

ARBEITSPAPIER

DATA SCIENCE: LERN- UND AUSBILDUNGSGEHÄLTE

DEZEMBER 2019

ARBEITSPAPIER

DATA SCIENCE: LERN- UND AUSBILDUNGSGEHALTE

DEZEMBER 2019

Vorschläge des GI-Arbeitskreises „Data Science/Data Literacy“ für die inhaltliche Ausgestaltung von Data-Science-Studiengängen und -Weiterbildungsangeboten in Zusammenarbeit mit der vom Bundesministerium für Bildung und Forschung (BMBF) und acatech initiierten Plattform Lernende Systeme

INHALT

Geleitwort	3	4. Lern- und Ausbildungsinhalte für Data Science	17
Zusammenfassung	4	4.1 Persona A: Master of Data Science (M.Sc.)	19
1. Hintergrund	5	4.2 Persona B.1: Master of Data Science (M.Sc.) (in der Domäne)	21
1.1 Was ist Data Science?	5	4.4 Persona C.1: Basic Data Scientist	23
1.2 „Data Science/Data Literacy“ in der Gesellschaft für Informatik	6	4.5 Persona C.2: Advanced Data Scientist	25
1.3 Plattform Lernende Systeme: Technologische Wegbereiter und Data Science ...	6	5. Ausblick	27
1.4 Vorgehen des Arbeitskreises	7	Literatur	28
2. Data Science: Ein Überblick	8	Anhang	29
2.1 Begriffsklärung	8	Anhang 1: Aufschlüsselung der Data-Scientist-Lerninhalte	30
2.2 Data-Science-Kompetenzen	9	Anhang 2: Notwendige Voraussetzungen Persona B und C	33
2.2.1 acatech Data-Science- Schlüsselkompetenzen	9	Autorinnen und Autoren	36
2.2.2 EDISON Data Science Framework (EDSF)	9	Impressum	37
2.2.3 Data Literacy	10		
2.2.4 CRISP-DM	10		
2.2.5 IT-Skills-Studie	10		
2.3 Zusammenfassung der Strukturierungsansätze	10		
3. Data-Science-Perspektiven	12		
3.1 Persona A: Master of Data Science (M.Sc.)	12		
3.2 Persona B: Master of Data Science (M.Sc.) außerhalb der Informatik und Mathematik	13		
3.2.1 Persona B.1: M.Sc. Data Science (in Domäne) ..	14		
3.2.2 Persona B.2: M.Sc. in Domäne mit Data- Science-Kompetenzen	14		
3.3 Persona C: Weiterbildung zum Data Scientist	15		
3.3.1 Persona C.1: Basic Data Scientist	15		
3.3.2 Persona C.2: Advanced Data Scientist	15		

GELEITWORT

Sehr geehrte Damen und Herren,
liebe Leserin, lieber Leser,

die Begriffe „Künstliche Intelligenz“, „Big Data“ und „Data Science“ gehören zu den am meisten „gebrauchten“ der letzten Jahre. Die sogenannte „Künstliche Intelligenz“ (Artificial Intelligence) beschreibt dabei – jenseits des Marketingsprechers – die Erforschung „intelligenter“ Problemlösungsverhaltens sowie die Entwicklung „intelligenter“ Computersysteme. Dabei sind Methoden der Künstlichen Intelligenz, insbesondere des maschinellen Lernens und des Deep Learning mit künstlichen neuronalen Netzen, auf große Datenmengen in hinreichender Qualität angewiesen.

„Big Data“ bezeichnet Datenmengen, die zu groß, zu komplex, zu schnelllebig oder zu schwach strukturiert sind, um sie mit manuellen und herkömmlichen Methoden der Datenverarbeitung auszuwerten. Der große Datenumfang (Volume), die Geschwindigkeit, mit der die Datenmengen generiert und transferiert werden (Velocity), die Bandbreite der Datentypen und -quellen (Variety) sowie die Echtheit von Daten (Veracity) zeichnen dabei diese Daten aus. Erweitert wird diese Definition häufig um die den unternehmerischen Mehrwert (Value) und die Sicherstellung der Datenqualität (Validity).

Aufgrund der wissenschaftlichen und technologischen Entwicklungen in den Bereichen Big Data und der Künstlichen Intelligenz, die auch Methoden jenseits des maschinellen Lernens beinhalten – etwa die Mensch-Technik-Interaktion, wissensbasierte Systeme oder mathematische Logik – rückt insbesondere die Wissenschaft im Umgang mit Daten zunehmend in den Fokus.

Der Schwerpunkt dieser sogenannten „Data Science“ liegt dabei nicht bei den Daten selbst, sondern auf der Art und Weise, wie diese verarbeitet, aufbereitet, analysiert und in Entscheidungen umgesetzt werden. Data Science beschäftigt sich mit einer zweckorientierten Datenanalyse und der systematischen Generierung von Entscheidungshilfen und -grundlagen, um Wettbewerbsvorteile erzielen zu können.

Dieses neue Wissenschaftsfeld an der Schnittstelle zu verschiedenen Anwendungsbereichen – sowohl für Forschung als auch für die Lehre – erfährt einen enormen Bedeutungszuwachs. Deshalb hat die Gesellschaft für Informatik e.V. vor zwei Jahren die Task-Force „Data Science“ ins Leben gerufen. Diese interdisziplinäre Arbeitsgruppe geht der Frage nach, was einen Data Scientist in Abgrenzung zu bestehenden Wissenschaftsdisziplinen wie der Informatik ausmacht und



Michael Goedicke



Peter Liggesmeyer

welche Kompetenzen ein Datenwissenschaftler und eine Datenwissenschaftlerin mitbringen müssen.

Im November 2019 hat das Kabinett Eckpunkte einer Datenstrategie der Bundesregierung mit vier Handlungsfeldern beschlossen: So sollen die Datenbereitstellung und der Datenzugang verbessert, eine verantwortungsvolle Datennutzung befördert, die Datenkompetenz in der Gesellschaft erhöht und der Staat zum Vorreiter einer Datenkultur gemacht werden. Die Gesellschaft für Informatik e.V., die mit 20.000 Mitgliedern die größte Fachgesellschaft für Informatik im deutschsprachigen Raum ist, will diese Entwicklungen maßgeblich mitgestalten.

Diese Publikation einer interdisziplinären Autorenschaft ist in Zusammenarbeit mit der Plattform Lernende Systeme entstanden und soll mit der Ausgestaltung von Studiengängen sowie von Aus- und Weiterbildungsangeboten helfen, die richtigen Themen im Bereich Data Science zu adressieren.

Wir wünschen viel Spaß bei der Lektüre.

Ihr
MICHAEL GOEDICKE
Vize-Präsident der Gesellschaft für Informatik e.V.

Ihr
PETER LIGGESMEYER
Past-President und Sprecher der Präsidiums-Task-Force
„Data Science“ der Gesellschaft für Informatik e.V.

ZUSAMMENFASSUNG

Data Science wird sowohl in der Wirtschaft als auch in den (angewandten) Wissenschaften als eine der Schlüsseldisziplinen unserer Zeit angesehen. Die Politik hat das aufgegriffen und die Ausbildung von Data-Science-Expertinnen und -Experten zu einem Kernanliegen erklärt. Ziel ist es, Data Science und den Umgang mit Daten in allen Bereichen, insbesondere aber in den Hochschulen und Universitäten, zu einem zentralen Wissenschaftsfeld auszubauen.

Aufgrund der wachsenden Bedeutung dieses neuen Wissenschaftsfeldes und der großen Relevanz für die Informatik hat der Vorstand der Gesellschaft für Informatik (GI) 2018 den Arbeitskreis „Data Science/Data Literacy“ ins Leben gerufen, der es sich gemeinsam mit Partnern anderer Wissenschaftsdisziplinen aus den Natur- und Lebenswissenschaften zum Ziel gemacht hat, Empfehlungen für die Entwicklung von Studien- und Weiterbildungsangeboten auszusprechen.

Unter Beteiligung der Plattform Lernende Systeme wurde dieses Arbeitspapier zu Lern- und Ausbildungsinhalten im Bereich Data Science entwickelt. Dazu wurde einerseits ein Set an Lern- und Ausbildungsinhalten in 14 Kompetenzfeldern identifiziert: Grundlagen sowie Fortgeschrittene Mathematik (1&2) und Informatik (3&4), Kryptographie und Sicherheit (5), Datenethik und Data Privacy (6), Data Governance (7), Datenintegration (8), Datenvisualisierung (9), Data Mining (10), Maschinelles Lernen (11), Business Intelligence (12), Domänenspezifische Anwendungen und Kommunikation mit Fachexperten (13) und die Implementierung von Data Science in der Organisation (14).

Diese Kompetenzfelder wurden in Anlehnung an die Anderson-Krathwohl-Taxonomie zur Profilierung von drei idealtypischen Personengruppen jeweils nach den drei kognitiven Prozessdimensionen (Verständnis-Level) Verstehen (L1), Anwenden (L2) und Analysieren (L3) bewertet:

- Persona A besitzt demnach einen Bachelor in Informatik, Mathematik / Statistik oder ggf. in Data Science und verfügt damit über Kenntnisse in Statistik, in Information Engineering oder Künstliche Intelligenz (KI), kann dies über entsprechende ECTS nachweisen und möchte einen Master in Data Science erwerben, um später als Data Scientist in der Industrie oder Forschung tätig zu sein.
- Persona B hat Kompetenzen eines Bachelor in einer Domänenwissenschaft (dies kann ein technisches oder naturwissenschaftliches Fach sein, aber auch ein Fach im Bereich der Geistes- bzw. Kulturwissenschaften) und will Data-Science-Kompetenzen für die Domäne erwerben.
- Persona C steht mitten im Beruf und kann bereits einschlägige informatische und mathematische Kenntnisse nachweisen. Sie will Data-Science-Kompetenzen für die praktische Anwendung im Job erwerben.

Diese Herangehensweise soll es den Leserinnen und Lesern ermöglichen, einen schnellen Überblick über die Anforderungen an mögliche Hochschul- und Weiterbildungsprogramme im Bereich Data Science zu erlangen. So entstehen unterschiedliche Profile für unterschiedliche bildungsbiografische Hintergründe und Ziele.

1. HINTERGRUND

Wikipedia definiert Data Science nüchtern und allgemein als „Extraktion von Wissen aus Daten“¹. Einer weiteren Definition in der wissenschaftlichen Literatur nach ist Data Science ein Gebiet, das Wissen in Statistik, Hard- und Software sowie Anwendungsdomänen umfasst [Cleveland 2001]. Im Koalitionsvertrag der aktuellen Bundesregierung, und darüber hinaus, wird Data Science als eine der Schlüsseldisziplinen unserer Zeit angesehen. Data Science, so das erklärte Ziel, soll in Zukunft in allen Bereichen, insbesondere aber in den Hochschulen, zu einem zentralen Wissenschaftsfeld ausgebaut werden.

Dieses Papier geht der Frage nach, was Data Science eigentlich ist und welche Fertigkeiten damit verbunden sind. Ziel dieses Papiers ist es, den Verantwortlichen für Lehre in Hochschulen und Universitäten sowie für Aus- und Weiterbildung in Unternehmen Orientierung zum Thema Data Science zu geben. Es richtet sich an Verantwortliche, die mit der Entwicklung von Data-Science-Curricula befasst sind, an Entwicklerinnen und Entwickler von Weiterbildungsprogrammen, Praktikerinnen und Praktiker, die im Bereich datengestützter Geschäftsentwicklung und -prozesse tätig sind sowie an die interessierte Öffentlichkeit.

1.1 WAS IST DATA SCIENCE?

Data Science ist ein interdisziplinäres Wissenschaftsfeld, welches durch die Anwendung wissenschaftlich fundierter Methoden, Prozesse, Algorithmen und Systeme die Extraktion von Erkenntnissen, Mustern und Schlüssen sowohl aus strukturierten als auch aus unstrukturierten Daten ermöglicht. Nach einer Definition der acatech beschäftigt sich Data Science damit, wie sehr große Datenmengen erhoben, verarbeitet, aufbereitet und analysiert werden können. Laut acatech² lässt sich Data Science in vier Kernbereiche einteilen:

1. **Data Engineering** umfasst alle Methoden und Prozesse, die für die Speicherung, den Zugriff sowie die Rückverfolgbarkeit von Daten nötig sind.
2. **Data Analytics** beschäftigt sich mit der Datenanalyse.
3. **Data Prediction** befasst sich mit der Vorhersage von Themen und Situationen auf Basis von Erfahrungswissen.
4. **Maschinelles Lernen** ist ein Querschnittsbereich zu den anderen drei Bereichen und steht für die Entwicklung von

Algorithmen, die aus Daten (Erfahrungswissen) lernen, dabei Muster erkennen, Modelle generieren und darauf aufbauend Themen und Situationen vorhersagen können.

