtigung ist damit eine zentrale Voraussetzung für die erfolgreiche Durchführung von Data-Science-Projekten.

Die nachfolgende Übersicht stellt die übergreifenden Aspekte in Bezug zu den einzelnen Phasen dar und bietet eine kompakte Orientierung. In den folgenden Unterkapiteln werden die Aspekte und ihre Bedeutung für die Projektpraxis detaillierter beschrieben.

	Domain	Methodology	IT Infrastructure
Problem Statement	Fachliche Zielsetzung und Problemdefinition prägen die Auswahl und Ausgestaltung von Use Cases	Klare Definition des Untersuchungsgegenstands und nachvollziehbare Zielsetzung	Prüfung der grundsätzlichen technischen Umsetzbarkeit
Data Provision	Bewertung der Datenrelevanz und Datenqualität erfolgt im Domänenkontext	Transparente, replizierbare Datenaufbereitung und fundierte Datenanalyse	Verfügbare Datenzugriffe, Schnittstellen und Rechenressourcen bestimmen die Datenverarbeitung
Analysis	Auswahl, Anwendung und Bewertung der Verfahren richten sich nach domänenspezifischen Anforderungen	Methodisch korrekte Auswahl, Anwendung, Parametrisierung und Evaluation von Verfahren	Infrastruktur beeinflusst Auswahl und Durchführbarkeit der Analyseverfahren
Deployment	Aufbereitung und Bereitstellung orientieren sich an fachlichen Anforderungen der Zielgruppe	Nachvollziehbare Bereitstellung und Dokumentation der Ergebnisse	Technische Integration, Betrieb und Skalierbarkeit der Lösungen
Usage	Nutzung und Bewertung der Ergebnisse erfolgen im Anwendungskontext	Systematische Bewertung der Nutzung und Überprüfung von Modellannahmen	Laufende Überprüfung von Leistungsfähigkeit und Effizienz im Betrieb

10.1 Domain

Die Domäne bildet den fachlichen Kontext, in dem ein Data-Science-Projekt durchgeführt wird. Sie bestimmt maßgeblich, welche Fragestellungen relevant sind, wie Daten interpretiert werden und welche Anforderungen an Analyseverfahren und deren Ergebnisse gestellt werden. Damit beeinflusst die Domäne sämtliche Phasen des Projekts.

Problem Statement

Die Domäne prägt die Identifikation und Ausgestaltung von Use Cases. Fachliche Zielsetzungen, Rahmenbedingungen und Anforderungen ergeben sich aus dem Anwendungskontext und bestimmen, welche Problemstellungen sinnvoll adressiert werden können.

Data Provision

Die Bewertung der Relevanz von Daten ist nur im Domänenkontext möglich. Ein fundiertes Datenverständnis kann ebenfalls nur unter Berücksichtigung der fachlichen Zusammenhänge aufgebaut werden.

Domänenspezifische Anforderungen – etwa regulatorische Vorgaben oder etablierte Datenstandards – sind bei der Datenaufbereitung zu berücksichtigen. Auch Transformationen, wie die Vergleichbarmachung von Messwerten oder die Einordnung von Ausreißern, müssen im Kontext der Domäne erfolgen.

In der explorativen Datenanalyse wird die fachliche Problemstellung konkretisiert und in die Untersuchung der Daten überführt, mit dem Ziel, domänenrelevante Erkenntnisse zu gewinnen.

Analysis

Domänenspezifische Rahmenbedingungen können die Auswahl von Analyseverfahren einschränken oder bestimmte Verfahren ausschließen. Gleichzeitig ergeben sich aus der Domäne Anforderungen an die Ausgestaltung der Verfahren, etwa hinsichtlich Nachvollziehbarkeit oder der Berücksichtigung kausaler Zusammenhänge.

In vielen Domänen existieren etablierte Verfahren, die als Referenz für die Auswahl geeigneter Ansätze dienen können. Darüber hinaus beeinflusst die gewünschte Form der Ergebnisse die Auswahl der Verfahren.

Die Bewertung der Analyseergebnisse erfordert eine Einordnung in den fachlichen Kontext. Relevante Metriken sowie Anforderungen an die Darstellung der Ergebnisse ergeben sich ebenfalls aus der Domäne.