Der Schwerpunkt der Data Science liegt dabei nicht auf den Daten selbst, sondern auf der Art und Weise, wie diese verarbeitet, aufbereitet und analysiert werden. Data Science beschäftigt sich mit einer zweckorientierten Datenanalyse und der systematischen Generierung von Entscheidungshilfen und -grundlagen, um Wettbewerbsvorteile erzielen zu können.

Wie bereits deutlich wurde, weist das Feld der Data Science starke Bezüge zum Maschinellen Lernen auf. Häufig werden weitere Bezüge zur Künstlichen Intelligenz hervorgehoben. „Künstliche Intelligenz definiert Herausforderungen, die es zu lösen gilt, und entwickelt Lösungsansätze. Beim Maschinellen Lernen steht das Erlernen der Lösungen im Vordergrund.“ [KerstinTresp2019] Die einzelnen Schritte im Data-Science-Prozess können daher dazu dienen, die solcherart definierten Herausforderungen z. B. über maschinelle Lernverfahren zu meistern.

Das Thema Data Science und die Identifizierung der notwendigen Kompetenzen ist sowohl für die Unternehmenspraxis als auch für Lehre und Forschung an den Hochschulen von großer Relevanz. Im Unternehmensumfeld ist das Thema häufig im Bereich Business Intelligence angesiedelt. Unternehmen aus allen Branchen suchen händierend die auf große Datenmengen spezialisierten Analysten oder sehen die Notwendigkeit, diese selbst aus- und weiterzubilden. Bereits vor zwei Jahren ging die Unternehmensberatung McKinsey von 150.000 offenen Stellen allein in den Vereinigten Staaten aus. Diese Zahl dürfte heute noch größer sein.

In der Wissenschaft beschäftigt sich Data Science mit unterschiedlichen Teilbereichen und kann daher vor dem Hintergrund verschiedener akademischer Disziplinen betrieben werden: Informatik, Statistik, Mathematik, Natur- oder Wirtschaftswissenschaften, einschließlich des Maschinellen Lernens, des Statistischen Lernens, der Programmierung, der Datentechnik, der Mustererkennung, der Prognostik, der Modellierung von Unsicherheiten und der Datenlagerung. Mittlerweile existiert eine Reihe von Data-Science-Bachelor- und -Masterstudiengängen. [Lübcke2018] konstatiert,

1 https://de.wikipedia.org/wiki/Data_Science#cite_note-1 (01. Juli 2019).

2 https://www.acatech.de/wp-content/uploads/2019/02/acatech_NKM_Data_Science_WEB-2.pdf (01. Juli 2019).

dass die Hochschulen und Universitäten in Deutschland seit dem Jahr 2014 verstärkte Anstrengungen unternehmen, um das Spektrum an Studienangeboten im Bereich Data Science kontinuierlich zu erweitern.³

1.2 „DATA SCIENCE/DATA LITERACY“ IN DER GESELLSCHAFT FÜR INFORMATIK

Aufgrund der wachsenden Bedeutung dieses neuen Wissenschaftsfeldes an der Schnittstelle zu verschiedenen Anwendungsbereichen – sowohl für die Industrie, die Forschung und Wissenschaft als auch für die Lehre – hat der Vorstand der Gesellschaft für Informatik (GI) 2018 die Task-Force „Data Science/Data Literacy“⁴ ins Leben gerufen. Dieser interdisziplinäre Vorstandsarbeitskreis ging der Frage nach, was die Data Science in Abgrenzung zu bestehenden Wissenschaftsdisziplinen ausmacht und welche Kompetenzen ein Datenwissenschaftler und eine Datenwissenschaftlerin mitbringen müssen.

Unter Beteiligung von mehr als 80 Expertinnen und Experten und in mehreren Workshops entstanden diverse Publikationen:

- In einem Studienprojekt für das Hochschulforum Digitalisierung wurden zunächst die grundlegenden digitalen Kompetenzen in der Breite der Studierendenschaft identifiziert: Dazu gehört **Data Literacy** bzw. die grundsätzliche Fähigkeit des planvollen Umgangs mit Daten. Im Gegensatz zu Data Science liegt der Fokus auf der Vermittlung von bedarfsgerechtem, disziplinenübergreifendem Know-how, um datengestützt arbeiten und entscheiden zu können.⁵
- Ein **Policy-Paper** der Task-Force in Zusammenarbeit mit Vertreterinnen und Vertretern der Deutschen Mathematiker-Vereinigung e.V., der Deutschen Physikalischen Gesellschaft e.V. und der Gesellschaft Deutscher Chemiker e.V. beschäftigt sich mit „Data Literacy und Data Science Education: Digitale Kompetenzen in der Hochschulausbildung“.⁶

Im Vorfeld der Erstellung dieses White Papers hat sich der Arbeitskreis noch einmal auf eine kleine Expertengruppe kondensiert. Dabei hat sich insbesondere die breite Beteiligung aus unterschiedlichen wissenschaftlichen Disziplinen

sowie der Unternehmenspraxis als Herausforderung herausgestellt. All diese Gruppen haben unterschiedliche Interessen am Thema und einen unterschiedlichen Blick darauf. Konsens bestand aber bei allen Beteiligten darüber, dass es wünschenswert sei, Data-Science-Kernkompetenzen zu identifizieren – sowohl für die Weiterbildung in Unternehmen als auch für die Lehre an Universitäten und Hochschulen.

1.3 PLATTFORM LERNENDE SYSTEME: TECHNOLOGISCHE WEGBEREITER UND DATA SCIENCE

Auch die Plattform Lernende Systeme⁷, die vom Bundesministerium für Bildung und Forschung (BMBF) auf Anregung des Fachforums Autonome Systeme des Hightech-Forums und acatech – Deutsche Akademie der Technikwissenschaften initiiert wurde, sieht Data Science als eine Schlüsselqualifikation in der akademischen Ausbildung. Die Plattform Lernende Systeme vereint Expertise aus Wissenschaft, Wirtschaft, Politik und Gesellschaft, und unterstützt den weiteren Weg Deutschlands zu einem international führenden Technologieanbieter für Lernende Systeme. Sie versteht sich als ein Ort des Austauschs und der Kooperation. In sieben interdisziplinären und branchenübergreifenden Arbeitsgruppen (AG) arbeiten rund 200 Expertinnen und Experten aus Wissenschaft, Unternehmen unterschiedlicher Größe, Politik und Zivilgesellschaft zusammen und erörtern im regelmäßigen Austausch technologische, wirtschaftliche, rechtliche und gesellschaftliche Fragen, die mit der Entwicklung und Einführung von Lernenden Systemen und Künstlicher Intelligenz verbunden sind.

Die Arbeitsgruppe 1 „Technologische Wegbereiter und Data Science“ befasst sich mit den technologischen Grundlagen und Enablern von Lernenden Systemen. Sie übernimmt innerhalb der Plattform eine Querschnittsfunktion und gibt Impulse an alle weiteren Arbeitsgruppen. Eine Leitfrage ihrer Arbeit ist, wie die Ausbildung von Forscherinnen, Forschern und Fachkräften für Maschinelles Lernen und Data Science an Hochschulen weiter verbessert werden kann, damit deutschlandweit Faktoren geschaffen werden, die den schnellen und erfolgreichen Einsatz von Maschinellen Lernen und Data Science begünstigen.

3 https://his-he.de/fileadmin/user_upload/Publikationen/Forum_Hochschulentwicklung/Forum_HE_201801_Web.pdf (08. November 2019).

4 <https://gi.de/datascience/> (08. November 2019).

5 <https://gi.de/datascience/> (08. November 2019).

6 https://gi.de/fileadmin/GI/Hauptseite/Aktuelles/Aktionen/Data_Literacy/GI_DataScience_2018-04-20_FINAL.pdf (08. November 2019).

7 www.plattform-lernende-systeme.de (11. November 2019).

Weitere zentrale Fragestellungen der Arbeitsgruppe 1 befassen sich mit dem Potenzial wichtiger Forschungsfelder bei Künstlicher Intelligenz, maschinellem Lernen und Data Science für disruptive Anwendungen, den Stärken und Schwächen der KI-Forschung in Deutschland sowie dem Wissenstransfer von der Forschung in die Anwendung.

Das vorliegende Papier ist in Kooperation mit der Arbeitsgruppe 1 „Technologische Wegbereiter und Data Science“ der Plattform Lernende Systeme entstanden und wird explizit von ihr befürwortet.

1.4 VORGEHEN DES ARBEITSKREISES

Der GI-Arbeitskreis „Data Science/Data Literacy“ hat eine Reihe von Lerninhalten, Kompetenzen und Fähigkeiten identifiziert und diese drei verschiedenen Personas zugeordnet. So erhalten Leserinnen und Leser Anhaltspunkte, welche Qualifikationen auf dem Weg zum Data Scientist vermittelt werden sollten. Diese drei Personas repräsentieren unterschiedliche Adressaten mit unterschiedlichen Bildungsbiografien.

Die Ergebnisse der Arbeit des Arbeitskreises halten Sie in Ihren Händen. Das vorliegende White Paper gibt Empfehlungen zur Ausgestaltung und Professionalisierung von Kompetenzprofilen im Bereich Data Science. Die Autorinnen und Autoren freuen sich über Feedback unter berlin@gi.de und wollen in jährlichem Turnus auf das Papier blicken, um ggf. neue Entwicklungen in den Datenwissenschaften, den Technologien und ihren Anwendungsbereichen aufzunehmen.

2. DATA SCIENCE: EIN ÜBERBLICK

Die Datenwissenschaft ist keine Erfindung der letzten Jahre: Der Begriff „Data Science“ stammt aus den Anfängen der Datenhaltung und -analyse, die bis in die 1960er-Jahre zurückgehen. Mit der zunehmenden Bedeutung von „Big Data“ rückte die Wissenschaft der Daten verstärkt in den Fokus. Die Bedeutung der statistischen Datenanalyse für ein Verständnis von Daten – gerade für große Datenmengen – wurde etwa von John Tukey bereits in einem Artikel von 1962 vorhergesehen. Ein Informatik, Statistik, Mathematik und Anwendungsdomänen umspannendes Curriculum für Data Science wurde 2001 vorgeschlagen [Cleveland2001]. Der Schwerpunkt der Data Science liegt dabei nicht bei den Daten selbst, sondern auf der Art und Weise, wie diese verarbeitet, aufbereitet und analysiert werden. Data Science beschäftigt sich mit einer zweckorientierten Datenanalyse und der systematischen Generierung von Entscheidungshilfen und -grundlagen.

Im Unternehmensumfeld ist das Thema häufig im Bereich Business Intelligence angesiedelt; in einem Beitrag im Harvard Business Manager hat Tom Davenport Data Science zum „attraktivsten Beruf des 21. Jahrhunderts“ gekürt.⁸ Mittlerweile existiert in Deutschland und international eine Reihe von Data-Science-Studiengängen auf Bachelor- und auf Master-Niveau.

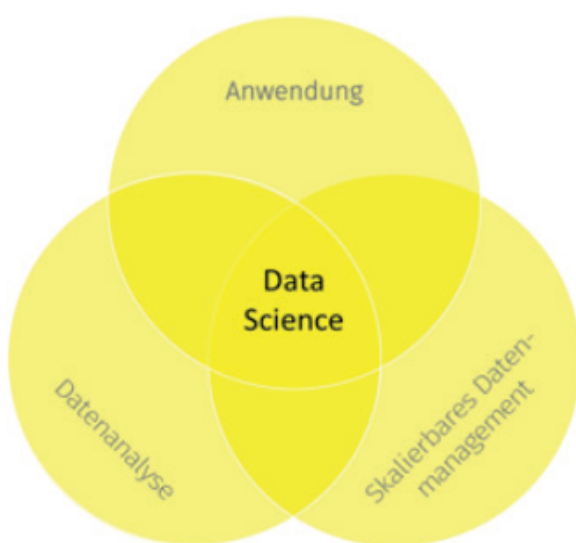


Abbildung 1: Die Anforderungen an Data Science an der Schnittstelle zwischen Datenmanagement, Datenanalyse und Anwendungsbezug

Volker Markl weist zu Recht darauf hin, dass in der Vergangenheit die Disziplinen der Datenanalyse und der skalierbaren Datenverarbeitung nicht eng genug miteinander verzahnt waren, was jedoch für einen souveränen Umgang mit großen Datenmengen von geringer Latenz erforderlich ist [Markl2015]. Zudem sind das Wissen aus der Anwendungsdomäne und die juristischen und gesellschaftlichen Implikationen zu beachten. Markl vergleicht deshalb die anspruchsvollen Anforderungen, die auf den Data Scientist projiziert werden, mit der Suche nach der eierlegenden Wollmilchsaue.