Deployment

Die Aufbereitung und Bereitstellung der Analyseergebnisse müssen sich an den fachlichen Anforderungen der Zielgruppe orientieren. Sowohl die Anwendbarkeit als auch die gewählte Form der Bereitstellung sind im Domänenkontext zu gestalten.

Zudem sind domänenspezifische Rahmenbedingungen, insbesondere nicht-funktionale Anforderungen, bei der technischen Umsetzung zu berücksichtigen.

Usage

Die tatsächliche Nutzung der Analyseergebnisse erfolgt im fachlichen Anwendungskontext. Die Bewertung ihres Nutzens sowie die Ableitung weiterer Anforderungen oder Verbesserungen sind daher eng an die Domäne gebunden.

10.2 Methodology

Die Durchführung von Data-Science-Projekten erfordert ein strukturiertes und methodisch fundiertes Vorgehen, das sich an grundlegenden wissenschaftlichen Prinzipien orientiert. Der Grad der methodischen Strenge kann dabei je nach Projektkontext variieren, insbesondere hinsichtlich der Tiefe der theoretischen Einbettung oder der systematischen Aufarbeitung bestehender Literatur. In praxisnahen Projekten kann diese bewusst reduziert werden, sofern eine angemessene Kosten-Nutzen-Abwägung erfolgt und die damit verbundenen Risiken berücksichtigt werden. Zu diesen Risiken zählt insbesondere, dass durch eine unzureichende Auseinandersetzung mit bestehenden Methoden oder Erkenntnissen relevante Lösungsansätze übersehen werden und/oder der Lösungsraum unnötig eingeschränkt wird.

Grundsätzlich gelten für Data-Science-Projekte dieselben Anforderungen wie für andere wissenschaftlich geprägte Vorhaben. Der Untersuchungsgegenstand ist klar zu definieren, sodass Zielsetzung und Kontext für Dritte nachvollziehbar sind. Die erzielten Ergebnisse müssen über bereits bekannte Erkenntnisse hinausgehen oder einen neuen Blickwinkel eröffnen. Gleichzeitig ist sicherzustellen, dass die Ergebnisse einen erkennbaren Nutzen besitzen und in geeigneter Form dokumentiert werden, sodass sie überprüfbar und reproduzierbar sind. Dabei sind insbesondere die Anforderungen an Objektivität, Reliabilität und Validität zu berücksichtigen.

Problem Statement

Die Problemstellung ist so zu formulieren, dass sie klar abgegrenzt und nachvollziehbar ist. Eine präzise Definition des Untersuchungsgegenstands bildet die Grundlage für die Auswahl geeigneter Methoden und für die Einordnung der Ergebnisse.

Data Provision

Die Eignung der Daten zur Bearbeitung der Fragestellung ist systematisch zu prüfen. Datenaufbereitungsschritte müssen transparent, nachvollziehbar und replizierbar gestaltet werden. Dies umfasst insbesondere die Dokumentation aller Transformationen sowie die Sicherstellung der langfristigen Verfügbarkeit der Rohdaten. In der explorativen Datenanalyse ist ein fundiertes Verständnis der Daten sicherzustellen. Potenzielle Fehlerquellen sind zu identifizieren und geeignete statistische Verfahren zur Bewertung der Daten anzuwenden.

Analysis

Für die Auswahl wie auch die Anwendung von Analyseverfahren gilt: Anforderungen und Zielsetzungen sind nachvollziehbar festzulegen und die Auswahl geeigneter Verfahren ist unter Berücksichtigung bestehender Ansätze sowie aktueller wissenschaftlicher Erkenntnisse zu treffen. Bei der Entwicklung neuer oder angepasster Verfahren ist darzulegen, inwiefern bestehende Ansätze genutzt, erweitert oder ersetzt werden und warum eine Neuentwicklung erforderlich ist. Bei der Parametrisierung von Analyseverfahren ist zielgerichtet vorzugehen. Dabei ist sicherzustellen, dass die zugrunde liegenden Annahmen der Verfahren erfüllt sind. Zur korrekten Anwendung sind geeignete wissenschaftliche Quellen heranzuziehen.

Analyseprozesse und Ergebnisse sind umfassend zu dokumentieren und nachvollziehbar zu interpretieren. Die Evaluation ist von Beginn an mitzudenken, wobei geeignete Test- und Validierungsansätze einzusetzen sind. Dies ist insbesondere erforderlich, um sicherzustellen, dass die Ergebnisse nicht lediglich statistische Artefakte der betrachteten Daten darstellen.