2.1 BEGRIFFSKLÄRUNG

Der Begriff „Data Science“ wird in der Praxis häufig recht schwammig verwendet und adressiert unterschiedlichste Kompetenzfelder in verschiedener Tiefe. Einigkeit herrscht allerdings darüber, dass es im Wesentlichen um die Vermittlung von Fachwissen in Bezug auf den Umgang mit sehr großen Datenmengen (Big Data) geht. Die GI definiert drei Kerndimensionen von Big Data: Diese beziehen sich „[...] auf ein ansteigendes Volumen (engl. volume) der Daten, auf eine ansteigende Geschwindigkeit (engl. velocity), mit der Daten erzeugt und verarbeitet werden, und auf eine steigende Vielfalt (engl. variety) der erzeugten Daten“ [GI2013]. Data Science formt sich international als interdisziplinäres Wissenschaftsgebiet, das wissenschaftliche Methodiken für die Informations- und Erkenntnisgewinnung aus Daten durch Aufbereitung, Analyse und Inferenz von sehr großen, hochdimensionalen Datenbeständen anwendet und erforscht [DStatG2019].

Nach [acatech2017] beschäftigt sich Data Science mit der Art und Weise, wie Big Data erhoben, verarbeitet, aufbereitet und analysiert werden. Eine der vollständigsten Betrachtungen von Data Science und damit in Bezug stehenden Kompetenzen findet sich im EDISON Data Science Framework (EDSF). Demnach bezeichnet der Begriff „Data Scientist“ einen „Anwender, welcher über ausreichendes Wissen und Expertise in den Bereichen Business Needs, Domänenwissen, analytische Fähigkeiten, Programmierung und Systems Engineering verfügt, um den wissenschaftlichen Prozess durchgehend über alle Stufen des Big-Data-Lifecycles bis zur Lieferung eines erwarteten wissenschaftlichen oder geschäftlichen Nutzens für eine Organisation oder ein Projekt durchführen zu können“ [Edison2019]. Darüber hinaus wird der Begriff des „Data Steward“ eingeführt als „ein Profi im Umgang und

⁸ <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century> (07. November 2019).

Management von Daten, dessen Verantwortung die Planung, Umsetzung und die Verwaltung von (Forschungs-) Daten in Bezug auf Zugang, Speicherung, Suche und Präsentation umfasst“ [Edison 2019]. Der Data Steward stellt folglich das Bindeglied zwischen dem Data Scientist und den in den Domänen Forschenden dar.

Eine von vielen Praktikerinnen und Praktikern genutzte Definition stammt von Drew Conway. Diese stellt Data Science als eine domänenübergreifende Disziplin dar, die in der Schnittmenge grundlegende informatische und mathematische Kenntnisse erfordert sowie Sachkenntnis in Bezug auf die Anwendungsdomäne. Dies wird u.a. in [Kauermann2019] und [Weihsickstadt2018] näher begründet diskutiert.

Neben dem Begriff „Data Science“ gibt es in der Literatur eine ganze Reihe von Begriffen, die sich teilweise überlappen [Heidrich2018]:

- „Data Management“ bezeichnet eine Disziplin in der Informatik, die sich mit dem Steuern, Schützen, Ausliefern und Verbessern des Werts von Daten beschäftigt.
- „Data Literacy“ bezeichnet die grundlegende Fähigkeit, Daten auf kritische Art und Weise zu sammeln, zu managen, zu bewerten und anzuwenden.
- „Information Literacy“ bezeichnet die Fähigkeit, Informationen aus verschiedenen Formaten zu finden, zu managen und zu verwenden.
- „Data Information Literacy“ bezeichnet die Anwendung von Information Literacy im Forschungskontext.
- „Science Data Literacy“ bezeichnet die Fähigkeit, wissenschaftliche Daten zu verstehen, zu verwenden und zu managen.
- „Digital Literacy“ bezeichnet die Fähigkeit, Informationen unter Nutzung digitaler Technologien finden, organisieren, verstehen, evaluieren und erzeugen zu können.
- „Statistical Literacy“ bezeichnet die Fähigkeit, auszuwählen, was gezählt bzw. gemessen wird, wie eine zusammenfassende Statistik erzeugt wird, welche Vergleiche damit angestellt werden dürfen und wie die Ergebnisse kommuniziert werden sollen.

Die oben aufgeführten Begriffe stellen teilweise Teilaspekte von Data Science dar (z.B. Management von Big Data) oder bezeichnen eher grundlegende Kompetenzen im richtigen Umgang mit Daten – ohne die speziellen Herausforderungen von Big Data (wie z.B. die oben aufgeführten Literacy-Definitionen).

2.2 DATA-SCIENCE-KOMPETENZEN

Um den Begriff „Data Science“ greifbarer zu machen und abzuleiten, welche Inhalte tatsächlich vermittelt werden sollten, ist es sinnvoll, Kompetenzfelder und Wissensbausteine zu definieren. Dazu wurden verschiedene existierende Strukturierungsansätze betrachtet und bewertet.

2.2.1 ACATECH DATA-SCIENCE-SCHLÜSSELKOMPETENZEN

Ein Expertenpanel der acatech hat im Rahmen eines vom BMBF geförderten Projekts Schlüsselkompetenzen für Data Science identifiziert, die es neben den Kernbereichen Data Engineering, Data Analytics, Data Prediction und Maschinelles Lernen zu fördern gilt. Dazu wurden auch verschiedene Länder bzgl. dieser Kompetenzen verglichen [acatech2017]. Die Schlüsselkompetenzen umfassen (1) Informatik auf Universitätsniveau, (2) Datengetriebene Geschäftsmodelle, (3) Forschung und Entwicklung für Data Analytics, (4) Open Data bei staatlichen Daten, (5) Klarheit in den Rechtsfragen.

Vorteile: einfache, leicht zu vermittelnde Kernbereiche

Nachteile: Fokus auf Bewertung der Schlüsselkompetenzen und nicht auf Vermittlung von Kompetenzfeldern und Wissensbausteinen

2.2.2 EDISON DATA SCIENCE FRAMEWORK (EDSF)

Das EDSF ist aus einem Forschungsprojekt der Europäischen Union im Rahmenprogramm Horizon 2020 (Grant 675419) entstanden. Seit Ende des Projektes wird das Framework durch die EDISON Community gewartet, welche durch die Universität von Amsterdam koordiniert wird [Edison2019].

In der Version 3 des EDSF werden fünf Kompetenzgruppen definiert: Data Analytics, Data Science Engineering, Data Management, Research Methods and Project Management und Business Analytics (bzw. eine domänenspezifisch auszugestaltende Kompetenzgruppe). Für diese Gruppen werden 30 Kompetenzen definiert, die 52 Fähigkeiten einschließen. Parallel dazu definiert der EDISON Body of Knowledge (BoK) 23 Knowledge Areas und 207 Knowledge Units. Darüber hinaus werden Empfehlungen für die Entwicklung von Curricula im Kontext Data Science gegeben.

Vorteile: umfassendes Kompetenz-Framework, Curricula-Empfehlungen

Nachteile: sehr komplex, Definitionen teils nicht intuitiv, schwer vermittelbar, teilweise Redundanzen

2.2.3 DATA LITERACY

[Ridsdale2015] fasst bestehende Definitionen von Data Literacy zusammen und definiert einen Kompetenzrahmen, der fünf Bereiche umfasst: (1) Einführung in Daten (konzeptioneller Rahmen), (2) Datensammlung, (3) Datenmanagement, (4) Datenevaluation und (5) Datenanwendung. In diesen Bereichen werden 22 Kompetenzen definiert.

Auch wenn sich der Kompetenzrahmen nicht direkt auf Data Science bezieht, so scheint die grundsätzliche Strukturierung durchaus übertragbar.⁹

Vorteile: einfach und intuitiv

Nachteile: Fokus auf sehr allgemeine Kompetenzen im Umgang mit Daten; einige Spezifika von Data Science fehlen

2.2.4 CRISP-DM

CRISP-DM steht für Cross-Industry Standard Process for Data Mining. Es beschreibt die für eine Data-Mining-Fragestellung essentiellen Schritte und wird branchenübergreifend in Forschung und Industrie in der Breite eingesetzt. Es wurde 1996 im Rahmen eines Forschungsprojekts der Europäischen Union unter Beteiligung namhafter Industriepartner entwickelt [SPSS2000]. CRISP-DM definiert sechs Phasen bzw. Schritte, die durchlaufen werden sollten: (1) Business Understanding, (2) Data Understanding, (3) Data Preparation, (4) Modelling, (5) Evaluation und (6) Deployment.

Vorteile: einfach und intuitiv

Nachteile: Fokus auf den durchzuführenden Prozess und nicht auf die Vermittlung von Kompetenzfeldern und Wissensbausteinen

2.2.5 IT-SKILLS-STUDIE

Die von der Firma Data Assessment Solutions (DAS) durchgeführte IT-Skills-Studie [DAS2014] untersucht die in einem Unternehmen vorrangig benötigten Fähigkeiten und Kompetenzen (sog. Skills), um Big-Data-Projekte realisieren zu können. Es werden fünf Skills definiert, die auf bestimmten Kompetenzniveaus (Laie/in, Kenner/in, Könnler/in, Exper-

te/in) erlangt werden sollen: (1) Statistik und statistische Programmiersprachen, (2) Big-Data-Infrastruktur, (3) Business-Domänenwissen, (4) Datenintegration und -transformation und (5) Präsentation und Visualisierung.

Vorteile: einfach und intuitiv, Definition von Skill-Niveaus

Nachteile: Fehlen einiger Spezifika von Data Science

2.3 ZUSAMMENFASSUNG DER STRUKTURIERUNGSANSÄTZE

Die folgende Tabelle bildet die betrachteten Ansätze acatech, CRISP-DM, Data Literacy und IT-Skills auf das EDISON Data Science Framework (Version 3) grob ab. Es wird deutlich, dass sich die Ansätze auf der Ebene der Gruppierung von Kompetenzen zwar recht gut aufeinander abbilden lassen, dass zugleich allerdings im Wesentlichen nur Teilaspekte der Kompetenzgruppen des EDSF von acatech, CRISP-DM, Data Literacy und IT-Skills abgedeckt werden. Beispielsweise umfasst der Bereich „Data Analytics (DSDA)“ des EDSF deutlich mehr Kompetenzen als der Bereich „Hypothesis and Modelling“ des CRISP-DM.

Über diese Initiativen und Projekte hinaus gibt es eine Reihe weiterer, die das Thema adressieren, wie etwa das Da.Re.-Projekt¹⁰ (Data Science Pathways to Re-imagine Education). In dem von der Europäischen Kommission finanzierten Erasmus+-Projekt, hat ein Team aus Universitäten, Unternehmen und einem Verbund aus Italien, Portugal, dem Vereinigten Königreich, Slowenien und Serbien den Bedarf an Bildung in der Datenwissenschaft ermittelt, um Europas wachsenden sozialen und wirtschaftlichen Bedarf im öffentlichen und privaten Sektor an Data Scientists zu decken.

Die Deutsche Statistische Gesellschaft stellt in einem Positionspapier¹¹ heraus, dass Data Science Kompetenzen und Fähigkeiten erfordert, die bisher in der Regel nur verteilt über die Fächer Informatik, Mathematik und Statistik vorlagen. Die Statistik befasst sich mit den Kernfragen für Datenverständnis und Wissensextraktion. Dies sind Datendeskription, Datenexploration und Datenanalyse sowie Stichprobentheorie und Inferenzstatistik.

⁹ Der Kurs „Data8: The Foundations of Data Science“ der UC Berkeley wiederum kombiniert zur Vermittlung einer Data Literacy drei Perspektiven: Inferential Thinking, Computational Thinking und Real-World-Relevanz (<http://data8.org>). Einen ähnlichen Ansatz verfolgt die TU Berlin mit ihrem Kurs „Data Science - Essentials of Data Programming“ https://www.bigdama.tu-berlin.de/menue/teaching/ss_2019/data_science_essentials_of_data_programming/ (4. Dezember 2019).

¹⁰ <http://dare-project.eu> / <https://www.kdnuggets.com/2019/10/european-approach-masters-data-science.html> (08. November 2019).