Deployment

Analyseverfahren und -ergebnisse sind so bereitzustellen, dass ihre Funktionsweise nachvollziehbar bleibt. Dies umfasst insbesondere eine angemessene Dokumentation sowie die transparente Beschreibung der eingesetzten Methoden und ihrer Grenzen.

Usage

Die Nutzung der Analyseartefakte ist systematisch zu beobachten und zu bewerten. Hypothesen über die Leistungsfähigkeit der Modelle sind im Nutzungskontext zu überprüfen. Dabei sind Unterschiede zwischen Entwicklungs- und Einsatzumgebung sowie Veränderungen in der Nutzungsumgebung zu dokumentieren.

10.3 IT Infrastructure

Die IT-Infrastruktur bildet die technische Grundlage für die Durchführung von Data-Science-Projekten. Sie beeinflusst sämtliche Phasen des Projekts, z. B. durch Einschränkungen hinsichtlich verfügbarer Technologien, Datenzugriffe oder Rechenkapazitäten.

Die konkrete Bedeutung der IT-Infrastruktur hängt stark von Art und Umfang des Projekts ab. Insbesondere die Komplexität der Daten, die Anforderungen an die Analyseverfahren sowie die geplante Form der Nutzbarmachung bestimmen, welche infrastrukturellen Voraussetzungen erforderlich sind. Gleichzeitig sind bestehende Systeme und organisatorische Rahmenbedingungen zu berücksichtigen, da sie die Gestaltung und Umsetzung von Lösungen maßgeblich beeinflussen. 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.

Problem Statement

Bereits in einer frühen Phase ist zu prüfen, ob das geplante Vorhaben mit der vorhandenen oder unter Berücksichtigung wirtschaftlicher Rahmenbedingungen realisierbaren IT-Infrastruktur umsetzbar ist. Dies umfasst insbesondere die Bewertung verfügbarer Systeme, Schnittstellen und technischer Ressourcen.

Data Provision

Die verfügbaren Zugriffsmöglichkeiten und Schnittstellen zu Datenquellen bestimmen maßgeblich, wie Daten erschlossen und verarbeitet werden können. Einschränkungen durch Quell- oder Zielsysteme können die Auswahl von Technologien beeinflussen.

Darüber hinaus ist die verfügbare Rechenleistung bei der Datenaufbereitung und explorativen Datenanalyse zu berücksichtigen. Insbesondere bei großen oder komplexen Datenbeständen kann es notwendig sein, Daten direkt in bestehenden Systemen zu verarbeiten oder geeignete Datenextraktionen vorzunehmen.

Analysis

Die Auswahl und Anwendung von Analyseverfahren sind eng mit den verfügbaren technischen Möglichkeiten verknüpft. Es ist zu bewerten, welche Infrastruktur für die Durchführung der Analysen erforderlich ist und ob diese vorhanden oder beschaffbar ist.

Zudem ist zu prüfen, ob Analysen direkt auf den vorhandenen Daten durchgeführt werden können oder ob eine separate Verarbeitung notwendig ist.

Deployment

Die IT-Infrastruktur muss geeignet sein, die entwickelten Analyseverfahren in der vorgesehenen Form bereitzustellen und zu betreiben. Dies umfasst insbesondere Aspekte wie Integration in bestehende Systeme, Zugriffsmöglichkeiten sowie Anforderungen an Verfügbarkeit, Skalierbarkeit und Sicherheit.

Darüber hinaus sind Möglichkeiten zur Wartung, Aktualisierung und Weiterentwicklung der bereitgestellten Lösungen zu berücksichtigen.

Usage

Im laufenden Betrieb ist zu überprüfen, ob die gewählte IT-Infrastruktur den Anforderungen weiterhin gerecht wird. Dies betrifft sowohl die Leistungsfähigkeit als auch die Effizienz der eingesetzten Systeme. Anpassungen können erforderlich werden, wenn sich Nutzungsanforderungen oder Rahmenbedingungen ändern.

Literatur

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.

Donoho, D. (2017). 50 Years of Data Science. Journal of Computational and Graphical Statistics, 26(4), 745–766.