¹¹ Siehe DStatG-Positionspapier „Die Rolle der Statistik für Big Data, Data Literacy, Machine Learning, KI, Analytics und Data Science“, 2019 (<https://dstatg.de/positionspapier-die-rolle-der-statistik-fuer-big-data-data-literacy-machine-learning-ki-analytics-und-data-science>).

EDSF V3	ACATECH	CRISP-DM	DATA LITERACY	IT-SKILLS
Business Analytics (DSBA) und Domänen-Spezifika	Datengetriebene Geschäftsmodelle	Business Understanding	-	Business- Domänen-wissen
Data Management (DSDM)	Data Engineering	Data Understanding	Datensammlung	Datenintegration und -transformation
		Data Preparation	Datenmanagement	
			Datenanwendung	
Data Analytics (DSDA)	Data Analytics	Modelling	Datenevaluation	Statistik und statistische Programmiersprachen
	Data Prediction	Evaluation		Präsentation und Visualisierung
	Maschinelles Lernen			
Research Methods and Project Management (DSRMP)	Forschung und Entwicklung für Data Analytics	Modelling	Konzeptioneller Rahmen	-
	Open Data	Evaluation	Datenevaluation	
Data Science Engineering (DSENG)	Data Engineering	Deployment	Datenevaluation	Big-Data-Infrastruktur
	Rechtsfragen		Datenanwendung	

Abbildung 2: Vergleich unterschiedlicher Data-Science-Strukturierungsansätze

In einigen der bisher angeführten Strukturierungsansätzen für Data-Science-Kompetenzen werden die Kompetenzen des kritischen Hinterfragens der Datenqualität und des ethischen Blicks auf Data Science als Teilaspekte erwähnt. Die Auswirkungen mangelhafter Datenqualität und daraus resultierende mangelhafte Modelle waren in der Vergangenheit häufig Gegenstand öffentlicher Debatten [Beck2019]. Data Scientists sollten daher darin geschult sein, ethische Problemstellungen im Data-Science-Prozess zu identifizieren. Künftige Data Scientists sollten etwa in der Lage sein, die Datenqualität aus ethischer Perspektive hinterfragen zu können, beispielsweise hinsichtlich systematischer Verzerrungen in Datensätzen, die zu diskriminierenden Modellen führen, oder bzgl. mangelndem Bewusstsein über die Schutzwürdigkeit von Daten. Der kontinuierliche kritische Blick auf die Datenqualität über ethische Gesichtspunkte hinaus ist

als wichtige Kompetenz eines Data Scientists zu werten, da die Qualität der Ergebnisse einer Datenanalyse nicht zuletzt auch von der Datenqualität abhängt. Datenethik ist somit zu den Kompetenzfeldern der Data Science zu zählen, ebenso wie die Fähigkeit zur kritischen Reflexion der Datenqualität. Insbesondere bei Forschungsdaten, bei denen es wichtig ist, dass diese entlang des Lebenszyklus ungehindert und verlustfrei „fließen“ können, wurden sogenannte „FAIR Data Principles“ formuliert. Diese Grundsätze formulieren Prinzipien, die nachhaltig nachnutzbare Forschungsdaten erfüllen müssen und die Forschungsdateninfrastrukturen dementsprechend im Rahmen der von ihnen angebotenen Services implementieren sollten. Gemäß der FAIR-Prinzipien sollen Daten „Findable, Accessible, Interoperable, and Re-usable“ sein.

3. DATA-SCIENCE-PERSPEKTIVEN

Unabhängig vom verwendeten Kompetenzrahmen hängt die Wichtigkeit und Tiefe der Vermittlung einzelner Kompetenzbausteine von der konkreten Zielgruppe ab. Dazu unterscheiden wir grundlegend drei Personas, die in diesem Kapitel etwas differenzierter dargestellt werden:

1. Persona A besitzt in der Regel einen Bachelor in Informatik oder Mathematik bzw. Statistik oder ggf. in Data Science sowie Vorwissen in Information Engineering oder KI und kann dies über entsprechende ECTS nachweisen. Sie möchte einen Master of Data Science (M.Sc.) erwerben und später als Data Scientist in der Industrie oder Forschung tätig sein.
2. Persona B besitzt in der Regel einen Bachelor in einer Domänenwissenschaft (die von einem technisch-naturwissenschaftlichen Fach bis zu Geistes- bzw. Kulturwissenschaften reichen kann). Persona B.1 möchte einen Master in Data Science mit Spezialisierung in der Domäne des Bachelors erwerben und später im Bereich der Domäne als Data Scientist für Industrie und Forschung tätig sein. Persona B.2 möchte den Master in der Domänenwissenschaft mit Spezialisierung in Data Science erwerben und in Industrie oder Forschung als Domänenexpertin bzw. -experte mit einer Data-Science-Zusatzqualifikation tätig sein.
3. Persona C steht mitten im Beruf und kann gewisse informatische und mathematische Kenntnisse nachweisen. Persona C.1 hat das Ziel, grundlegende Basiskenntnisse zu erwerben, die jeder Data Scientist mitbringen sollte, um sie im Unternehmen einsetzen zu können. Persona C.2 beabsichtigt, Data-Science-Kompetenzen in der Breite zu vertiefen und sich als Expertin oder Experte für das Thema mit Anwendungserfahrung zu positionieren.

3.1 PERSONA A: MASTER OF DATA SCIENCE (M.SC.)

Die hier beschriebene Persona wird nur für den Data-Science-Master betrachtet. Obwohl es in Deutschland bereits den Bachelor gibt (etwa an der TU Dortmund, FH Zwickau, Uni Potsdam, FH Kiel, TH Ostwestfalen-Lippe, Uni Marburg u.a.), wird dieser hier nicht betrachtet. Die Arbeitsgruppe erachtet es als sinnvoll, Schulabsolventinnen und -absolventen zunächst in einem der traditionellen Studienfächer zu beheimaten und ihnen dann die Möglichkeit zur Vertiefung im Master of Data Science (M.Sc.) zu ermöglichen. Dieses Vorgehen betont die Notwendigkeit einer engen Verzahnung der Data Science mit den fachlichen Disziplinen aller Richtungen (Naturwissenschaften, Wirtschaftswissenschaften, Geisteswissenschaften), die im Zuge der Digitalisierung und der Datafizierung einem disruptiven Wandel unterzogen

sind. Zudem erlaubt dieses Vorgehen die graduelle Öffnung des Data-Science-Masters für Quereinsteiger und Quereinsteigerinnen, die bei einem konsekutiven Bachelor-/Masterstudiengang nicht möglich wäre. Für einen Masterstudiengang sollten in der Regel eine Vorqualifikation in einem der Teilbereiche (Informatik, Mathematik, Statistik, ggfs. Naturwissenschaften mit starkem Datenbezug, Data-Science-Bachelor) und Kenntnisse in wissenschaftlicher Methodik vorausgesetzt werden. Andererseits ist Data Science oft auch im Kontext mit einer Anwendungswissenschaft (Domain) zu sehen, aus der die Studierenden einen Bachelor-Abschluss oder Arbeitserfahrungen aus der Industrie mitbringen. Grundlegende Mathematik-, Statistik- und Informatik-Kenntnisse bilden die Voraussetzung für eine Vertiefung im Masterstudiengang. Der Master ist daher nur in diesem basalen Sinne konsekutiv.

Der Master of Data Science (M.Sc.) unterscheidet sich vom Informatik-Master. Beide Studiengänge teilen in der Regel lediglich Grundlagen und einige fortgeschrittene Themen der Informatik und Statistik/Mathematik gemäß der unten aufgeführten Thementabelle.

Es ergeben sich daher die folgenden Zielrichtungen:

- **Abschluss/Studiengang:** Master of Data Science (M.Sc.)
- **Vorqualifikation/Technologie- und Datenkompetenz:** In der Regel sollte ein Bachelor of Science (B.Sc.) in Informatik/Mathematik/Statistik (ggfs. auch Naturwissenschaften mit starkem Datenbezug oder Data-Science-Bachelor) und Vorwissen in Statistik, Data Engineering oder KI vorausgesetzt werden. Einige Studiengänge setzen dies voraus.
- **Ausbildung:** Universitäten und Hochschulen der angewandten Wissenschaften
- **Berufsbild/Erwartung Arbeitsmarkt:** Die beiden wichtigsten Berufsbilder sind hier Data Engineers bzw. Data Scientists in der Industrie und Wissenschaftler/Wissenschaftlerinnen in der Forschung (und der Lehre). Darüber hinaus gibt es auch hier zusätzliche Berufszweige im Management oder im Bereich Business Analytics, die Erfahrung und Kenntnisse im Umgang mit Daten/Maschinellen Lernen erfordern.
- **Fokus:** Im Fokus stehen daher gleichberechtigt die Anwendung in der Industrie – darunter auch Consulting – und die Wissenschaft/Forschung. Aber auch in anderen Bereichen wie Ethik und Politik oder in angrenzenden Feldern wie IT-Sicherheit werden Data-Science-Absolventinnen und -Absolventen benötigt.
- **Interesse/Motivation:** Aus den genannten Zielrichtungen ergeben sich verschiedene denkbare Motivationen für ein

Data-Science-Studium: Qualifikation für eine wissenschaftliche Karriere, ggf. mit Promotion, oder Karrieremöglichkeit in der Industrie, in der Absolventinnen und Absolventen mit verschiedenen Domänen (Pharmaindustrie, Handel/Marketing, Materialtechnik, Medizintechnik, Finanzwirtschaft/E-Commerce, Transport/Logistik etc.) in Berührung kommen und dort ihr Wissen anwenden.

3.2 PERSONA B: MASTER OF DATA SCIENCE (M.SC.) AUSSERHALB DER INFORMATIK UND MATHEMATIK / STATISTIK¹²

Data Scientists werden sich künftig auch aus Anwendungsdomänen außerhalb der Mathematik / Statistik und Informatik herausbilden. Dieser Personenkreis lässt sich wie eingangs beschrieben weiter differenzieren in B.1: Bachelor-Absolventinnen und -Absolventen in einer Domain-Wissenschaft (Natur-/Ingenieur- oder Kultur-/Geisteswissenschaften, d.h. mit sehr unterschiedlicher Vorbildung), die auf Basis grundlegender Kenntnisse in Datenwissenschaften (die ebenfalls erst in den Bachelorstudiengängen verankert werden müssen) ein Masterprogramm in Data Science absolvieren möchten.

Persona B.2 betrifft analog vorgebildete Bachelor-Absolventinnen und -Absolventen, die im Master innerhalb der Domäne verbleiben, aber im Masterstudium vertiefte Data-Science-Kompetenzen erwerben möchten. Die Heterogenität der Vorkenntnisse und auch der Anforderungen im Domain-Master impliziert sehr differenzierte Betrachtungen der notwendigen bzw. sinnvollen Ausbildungsinhalte, die in geeigneten Modulen abgebildet werden müssen.

In der Konsequenz erfordert diese Personengruppe eine komplexe Aufschlüsselung von Inhalten entsprechend der jeweiligen Vorbildung. Die Arbeitsgruppe sieht es als notwendig an, diese ggf. in Vorkursen anzugleichen, sodass ein gleichwertiger Einstieg in Masterprogramme möglich wird. In der Übersichtstabelle (Abbildung 4 auf Seite 34) beinhaltet dies den gesamten Bereich „Grundlagen Mathematik“, den Punkt „Programmierung“ im Bereich „Grundlagen Informatik“ sowie den Aspekt „ML Languages“ im Bereich „Maschinelles Lernen: Sprachen und Werkzeuge“. Gleichzeitig bedeutet dies, dass – im Gegensatz zu den Mathematik-/Statistik-/Informatik-B.Sc.-Studierenden – ein Gutteil der in der Tabelle (Abbildung 4) bei Persona A verankerten Inhalte

erst im Masterprogramm eingeführt werden muss. Zudem werden die Inhalte bei Persona B.1 (Master of Data Science in der Domäne) und Persona B.2 (Data-Science-Kompetenzen bei Schwerpunktausbildung innerhalb der Domäne) jeweils unterschiedlich gewichtet (auch je nach Pflicht- bzw. Wahlpflichtbereich und im Sinne der verschiedenen Kompetenzlevel), um allein dem zeitökonomischen Faktor Rechnung zu tragen.

Besondere Herausforderungen bestehen bezüglich der Zulassungsmodalitäten. Masterstudiengänge in Data Science müssen explizit für Nicht-Mathematiker/innen, Nicht-Statistiker/innen und Nicht-Informatiker/innen geöffnet werden, wobei die konkrete Ausgestaltung der notwendigen, je nach akademischer Herkunft sehr heterogenen Vorbereitungskurse besondere Betreuung durch Studienkoordinierende erfordert.