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

Europ. Parl. (2026). Was ist künstliche Intelligenz und wie wird sie genutzt? https://www.europarl.europa.eu/topics/de/article/20200827STO85804/was-ist-kunstliche-intelligenz-und-wie-wird-sie-genutzt, aufgerufen am 22.03.2026.

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

Jayawardene, V., Sadiq, S., Indulska, M. (2013). The Curse of Dimensionality in Data Quality. *Australasian Conference on Information Systems (ACIS) 2013 Proceedings*, paper 165.

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.

Neuhaus, U., Schulz, M., Schröder, H. et al. Kompetenzfelder künftiger Beschäftigter im Bereich Künstlicher Intelligenz. HMD 61, 471–484 (2024). https://doi.org/10.1365/s40702-024-01046-7.

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.

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.

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.

Wicht, P.; Hausmann, A.; Hoseini, S.; Kaufmann, J. (2021): Aufbau eines Data-Science-Teams – „Lessons learned". In: *Wirtschaftsinformatik & Management* 13, 266–275. https://doi.org/10.1365/s35764-021-00350-x.

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.

Weitere Veröffentlichungen zum DASC-PM

Deutschsprachige Veröffentlichungen

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. (Hrsg.): Data Science anwenden – Einführung, Anwendungen und Projekte, ISBN: 978-3-658-33813-8, S. 127-144, https://doi.org/10.1007/978-3-658-33813-8_8.

Kaufmann, J., Kühnel, S., Theuerkauf, R., Alekozai, E. M., Hoseini, S., Neuhaus, U., & Schulz, M. (2021). Where is the Science in Data Science Projects? Online-Workshop über die Wissenschaftlichkeit von Vorgehensmodellen für Data-Science-Projekte (WISDAP), Gesellschaft für Informatik e.V. (GI) (Hrsg.): INFORMATIK 2021, Lecture Notes in Informatics (LNI), Bonn 2021, S. 1729-1741, https://doi.org/10.18420/informatik2021-150.

Kaufmann, J., Schulz, M., & Kühnel, S. (2025). Data Science – Projekte mit DASC-PM. In: WISU - Das Wirtschaftsstudium, Sektion Wirtschaftsinformatik, 54 Jg., Heft 3/2025, S. 263.

Schulz, M. (2020). Data-Science-Projekte und ihre Besonderheiten. In: Wirtschaftsinformatik & Management, 12, S. 376-381, https://doi.org/10.1365/s35764-020-00281-z.

Schulz, M., Neuhaus, U., Kaufmann, J., Badura, D., Kerzel, U., Welter, F., Prothmann, M., Kühnel, S., Passlick, J., Rissler, R., Badewitz, W., Dann, D., Gröschel, A., Kloker, S., Alekozai, E. M., Felderer, M., Lanquillon, C., Brauner, D., Gölzer, P., Binder, H., Rohde, H., & Gehrke, N. (2020). DASC-PM v1.0 - Ein Vorgehensmodell für Data-Science-Projekte, NORDAKADEMIE gAG Hochschule der Wirtschaft, Hamburg, Elmshorn 2020, ISBN: 978-3-00-064898-4, https://doi.org/10.25673/32872.

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, NORDAKADEMIE gAG Hochschule der Wirtschaft, Elmshorn 2022, ISBN: 978-3-9824465-0-9, https://doi.org/10.25673/85296.

Schulz, M., Neuhaus, U., & Kühnel, S. (2020). Data-Science-Prozessmodell (DASC-PM). In: WISU, Basiswissen Wirtschaftsinformatik, 49 Jg., Heft 4/2020, S. 387 ff.

Schulz, M., Neuhaus, U., Kühnel, S., Rohde, H., Hoseini, S., & Theuerkauf, R. (Hrsg.) (2023): DASC-PM v1.1 Fallstudien, NORDAKADEMIE gAG Hochschule der Wirtschaft, Hamburg 2023.

Theuerkauf, R., Daurer, S., Hoseini, S., Kaufmann, J., Kühnel, S., Schwade, F., Alekozai, E. M., Neuhaus, U., Rohde, H., & Schulz, M. (2022). Vorschlag eines morphologischen Kastens zur Charakterisierung von Data-Science-Projekten. In: Informatik Spektrum 45, https://doi.org/10.1007/s00287-022-01508-6.