Im Vergleich der beiden Personas B.1 und B.2 lässt sich die Persona B.2 voraussichtlich schwieriger umsetzen, da die notwendigen Module stärker gestrafft und auf die Domäne ausgerichtet sein müssen. Dies erfordert spezielle Strukturen, die nicht ohne Weiteres aus dem Data-Science-MA-Programm übernommen, sondern speziell adaptiert werden müssen. Dafür wird in der Anfangszeit ein erhöhter Personaleinsatz von Seiten der Data-Science-orientierten Fakultäten verlangt, ggf. unterstützt durch externe Lehraufträge. Nach der Etablierungsphase wird es möglich sein, auch aus der Domänenwissenschaft heraus geeignetes Lehrpersonal zu rekrutieren und auszubilden. Gleichzeitig erfordert die Etablierung von Persona B.2 die aktive Mitwirkung von Curricularkommissionen der Fachgesellschaften sowie der Studienausschüsse an den Fakultäten selbst, um den Mehraufwand durch geeignete Reduktion an anderer Stelle in Grenzen zu halten. Da in Master-Programmen in der Regel große Teile dem Wahlpflichtbereich zuzuordnen sind, sollte dies flexibel möglich sein.

Persona B.1 hingegen lässt sich strukturell leichter realisieren, wenn auf Master-Module für Persona A zurückgegriffen werden kann. Die Herausforderung besteht vielmehr auf Seiten der Domänenwissenschaft, die kompakte Kurse bzw. Module bereitstellen muss, die datenwissenschaftlichen Bezug aufweisen. Auch wenn dies auf Basis aktueller Fragestellungen sicher möglich ist und weiter vertieft werden

¹² Die Arbeitsgruppe sieht einen Data-Science-Bachelor mit anschließendem Master in einem Domänen-Bereich als nicht zielführend an, da die fachspezifischen Grundkenntnisse nur im Bachelorstudiengang erworben werden können; zudem würde dies auch mit den typischen Zugangsvoraussetzungen kollidieren.

kann, bietet sich insbesondere die gemeinsame Projekt- bzw. Praktikumsarbeit im Team an, sofern die Personas B.1 und B.2 faktisch wegen ihrer unterschiedlichen Studienschwerpunkte zusammengebracht werden können. Dies kann somit gleichzeitig auch ein sinnvolles Modell für die Ausbildung von Persona B.2 darstellen.

3.2.1 PERSONA B.1: M.SC. DATA SCIENCE (IN DOMÄNE)

Hierbei handelt es sich um Bachelor-Absolventinnen und -Absolventen in einer Domain-Wissenschaft (Natur-/Ingenieur- oder Sozial-/Geisteswissenschaften, d.h. mit sehr unterschiedlicher Vorbildung), die auf Basis grundlegender Kenntnisse in Datenwissenschaften ein Masterprogramm in Data Science absolvieren möchten:

- **Abschluss/Studiengang:** M.Sc. (Domain) Data Science (mit Schwerpunkt in den Natur- und Lebenswissenschaften, z.B. in Physik, Chemie, Biologie, aber auch Sozial- und Geisteswissenschaften, z.B. Rechtswissenschaft, Literaturwissenschaft, Psychologie, Journalistik etc.)
- **Vorqualifikation/Technologie- und Datenkompetenz:** Selbst in der Gruppe der Bachelor-Absolventinnen und -Absolventen aus den Natur- und Ingenieurwissenschaften bestehen große Unterschiede, gravierender ist jedoch die unzureichende Grundausbildung in mathematisch-/informatischen Techniken innerhalb der Geistes- und Kulturwissenschaften. Diese Voraussetzungen müssen im Bachelor-Bereich bzw. in Vorkursen entsprechend ausgeglichen werden. Eine große Rolle hierbei spielen „Studium Generale“-basierte Qualifikationskurse in „Data Literacy“, deren Einrichtung von verschiedenen Hochschulen angestrebt wird. Für erstere Gruppe kann häufig ein Defizit in Statistik und Informatik identifiziert werden, wohingegen die Grundlagen der Mathematik in der Regel bekannt sind bzw. gezielt vertieft werden können.
- **Ausbildung:** Universitäten und Hochschulen der angewandten Wissenschaften
- **Berufsbild/Erwartung Arbeitsmarkt:** Die beiden wichtigsten Berufsbilder sind hier die/der Data Engineer und Data Scientist für die Industrie und die Wissenschaftlerin/der Wissenschaftler für Forschungsprojekte/Promotion etc., zunehmend auch die/der Data Steward. Weiterhin gibt es auch hier Berufsbilder im Management oder als Business Analyst mit Erfahrung und Kenntnissen im Umgang mit Daten, Data Mining oder Machine Learning. Zunehmend wichtig werden Spezialisten im Bereich des (Forschungs-) Datenmanagements in forschungsintensiven Branchen, die eine Brückenfunktion zwischen „Datenerzeugern“ (aus

experimentellen bzw. empirischen Quellen) und „Datenverwertern“ (z.B. Modellierung) ausfüllen.

- **Fokus:** Im Vordergrund stehen daher gleichberechtigt die Anwendung in der Industrie – darunter auch Consulting – und die Wissenschaft/Forschung. Aber auch in anderen Bereichen wie Ethik und Politik oder in benachbarten Gebieten wie IT-Sicherheit werden Absolventinnen und Absolventen mit Data-Science-Qualifikationen dringend benötigt.
- **Interesse/Motivation:** Aus der obigen Zielrichtung ergibt sich einerseits eine Motivation, die auf eine wissenschaftliche Karriere, ggf. mit Promotion, ausgerichtet ist. Andererseits werden Absolventinnen und Absolventen in der Industrie immer mit verschiedenen Domänen in Berührung kommen und dort das gelernte Wissen anwenden, etwa in der Pharmaindustrie, Chemie, Handel/Marketing, Materialtechnik, Medizintechnik, Finanzwirtschaft/E-Commerce, Transport/Logistik etc.

3.2.2 PERSONA B.2: M.SC. IN DOMÄNE MIT DATA-SCIENCE-KOMPETENZEN

Hierbei handelt es sich um analog vorgebildete Bachelor-Absolventinnen und -Absolventen, die im Master innerhalb der Domäne verbleiben, aber im Rahmen des Masters vertiefte Data-Science-Kompetenzen erwerben möchten.

- **Abschluss/Studiengang:** M.Sc. Domänenwissenschaft (z.B. Physik, Chemie, Literaturwissenschaft, Journalistik, Jura etc.)
- **Vorqualifikation/Technologie- und Datenkompetenz:** Wie bei Persona B.1 bestehen selbst in der Gruppe der Bachelor-Absolventinnen und -Absolventen aus den Natur- und Ingenieurwissenschaften große Unterschiede, verschärft wird das Problem noch durch die oft unzureichende technische/formale Grundausbildung in den Geistes- und Kulturwissenschaften. Diese Voraussetzungen müssen im Bachelor-Bereich entsprechend ausgeglichen werden (s. bei Persona B.1). Gleiches gilt für die heterogene mathematische Ausbildung mit häufig anzutreffenden Defiziten im Bereich Statistik.
- **Ausbildung:** Universitäten und Hochschulen der angewandten Wissenschaften
- **Berufsbild/Erwartung Arbeitsmarkt:** Hier steht die erworbene Zusatzqualifikation im Vordergrund, sodass die üblichen Berufsbilder der in den Domänen Forschenden weitgehend erhalten bleiben, aber für die Absolventinnen und Absolventen neue Perspektiven entsprechend aktueller Anforderungen eröffnet werden, sowohl im Promotions-

als auch im Unternehmensbereich. Besonders Fähigkeiten im Umgang mit großen Datenmengen und grundlegende Kenntnisse im Machine Learning werden zunehmend eine große Rolle spielen, ebenso das Profil des Data Stewards.

- *Fokus:* Sowohl in der Industrie als auch im Bereich Wissenschaft werden entsprechend weiterqualifizierte Personen eine wichtige Rolle als Vermittler zwischen Informatik/Mathematik/Statistik und der Anwendung in der Domäne spielen. Der Fokus des Profils liegt somit insbesondere in der Teamarbeit. Gleichzeitig wird diese Personengruppe eine wichtige Rolle im Bereich der Translation moderner Methoden der Data Science in Grundlagen- und Anwendungsprojekte spielen und somit als Innovationsmotor wirken.
- *Interesse/Motivation:* Aus der obigen Zielrichtung ergibt sich eine Motivation, die auf eine wissenschaftliche Karriere, ggf. mit Promotion, oder eine Karriere im Unternehmen ausgerichtet ist. Grundsätzlich dominiert aber die intrinsische Motivation der Domäne, sodass Absolventinnen und Absolventen innerhalb der Domain-Peergroup weiterhin als gleichqualifiziert und -berechtigt angesehen werden.

3.3 PERSONA C: WEITERBILDUNG ZUM DATA SCIENTIST

Persona C fokussiert auf das Thema Weiterbildung im Kontext von Data Science. Schon jetzt zeichnet sich ab, dass der Bedarf an Data Scientists nicht alleine durch die Hochschulbildung von Studierenden gedeckt werden kann, sondern auch dedizierte Weiterbildungsprogramme aufgesetzt werden müssen, um Mitarbeitende von Unternehmen und Arbeitssuchende entsprechend qualifizieren zu können.

Hierbei kann man zwei Ausbildungsziele unterscheiden: Persona C.1 hat zum Ziel, grundlegende Basiskenntnisse zu vermitteln, die jeder Data Scientist mitbringen sollte. Persona C.2 baut auf dem Basic Data Scientist auf und hat zum Ziel, Kompetenzen in der Breite zu vertiefen und damit einem Hochschulstudiengang M.Sc. Data Science näher zu kommen.

3.3.1 PERSONA C.1: BASIC DATA SCIENTIST

Das Zielbild einer Persona für die Vermittlung grundlegender Kompetenzen im Weiterbildungsbereich wurde wie folgt entworfen:

- *Abschluss/Studiengang:* Basic Data Scientist (Weiterbildung)
- *Vorqualifikation/Technologie- und Datenkompetenz:* Als Vorqualifikation gibt es keine Einschränkung auf bestimm-

te Abschlüsse oder Studiengänge, sofern gewisse informatische Kenntnisse (wie Programmierkenntnisse, Umgang mit Skriptsprachen und Datenbankmanagementsystemen) und mathematische Kenntnisse (wie Stochastik, Kombinatorik, Statistik, Optimierung und Numerik) nachgewiesen werden können. Für die Erlangung dieser Vorqualifikation bieten sich z.B. Vorkurse oder Einstiegssemester mit dezidierten Modulen zu Grundlagen der Informatik und Mathematik an.

- *Ausbildung:* Universitäten und Hochschulen der angewandten Wissenschaften (z.B. als Fern- oder Ergänzungsstudium), Forschungseinrichtungen oder private Weiterbildungsanbieter
- *Berufsbild/Erwartung Arbeitsmarkt:* Die angestrebten Berufsbilder in der Industrie sind Data Scientist, Data Engineer, Data Manager, Data Analyst oder Daten-Strategie/in. Die Erwartungshaltung an Absolventinnen und Absolventen wäre die Rolle eines Berufsanfängers, der selbständig kleinere Anwendungsaufgaben lösen kann, aber von erfahrenen Mitarbeitern und Mitarbeiterinnen angeleitet werden muss.
- *Fokus:* Im Fokus steht die Anwendung in der Industrie – darunter auch Consulting. Aber auch in anderen Gebieten wie Ethik und Politik oder in angrenzenden Bereichen wie IT-Sicherheit werden datenwissenschaftlich qualifizierte Absolventinnen und Absolventen benötigt. Im Wesentlichen können sie Techniken, Methoden und Werkzeuge in kleinen, überschaubaren Anwendungskontexten verantwortlich einsetzen und verfügen über das dazu notwendige Grundlagenwissen. Ausgehend von diesen Basiskompetenzen kann dann eine Spezialisierung, z.B. im Bereich Deep Learning oder Machine Learning, erfolgen.
- *Interesse/Motivation:* Die Motivation zur Basic-Data-Science-Ausbildung liegt darin, die für das aktuelle oder zukünftige Betätigungsfeld notwendigen theoretischen Grundkompetenzen zu erlangen und sie in der Praxis in kleinen, überschaubaren Anwendungskontexten einsetzen zu können. Mit dieser Qualifikation kann das Einsatzfeld von Mitarbeiterinnen und Mitarbeitern im Unternehmen verbreitert werden. Entsprechend verbessern Arbeitssuchende ihre Chancen.