Englischsprachige Veröffentlichungen

Kuehnel, S., Neuhaus, U., Kaufmann, J., Schulz, M., & Alekozai, E. M. (2023). Using the Data Science Process Model Version 1.1 (DASC-PM v1.1) for Executing Data Science Projects: Procedures, Competencies, and Roles. In: Barton, T. & Müller, C. (Hrsg.): Apply Data Science: Introduction, Applications and Projects, S. 119-134, ISBN: 978-3-658-38797-6 , https://doi.org/10.1007/978-3-658-38798-3_8.

Schulz, M., Neuhaus, U., Kaufmann, J., Badura, D., Kühnel, S., Badewitz, W., Dann, D., Kloker, S., Alekozai, E. M., & Lanquillon, C. (2020). Introducing DASC-PM: A Data Science Process Model, Australasian Conference on Information Systems, ACIS Proceedings, paper 45, 2020, Wellington, New Zealand, https://aisel.aisnet.org/acis2020/45, http://dx.doi.org/10.25673/92266.

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 - A Process Model for Data Science Projects, NORDAKADEMIE gAG Hochschule der Wirtschaft, Elmshorn 2022, ISBN: 978-3-9824465-1-6, http://dx.doi.org/10.25673/91094.

Schulz, M., Neuhaus, U., Kühnel, S., Rohde, H., Hoseini, S., & Theuerkauf, R. (Hrsg.) (2023). DASC-PM v1.1 Case Studies, NORDAKADEMIE gAG Hochschule der Wirtschaft, Hamburg 2023.

Verzeichnis der Autor:innen

Als Autorinnen und Autoren der Version 2.0 werden alle aktiv an der Bearbeitung dieser Version Beteiligten geführt, die dieser Nennung zugestimmt haben.

Sie bedanken sich bei allen Mitwirkenden der Versionen 1.1 und 1.0 für die Arbeit an den bisherigen Ausarbeitungen.

Prof. Dr. Michael Schulz, NORDAKADEMIE Hochschule der Wirtschaft

Prof. Dr. Jens Kaufmann, Hochschule Niederrhein

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

Dipl.-Inform. Uwe Neuhaus, Europa-Universität Flensburg

Heiko Rohde (M.Sc.), valantic Business Analytics GmbH

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

Prof. Dr. Carsten Lanquillon, Hochschule Heilbronn

Prof. Dr. Stephan Daurer, DHBW Ravensburg

Jonas Dieckmann (M.Sc.), Philips

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

Felix Welter (M.Sc.), CGI

Prof. Dr. Uwe Haneke, Hochschule Karlsruhe

Prof. Dr. Christian Beecks, FernUniversität in Hagen

Dr. Martin Böhmer, Martin-Luther-Universität Halle-Wittenberg

Prof. Dr. Bahne Christiansen, NORDAKADEMIE Hochschule der Wirtschaft

Prof. Dr. André Drews, Technische Hochschule Lübeck

Prof. Dr. Arne Ewald, NORDAKADEMIE Hochschule der Wirtschaft

Dr. Viktor Harkov, HEON GmbH

Franziska Herrmann (B.Sc.), NORDAKADEMIE Hochschule der Wirtschaft

Tim Hilbig (M.Sc.), Euler Hermes

Prof. Dr. Dirk Johannßen, Fachhochschule Westküste

Dipl.-Ing. oec. Lutz Kretschmann, Hapag-Lloyd AG

Prof. Dr.-Ing. Bernhard Meussen, NORDAKADEMIE Hochschule der Wirtschaft

Dr. Christian Pasold, Datron Consulting GmbH

Stefan Rösl (MBA), Ostbayerische Technische Hochschule Amberg-Weiden

Dr. Klaus Schmerler, Martin-Luther-Universität Halle-Wittenberg

Theo Schnelle de Lourenco (M.Sc.), Aclue GmbH

Prof. Dr. Martin Schultz, Hochschule für Angewandte Wissenschaften Hamburg

Prof. Dr. Michael Seifert, Technische Hochschule Rosenheim

Marcus Soll (M.Sc.), NORDAKADEMIE Hochschule der Wirtschaft

Dr. Robert Stahlbock, Universität Hamburg

Niklas Ullmann (M.Sc.), Synvert GmbH

dasc°pm