3.3.2 PERSONA C.2: ADVANCED DATA SCIENTIST

Das Zielbild einer Persona für die Vermittlung tiefergehender Data-Science-Kompetenzen im Weiterbildungsbereich stellt sich wie folgt dar:

- *Abschluss/Studiengang:* Advanced Data Scientist (Weiterbildung)
- *Vorqualifikation/Technologie- und Datenkompetenz:* Als Vorqualifikation für den Advanced Data Scientist gilt eine erfolgreiche Absolvierung der Basic-Data Scientist-Ausbildung. Die praktische Anwendung der im Rahmen der Basic-Ausbildung vermittelten Techniken, Methoden und Werkzeuge auf Basis von realen Daten ist dabei wünschenswert.
- *Ausbildung:* Universitäten und Hochschulen der angewandten Wissenschaften (z.B. als Fern- oder Ergänzungsstudium), Forschungseinrichtungen oder private Weiterbildungsanbieter
- *Berufsbild/Erwartung Arbeitsmarkt:* Die angestrebten Berufsbilder in der Industrie sind Data Scientist, Data Engineer, Data Manager, Data Analyst, Data Architect oder Daten-Strategie/in. Die Erwartungshaltung an Absolventinnen und Absolventen wäre die Rolle eines vollwertigen Team-Mitglieds, welches selbständig auch komplexere Anwendungsaufgaben lösen und Berufsanfänger anleiten/unterstützen kann.
- *Fokus:* Im Fokus steht die Anwendung in der Industrie – darunter auch Consulting. Aber auch in anderen Bereichen wie Ethik und Politik oder in angrenzenden Feldern wie IT-Sicherheit werden datenwissenschaftlich qualifizierte Absolventinnen und Absolventen benötigt. Im Wesentlichen umfasst dies die Verbreiterung des theoretischen Wissens und die Beherrschung von Techniken, Methoden und Werkzeugen in komplexeren Anwendungskontexten.
- *Interesse/Motivation:* Die Motivation zur Advanced-Data-Scientist-Ausbildung liegt darin, für das aktuelle oder zukünftige Betätigungsfeld vertiefte theoretische Kompetenzen in der Breite zu erlangen und sie in der Praxis in komplexeren Anwendungskontexten einsetzen zu können. Mit dieser Qualifikation kann das Einsatzfeld von Mitarbeitern und Mitarbeiterinnen im Unternehmen verbreitert werden. Entsprechend verbessern Arbeitssuchende ihre Chancen.

4. LERN- UND AUSBILDUNGSIHINHALTE FÜR DATA SCIENCE

Diese Empfehlungen beschreiben Themenfelder, in denen die Studierenden und Beschäftigten im Verlauf ihres Studiums bzw. der Weiterbildung Kompetenzen erwerben sollen. Wichtig ist die Verzahnung von Kompetenzen aus den Fächern Informatik und Mathematik / Statistik in den Studiengängen. Mittels der Kompetenzbeschreibungen können gleichzeitig die curricularen Spezifika des jeweiligen Studiengangs bzw. des Weiterbildungsangebots kompetenzorientiert charakterisiert werden. Kompetenzen werden hier in Anlehnung an Weinert als „erlernbare kognitive Fähigkeiten und Fertigkeiten“ verstanden, die ein Individuum in einem Handlungskontext zur Problemlösung befähigen, einschließlich der dazu erforderlichen motivationalen, volitionalen und sozialen Handlungsdispositionen und Fähigkeiten [Weinert2001]. Kognitive und nicht-kognitive Facetten einer Kompetenz sind stets eng miteinander verbunden und können in fachbezogenen Handlungskontexten auch gemeinsam erworben werden.

Zur Beschreibung von Kompetenzfeldern orientieren sich diese Ausführungen an der sogenannten Anderson-Krathwohl-Taxonomie (AKT-Matrix) [Anderson2001] – in Anlehnung an die von der Gesellschaft für Informatik herausgegebenen Empfehlungen für Bachelor- und Masterprogramme im Studienfach Informatik an Hochschulen [GI2016].

Die etwas abgewandelte AKT-Matrix mit 4 Stufen und der Unterscheidung zwischen geringer und starker Kontextualisierung, die die [GI2016] für ihre curricularen Empfehlungen für Informatik-Studiengänge eingeführt und genutzt hat, wird hier zur Beschreibung der Kompetenzfelder vereinfacht, indem die Stufen 3 und 4 zusammengefasst werden und keine Unterscheidung zwischen geringer und starker Komplexität und Kontextualisierung gemacht wird. Die Stufen 3 (Analyse-

ren) und 4 (Erzeugen) wurden ebenfalls zusammengefasst, da der Fokus bei Data Science in der Analyse liegt und die Systemerzeugung in diesem Bereich keine relevante kognitive Leistung darstellt. Die verwendeten Stufen sind in Abbildung 3 dargestellt. Komplexität und Kontextualisierung als charakteristische Eigenschaften von Kompetenzfeldern ergeben sich jeweils aus der Natur der Kompetenzfelder oder aus der Verwendung, sodass sie nicht explizit herausgestellt werden müssen.

Die Kompetenzgruppen-Tabelle im Anhang ist im Rahmen von kollaborativen Workshops und Austauschformaten entstanden. Diese Kompetenzfelder wurden mit den drei definierten Personas in Kapitel 3 abgeglichen und in Bezug auf zwei Dimensionen bewertet:

- **Wichtigkeit: P=Pflichtinhalt oder W=Wahlpflichtinhalt**
Die Bedeutung der einzelnen Kompetenzfelder wurde entlang obligatorischer und optionaler Inhalte unterteilt. Dabei unterscheidet sich die Einteilung zwischen den drei Personas mitunter, da für unterschiedliche Ausbildungsziele unterschiedliche Schwerpunkte gelegt werden.
- **Kompetenzniveau: L1=Verstehen, L2=Anwenden, L3=Analysieren** (siehe Abbildung 3)
Das zu erreichende Kompetenzniveau entlang der Anderson-Krathwohl-Taxonomie gibt Auskunft darüber, welches Kompetenzniveau die jeweilige Persona im jeweiligen Kompetenzfeld erreichen sollte.

In Abbildung 4 werden den 14 zu vermittelnden Kompetenzfeldern jeweils konkrete Lerninhalte zugeschrieben. (Eine ausführliche Darstellung der Kompetenzfelder und der Lerninhalte, insbesondere hinsichtlich der Gewichtung für die unterschiedlichen Personas, erfolgt im Anhang.)

STUFE 1 (L1) VERSTEHEN	STUFE 2 (L2) ANWENDEN	STUFE 3 (L3) ANALYSIEREN
Lernende können Sachverhalte erklären, Beispiele anführen, Aufgabenstellungen interpretieren oder ein Problem in eigenen Worten wiedergeben.	Lernende können einen Arbeitsablauf, ein Verfahren oder eine Prozedur anwenden oder ausführen, ohne die Details des Verfahrens im Einzelnen kennen zu müssen.	Lernende können ein Problem in einzelne Teile zerlegen und so die Struktur des Problems verstehen; sie können Widersprüche aufdecken, Zusammenhänge erkennen und Folgerungen ableiten sowie zwischen Fakten und Interpretationen unterscheiden.

Abbildung 3: Vereinfachte Darstellung der AKT-Matrix mit den kognitiven Prozessdimensionen [Anderson2001]

KOMPETENZFELDER	PFLICHT-/WAHLINHALTE
(1) Grundlagen Mathematik & Statistik	Algebra, Stochastik, Analysis & Statistik (Deskriptive Statistik, Statistische Inferenz, Lineare Modelle, Simulation und Resampling)
(2) Fortgeschrittene Mathematik & Statistik	Fortgeschrittene Algebra, Stochastik, Analysis & Statistik (Nichtparametrische, Multivariate, Hochdimensionale, Bayesianische, Räumliche)
(3) Grundlagen der Informatik	Programmierung, Datenbanken, Algorithmen
(4) Fortgeschrittene Informatik	Virtualisierung, Container, NoSQL, Streaming, Cloud
(5) Kryptographie und Sicherheit	Daten- und Cybersicherheit, Kryptographie, Blockchain
(6) Datenethik und Data Privacy	Datenethik, DSGVO, Data Privacy & Compliance
(7) Data Governance	Data Policy, Strukturen, Metadatenmanagement
(8) Datenintegration	Instrumentation, Logging, Sensoren, verschiedene Datenquellen
(a) Datensammlung	
(b) Datenvorbereitung	Datenqualität, Labeling, Aggregation, Metrics, Segmentation, Feature Selection
(c) Datenpipelines	Data Flow, Infrastruktur und Tools, ETL
(9) Datenvisualisierung	Perception Theory, Colour, Editorial Thinking, Principles, Representations, Tools, Interactivity, Annotations
(10) Data Mining	KDD Prozess, IR Methods, Reporting, Text-, Web-, Process Mining, Time Series
(11) Maschinelles Lernen/Deep Learning	Languages, Libraries, Tools, Analytics Frameworks, Konzepte, Klassische Inferenz Methoden, Probabilistische Modelle
(a) Sprachen und grundlegende ML-Methoden	
(b) Deep Learning	Neural Networks, Reinforcement Learning
(12) Business Intelligence	Phases, Data Warehouse, ERP, ETL
(13) Domänenspezifische Anwendungen	Praktische Anwendungen und Tools für verschiedene Domänen, Research Data Management
(14) Data Science in der Organisation	Project Management, Communication und Soft Skills, Data Economy

Abbildung 4: Die vom Arbeitskreis identifizierten Kompetenzfelder und die entsprechenden Lerninhalte

Die folgenden fünf Tabellen fassen die Wichtigkeit (Pflicht/Wahl) und die Lernziele (L1, L2, L3) der Kompetenzfelder in

Bezug auf die Personas in Kapitel 3 und die ausführlichen Kompetenzfelder im Anhang zusammen.

4.1 PERSONA A: MASTER OF DATA SCIENCE (M.SC.)

Der Data-Science-Master für Bachelor-Absolventinnen und -Absolventen der Informatik, Mathematik oder mit

adäquaten Kenntnissen sollte folgende Inhalte und Lernziele umfassen:

KOMPETENZFELDER	PFLICHT-/WAHLINHALTE	LERNZIEL
(1) Grundlagen Mathematik & Statistik	Voraussetzung	-
(2) Fortgeschrittene Mathematik & Statistik	Pflicht	L3
(3) Grundlagen der Informatik	Pflicht (Voraussetzung: Programmierung)	L2
(4) Fortgeschrittene Informatik	Pflicht	L2
(5) Kryptographie und Sicherheit	Wahl	L1
(6) Datenethik und Data Privacy	Pflicht	L3
(7) Data Governance	Pflicht	L1
(8) Datenintegration	Pflicht	L2-L3
(9) Datenvisualisierung	Pflicht	L2
(10) Data Mining	Pflicht	L2
(11) Maschinelles Lernen/Deep Learning	Pflicht	L3
(12) Business Intelligence	Wahl	L2
(13) Domänenspezifische Anwendungen	Pflicht	L1
(14) Data Science in der Organisation	Pflicht	L2

Abbildung 5: Die Kompetenzfelder des MA Data Science mit einer Einschätzung der Wichtigkeit und den zu erreichenden Kompetenzniveaus

Zwischenfazit: Studierende des Masters Data Science bringen genug Vorkenntnisse mit und haben ca. vier Semester Zeit, die Kernelemente des Data Science mit einer bedeutenden Anzahl von Pflichtfächern eingehend zu vertiefen. Einzig die Bereiche Kryptographie und Sicherheit und Business Intel-

ligence werden als Wahlpflichtteile angesehen. Insbesondere in den Kompetenzfeldern Mathematik, Datenethik und Data Privacy und Maschinelles Lernen werden vertiefende Analysekompetenzen als erforderlich angesehen.

Ausgewählte Beispiele:

- **Data-Science-Master an der Beuth Hochschule:** Der Masterstudiengang Data Science an der Beuth Hochschule Berlin¹³ vermittelt exzellente Expertise in den Bereichen Machine Learning/Big Data Analytics, um diese praktisch im Unternehmen umsetzen zu können. Anwender für Data Science sind Unternehmen mit intelligenten Systemen/Maschinen, die große Datenströme verarbeiten, um Vorhersagen zu treffen. Dazu gehören Partnerunternehmen im Bereich Logistik, Marktforschung, Retail, Gesundheit oder auch Plattformbetreiber für Maschinelle Intelligenz. Der Studiengang bildet das gesamte Spektrum für Datenprodukte ab, u. a. Modellbildung, Datenbereinigung, Praxiserfahrung mit vielfältigen Datensätzen sowie Testen produktionsreifer Lösungen.
- **Master Data Science und Künstliche Intelligenz an der Universität des Saarlandes:** Die Studierenden des Masters arbeiten an konkreten Fragestellungen in Fächern

wie Computerlinguistik, Physik, Materialwissenschaften, Chemie, Psychologie und Biologie. Durch den Umgang mit sensiblen Daten sind auch Aspekte der IT-Sicherheit, der Rechtswissenschaften, des Datenschutzes, der Philosophie und der Ethik wichtige Studieninhalte. Ein Industriepraktikum und ein Masterpraktikum in einer Forschungsgruppe können in den Wahlpflichtbereich eingebracht werden. Neben den Professorinnen und Professoren der Fachrichtungen Informatik sowie Sprachwissenschaft und Sprachtechnologie der Universität des Saarlandes lehren im Masterstudiengang Wissenschaftlerinnen und Wissenschaftler der renommierten Forschungsinstitute auf dem Campus, die hier anwendungsnah die Verfahren der Gebiete Data Science, Künstliche Intelligenz, Maschinelles Lernen und Big Data erforschen und die Studierenden umfassend auf die Herausforderungen der digitalen Zukunft vorbereiten.¹⁴

¹³ <https://www.beuth-hochschule.de/m-ds> (08.November 2019).

¹⁴ <https://www.uni-saarland.de/master/studienangebot/mathinf/data-science/info.html> (08.November 2019).

4.2 PERSONA B.1: MASTER OF DATA SCIENCE (M.SC.) (IN DER DOMÄNE)

Der Data-Science-Master für Bachelor-Absolventinnen und

-Absolventen aus informatikfernen Studiengängen mit geringen Data-Science-Kompetenzen sollte folgenden Inhalte und Lernziele umfassen:

KOMPETENZFELDER	PFLICHT-/WAHLINHALTE	LERNZIEL
(1) Grundlagen Mathematik & Statistik	Voraussetzung	-
(2) Fortgeschrittene Mathematik & Statistik	Wahl	L3
(3) Grundlagen der Informatik	Wahl	L1
(4) Fortgeschrittene Informatik	-	-
(5) Kryptographie und Sicherheit	-	-
(6) Datenethik und Data Privacy	Pflicht	L2
(7) Data Governance	Pflicht	L1
(8) Datenintegration	Pflicht	L2-L3
(9) Datenvisualisierung	Pflicht	L1
(10) Data Mining	Pflicht	L2
(11) Maschinelles Lernen/Deep Learning	Wahl/Pflicht	L1-L3
(12) Business Intelligence	-	-
(13) Domänenspezifische Anwendungen	Pflicht	L3
(14) Data Science in der Organisation	Wahl	L2

Abbildung 6: Die Kompetenzfelder des M.Sc. Data Science in der Domäne mit Stellung im Studiengang (Wahl oder Pflicht) und den zu erreichenden Kompetenzniveaus

4.3 PERSONA B.2: MASTER IN DER DOMÄNE (M.SC.) MIT DATA-SCIENCE-KOMPETENZEN

Der Master für Bachelor-Absolventinnen und -Absolventen

außerhalb der Informatik und Mathematik / Statistik in Anwendungsdomänen mit geringen Data-Science-Kompetenzen sollte folgende Inhalte und Lernziele umfassen:

KOMPETENZFELDER	PFLICHT-/WAHLINHALTE	LERNZIEL
(1) Grundlagen Mathematik & Statistik	Voraussetzung	-
(2) Fortgeschrittene Mathematik & Statistik	-	-
(3) Grundlagen der Informatik	Wahl	L1
(4) Fortgeschrittene Informatik	-	-
(5) Kryptographie und Sicherheit	-	-
(6) Datenethik und Data Privacy	Wahl	L2
(7) Data Governance	Wahl	L1
(8) Datenintegration	Wahl/Pflicht	L1 - L2
(9) Datenvisualisierung	Wahl	L1
(10) Data Mining	Wahl	L1 - L2
(11) Maschinelles Lernen/Deep Learning	Wahl/Pflicht	L1 - L2
(12) Business Intelligence	-	-
(13) Domänenspezifische Anwendungen	Pflicht	L3
(14) Data Science in der Organisation	-	-

Abbildung 7: Die Kompetenzfelder des Master of Data Science (M.Sc.) in der Domäne mit Stellung im Studiengang (Wahl oder Pflicht) und den zu erreichenden Kompetenzniveaus

Zwischenfazit: Sowohl beim Master of Data Science (in der Domäne) (B.1) als auch beim Domänen-Master mit Data-Science-Kompetenzen (B.2) muss ein stärkerer Fokus auf Kompetenzen im Umgang mit Daten als auf fortgeschrittene mathematisch-informatische Grundlagen gelegt werden: Data Governance, Integration, Visualization und Mining sowie insbesondere Aspekte des Maschinellen Lernens stehen hierbei als anwendungsbezogene Bereiche naturgemäß im Vordergrund. Die Unterschiede zwischen den Personas B.1 und B.2 sind insbesondere dem verschiedenen Umfang an Lehrveranstaltungen geschuldet, da bei B.2 nach wie vor die

Ausbildung in der Domäne selbst den Vorrang hat. Somit entfällt z.B. bei Persona B.2 der Bereich der fortgeschrittenen Mathematik (2), und viele Gebiete mit Pflichtcharakter für B.1 werden bei B.2 dem Wahlpflichtbereich zugeordnet. Gleichzeitig wurden die Lernziele im Verhältnis zueinander unterschiedlich gewichtet, um den zeitlichen Lernaufwand auszubalancieren. Aus demselben Grund wurden in den Bereichen (8)-(11) einzelne untergeordnete Inhalte zusammengefasst bzw. als Auswahlmöglichkeiten deklariert, um die besondere Fokussierung auf spezielle Anforderungen der Domäne herauszuheben.

Ausgewählte Beispiele:

- **Leuphana Universität Lüneburg:** Das internationale Programm „Management & Data Science“ ist für Studierende aus allen (insbesondere auch informatikfernen) Disziplinen offen und beinhaltet einen Data-Science-Master für Persona B.1.
- **TU Dortmund:** Erste Ansätze zur Etablierung der Persona B.2 werden derzeit durch entstehende neue Modul-Angebote im Bereich Datenwissenschaften und deskriptive Statistik für Studierende der Chemie und Chemischen Biologie in enger Abstimmung mit der Fakultät für Statistik realisiert.

4.4 PERSONA C.1: BASIC DATA SCIENTIST

Für die Weiterbildung zum Basic Data Scientist sollten folgende Inhalte vermittelt und Lernziele erreicht werden:

KOMPETENZFELDER	PFLICHT-/WAHLINHALTE	LERNZIEL
(1) Grundlagen Mathematik & Statistik	Voraussetzung	L2
(2) Fortgeschrittene Mathematik & Statistik	-	-
(3) Grundlagen der Informatik	Voraussetzung	L1 - L2
(4) Fortgeschrittene Informatik	Pflicht	L1 - L2
(5) Kryptographie und Sicherheit	Wahl	L1
(6) Datenethik und Data Privacy	Pflicht	L1 - L3
(7) Data Governance	Pflicht	L1
(8) Datenintegration	Pflicht	L1 - L2
(9) Datenvisualisierung	Pflicht	L2
(10) Data Mining	-	-
(11) Maschinelles Lernen/Deep Learning	Wahl/Pflicht	L1 - L2
(12) Business Intelligence	Wahl/Pflicht	L1
(13) Domänenspezifische Anwendungen	-	-
(14) Data Science in der Organisation	Nur Projektmanagement (Pflicht)	L1

Abbildung 8: Die Kompetenzfelder eines Basic Data Scientist mit Stellung im Studiengang (Wahl oder Pflicht) und den zu erreichenden Kompetenzniveaus

Zwischenfazit: Teilnehmende an einem Basic-Data-Scientist-Weiterbildungsprogramm müssen über Grundlagen der Mathematik und Informatik verfügen. In den Bereichen Fortgeschrittene Informatik, Datenethik und Data Privacy, Data Governance, Datenintegration, Datenvisualisierung, Data Mining sowie Projektmanagement müssen Kompetenzen mindestens auf der Ebene Verstehen (L1) aufgebaut werden. Teilweise werden Kompetenzen auf Anwendungsebene (L2) gefordert – im Bereich Datenethik und Data Privacy sogar auf der Ebene Analysieren (L3). Kompetenzen in den Bereichen Maschinelles Lernen/Deep Learning sowie Business Intelligence werden nur auszugsweise auf den Ebenen Verstehen (L1) und Anwenden (L2) erworben.

Ausgewählte Beispiele:

- **Fraunhofer Data Scientist Basic Level**¹⁵: Durch die zertifizierte Schulung der Fraunhofer Gesellschaft soll

breitgefächertes Wissen vermittelt werden, um effizient in Data-Science-Teams mitarbeiten zu können. Die Schulungsinhalte decken die obigen Kompetenzfelder ab. Die Schulung ist auf 5 Tage begrenzt und fokussiert daher in der Breite hauptsächlich auf dem Lernziel Verstehen (L1) und teilweise Anwenden (L2).

- **Data-Science-Zertifikatsprogramm Hochschule Albstadt-Sigmaringen/Universität Mannheim**¹⁶: Über das Zertifikatsprogramm können die Einzelzertifikate CAS (Certificate of Advanced Studies) in den Bereichen Data Science Programmer, Data Miner, Data Engineer und Business Analyst und DAS (Diploma of Advanced Studies) in den Bereichen Data Scientist, Big Data Architect, Data Analyst und Management Scientist erworben werden, welche die Kompetenzfelder des Basic Data Scientist abdecken.

¹⁵ <https://www.bigdata.fraunhofer.de/de/datascientist/zertifizierungen/zertifizierung.html> (12. November 2019).

¹⁶ <https://www.hs-albsig.de/studienangebot/masterstudiengaenge/data-science/zertifikatsprogramm-data-science> (12. November 2019).

4.5 PERSONA C.2: ADVANCED DATA SCIENTIST

Der Advanced Data Scientist baut auf dem Basic Data Scientist auf. Das heißt, die dort vermittelten Inhalte stellen die Voraussetzung für die Teilnahme am Advanced-Data-Scientist-Weiterbildungsprogramm dar. Diese werden im

Folgenden nicht noch einmal gesondert als Voraussetzung aufgeführt.

Für die Weiterbildung zum Advanced Data Scientist sollten folgende Inhalte vermittelt und Lernziele erreicht werden:

KOMPETENZFELDER	PFLICHT-/WAHLINHALTE	LERNZIEL
(1) Grundlagen Mathematik & Statistik	-	
(2) Fortgeschrittene Mathematik & Statistik	-	
(3) Grundlagen der Informatik	Pflicht	L2
(4) Fortgeschrittene Informatik	Wahl/Pflicht	L2
(5) Kryptographie und Sicherheit	Wahl	L1
(6) Datenethik und Data Privacy	Pflicht	L2 - L3
(7) Data Governance	Nur MetaDatenmanagement (Pflicht)	L2
(8) Datenintegration	Pflicht	L2
(9) Datenvisualisierung	Pflicht	L1 - L2
(10) Data Mining	Pflicht/Wahl	L2
(11) Maschinelles Lernen/Deep Learning	Pflicht	L3
(12) Business Intelligence	Pflicht	L2
(13) Domänenspezifische Anwendungen	-	
(14) Data Science in der Organisation	Nur Projektmanagement (Pflicht)	L2

Abbildung 9: Die Kompetenzfelder eines Advanced Data Scientist mit Stellung im Studiengang (Wahl oder Pflicht) und den zu erreichenden Kompetenzniveaus

Zwischenfazit: Teilnehmende an einem Advanced-Data-Scientist- Weiterbildungsprogramm müssen entweder über einen Abschluss als Basic Data Scientist verfügen oder äquivalente Kompetenzen erworben haben. In den Bereichen Grundlagen der Informatik, Datenethik und Data Privacy, Metadatenmanagement, Datenintegration, Business Intelligence sowie Projektmanagement müssen Kompetenzen mindestens auf der Ebene Anwenden (L2) aufgebaut werden. Im Bereich

Datenvisualisierung werden Kompetenzen auf den Ebenen Verstehen (L1) und Anwenden (L2) verlangt, für Datenethik und Data Privacy auf den Ebenen Anwenden (L2) und Analysieren (L3) und für Maschinelles Lernen/Deep Learning auf der Ebene Analysieren (L3). Kompetenzen in den Bereichen Fortgeschrittene Informatik sowie Data Mining müssen auszugswise auf der Ebene Anwenden (L2) erworben werden.

Ausgewählte Beispiele:

- **Fraunhofer Data Scientist Advanced Level und Senior Data Scientist¹⁷:** Die weiterführende Schulung der Fraunhofer Gesellschaft zum Advanced Level setzt den Fraunhofer Data Scientist Basic Level sowie eine ausgewählte Vertiefung (z.B. Data Analyst, Data Manager oder Machine Learning Specialist) und darüber hinaus Berufserfahrung voraus. Das Ziel liegt in der praktischen Anwendung von Verfahren und der Vertiefung des Wissens. Darauf aufbauend erfolgt die Ausbildung zu Senior Data Scientists, welche innerhalb eines Unternehmens das Themenfeld voranbringen und Data-Science-Projekte leiten sollen. Die Schulungsreihe deckt die obigen Kompetenzfelder ab und fokussiert auf dem Lernziel Anwenden (L2) und teilweise Analysieren (L3).
- **Data-Science-Master Hochschule Albstadt-Sigmaringen/ Universität Mannheim¹⁸:** Der Weiterbildungs-Masterstudiengang Data Science enthält die Module des Basic Levels als Voraussetzung. Der Masterstudiengang vermittelt theoretische Kenntnisse in den Kompetenzfeldern des Advanced Data Scientists auf den Ebenen L2 und L3 sowie deren praktische Umsetzung.
- **Data Science Retreat¹⁹:** Das Retreat bietet datenwissenschaftliche Ausbildung und praktisches Coaching für Fachleute sowie eine Verbindung zu Unternehmen, die nach Expertinnen und Experten mit praktischer Erfahrung suchen. Das Retreat umfasst ein intensives 12-wöchiges Vollzeit-Bootcamp mit Unterricht und Projektarbeit.

¹⁷ <https://www.bigdata.fraunhofer.de/de/datascientist/zertifizierungen.html> (08. November 2019).

¹⁸ <https://www.hs-albsig.de/studienangebot/masterstudiengaenge/data-science> (08. November 2019).

¹⁹ <https://datascienceretreat.com/> (08. November 2019).

5. AUSBLICK

Die Karriere der Data-Science-Disziplin steht erst am Anfang. Mit zunehmender Datafizierung der Lebens- und Arbeitswelten steigt auch der Bedarf an Expertinnen und Experten, die in der Lage sind, sich auf wissenschaftlichem Niveau mit diesen Fragestellungen auseinanderzusetzen. Wie dieses Papier zeigt, sind die Anforderungen an Datenwissenschaftlerinnen und -wissenschaftler schon heute sehr vielfältig. Und diese Anforderungen werden – ungeachtet der Tatsache, dass viele der aktuellen Tätigkeiten eines Data Scientist künftig voraussichtlich eine weitere softwarebasierte Automatisierung erfahren werden – in Zukunft an Komplexität und Spezifizierung weiter zunehmen.

Der Bedeutungszuwachs von Data Science in den unterschiedlichen Anwendungsfeldern wird vor dem Hintergrund einer sich beschleunigenden technologischen Entwicklung zu einem weiter steigenden Bedarf an Expertise in diesem Bereich führen. Gleichzeitig werden auch die Data-Science-Grundlagendisziplinen, die Mathematik, die Statistik und die Informatik, einen weiteren Bedeutungszuwachs erfahren, weil es künftig noch viel stärker darauf ankommen wird, den systematischen und automatisierten Umgang mit den Daten, die die Welt vermessen, erklären und steuern sollen, zu entwickeln.

Der GI-Arbeitskreis „Data Science/Data Literacy“ wurde vom Vorstand der GI eingesetzt, um eine Hilfestellung für die inhaltliche Ausgestaltung von Data-Science-Studiengängen und -Weiterbildungsangeboten zu entwickeln. Diese Arbeit ist mit diesem Papier abgeschlossen und der Arbeitskreis hat seine Aufgabe erfüllt. Gleichzeitig ist aber auch klar, dass sich das Umfeld, in dem sich Datenwissenschaftlerinnen und -wissenschaftler bewegen, schnell ändert. Und obwohl der Arbeitskreis seine Aufgabe als erfüllt betrachten darf, ist dieses Papier doch nur eine erste Standortbestimmung, die künftig weitere Anpassungen, Ergänzungen und Erweiterungen erfahren wird.

Deshalb werden die Mitglieder des Arbeitskreises, die Mitwirkenden sowie die Autorinnen und Autoren dieses Papiers seitens der GI-Geschäftsstelle von Zeit zu Zeit aufgerufen werden, einen Blick auf den Status quo zu werfen und etwaige Anpassungen vorzunehmen. Die Gesellschaft für Informatik wird das Thema weiter verfolgen und aktiv werden, sollte Handlungsbedarf bestehen. Zudem wurde im Rahmen des Entwicklungsprozesses auch ein Vorgehen etabliert, das anderen Arbeiten an den Schnittstellen zur Informatik bei der Entwicklung von curricularen Empfehlungen dienlich sein kann, z.B. im Bereich Digital Design.

LITERATUR

[Anderson2001] Anderson, L.W.; Krathwohl, D. et. al. (Hrsg.) (2001): A Taxonomy for Learning, Teaching, and Assessing. A Revision of Bloom's Taxonomy of Educational Objectives, New York u.a.: Longman.

[acatech2017] Gausemeier, J.; Guggemos, M.; Kreimeyer, A. (2017): „Auswahl, Beschreibung, Bewertung und Messung der Schlüsselkompetenzen für das Technologiefeld Data Science“, acatech-Bericht, https://www.acatech.de/wp-content/uploads/2019/02/acatech_NKM_Data_Science_WEB-2.pdf. (Aufgerufen am: 08.11.2019).

[Beck2019] Beck, S. (et al.): „Künstliche Intelligenz und Diskriminierung. Herausforderungen und Lösungsansätze. Whitepaper der Plattform Lernende Systeme“ https://www.plattform-lernende-systeme.de/files/Downloads/Publicationen/AG3_Whitepaper_250619.pdf. (4. Dezember 2019).

[Cleveland2001] Cleveland, W.S. (2001): Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics, International Statistical Review, 69, 1, 21-26.

[Conway2013] Conway, D (2013): „The Data Science Venn Diagram“, <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>. (Aufgerufen am: 08.11.2019).

[DAS2014] Data Assessment Solutions GmbH (2014): „IT-Skills-Studie 2014: Big Data Projekte“, White-Paper, 2014.

[Edison2019] EDISON Community (2019): „EDISON Data Science Framework (EDSF)“, <https://github.com/EDISONcommunity/EDSF>. (Aufgerufen am: 08.11.2019).

[GI2013] Gesellschaft für Informatik e.V. (2013): Informatiklexikon: Big Data, <https://gi.de/informatiklexikon/big-data/>. (Aufgerufen am 08.11.2019).

[GI2016] Gesellschaft für Informatik e.V. (2016): Empfehlungen für Bachelor- und Masterprogramme im Studienfach Informatik an Hochschulen, https://dl.gi.de/bitstream/handle/20.500.12116/2351/58-GI-Empfehlungen_Bachelor-Master-Informatik2016.pdf?sequence=1&isAllowed=y. (Aufgerufen am: 08.11.2019).

[Heidrich2018] Heidrich, J.; Bauer, P.; Krupka, D. (2018): „Future Skills: Ansätze zur Vermittlung von Data-Literacy in der Hochschulbildung“, https://hochschulforumdigitalisierung.de/sites/default/files/dateien/HFD_AP_Nr37_DALI_Studie.pdf. (Aufgerufen am: 08.11.2019).

[Kauermann2019] Kauermann, G. (2019): Data Science – einige Gedanken aus Sicht eines Statistikers, Informatik Spektrum, online first.

[KerstinTresp2019] Kersting, K.; Tresp, V.: „Maschinelles und Tiefes Lernen. Der Motor für ‚KI made in Germany‘. Whitepaper der Plattform Lernende Systeme“ https://www.plattform-lernende-systeme.de/files/Downloads/Publicationen/AG1_Whitepaper_280619.pdf. (4. Dezember 2019).

[Lübcke2018] Lübcke, M.; Wannemacher, K.: Vermittlung von Datenkompetenzen an den Hochschulen: Studienangebote im Bereich Data Science, Forum Hochschulentwicklung 01/2018.

[Markl2015] Markl, V.: „Gesprengte Ketten“, in: Informatik Spektrum 01/2015.

[Ridsdale2015] Ridsdale, C.; Rothwell, J.; Smit, M. u.a. (2015): Strategies and Best Practices for Data Literacy Education: Knowledge Synthesis Report, Halifax (Canada) Dalhousie University.

[SPSS2000] Chapman, P.; Clinton, J.; Kerber, R. u.a. (2000): „CRISP-DM 1.0: Step-by-step data mining guide“, White-Paper, PSSS.

[Weihslckstadt2018] Weihs und Ickstadt (2018): Data Science: the impact of statistics, International Journal of Data Science and Analytics 6, 189–194).

[Weinert2001] Weinert, F. E. (2001): „Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit“, in: Weinert, F. E. (Hrsg.): Leistungsmessung in Schulen. Weinheim u. Basel: Beltz, S. 17-33.

ANHANG

Die erste nachfolgende Tabelle enthält in Spalte 1 die für ein Data-Science-Curriculum notwendigen Kompetenzgruppen. Die Spalten 2 bis 6 beschreiben die Wichtigkeit der Module (Pflichtmodul versus Wahlpflichtmodul) und die Lernziele (L1 bis L3: Verstehen, Anwenden, Analysieren) in Bezug auf die Personas in Kapitel 3.

Die vorliegenden Kompetenzfelder finden sich auch bei EDISON wieder:

• Grundlagen Mathematik	KU1.01.00ff
• Fortgeschrittene Mathematik	KU1.01.00ff
• Grundlagen der Informatik	KU2.03.00ff
• Fortgeschrittene Informatik	KU2.01.00ff, KU2.02.00ff, KU2.05.00ff, KU2.06.00, (KU3.01.00ff), (KU3.02.00ff), (KU3.03.00ff), (KU3.05.00ff), (KU3.06.00ff)
• Kryptographie und Sicherheit	KU2.04.00ff
• Datenethik und Data Privacy	KU3.01.06, KU3.04.w07, KU4.01.06
• Data Governance	KU3.04.00ff
• Datenintegration	KU1.03.00ff
• Data Mining	KU1.04.00ff, KU2.07.00ff
• Datenvisualisierung	KU2.06.06ff
• Maschinelles Lernen	KU1.02.00ff, KU1.05.00ff, KU4.01.00ff
• Business Intelligence	KU5.01.00ff, KU5.02.00ff
• Domänenspezifische Anwendungen	(KU2.05.06), (KU1.06.01)
• Data Science in der Organisation	KU4.02.00ff

ANHANG 1: AUFSCHLÜSSELUNG DER DATA-SCIENTIST-LERNINHALTE

PERSONA	A Master Data Science	B.1 Master Data Science (Domäne)	B.2 Master Domäne (mit Data- Science- Kompetenzen)	C.1 Basic Data Scientist (Weiterbildung)	C.2 Advanced Data Scientist (Weiterbildung)
KOMPETENZFELD	- P=Pflichtinhalt, W=Wahlpflichtinhalt				
LERNINHALT	- Lernziele: L1=Verstehen, L2=Anwenden, L3=Analysieren				
(1) Grundlagen Mathematik					
Lineare Algebra	-	-	-	-	-
Statistik (Deskriptive Statistik, Statistische Inferenz, Lineare Modelle, Simulation und Resampling)	-	-	-	-	-
Stochastik	-	-	-	-	-
Analysis/Calculus	-	-	-	-	-
(2) Fortgeschrittene Mathematik					
Fortgeschrittene Lineare Algebra	P, L3	W, L3	-	-	P, L2
Fortgeschrittene Statistik (Nichtparametrische, Multivariate, Hochdimensionale, Bayesianische, Räumliche)	P, L3	W, L3	-	-	P, L2
Fortgeschrittene Stochastik	P, L3	W, L3	-	-	P, L2
Fortgeschrittene Analysis/Calculus	P, L3	W, L3	-	-	P, L2
Simulationsverfahren	P, L3	W, L3	-	-	P, L2
(3) Grundlagen Informatik					
Programmierung	-	-	-	-	P, L2
Software-Engineering	P, L2	W, L1	W, L1	-	P, L2
Datenbanken/SQL	P, L2	W, L1	W, L1	-	P, L2
Komplexität von Algorithmen	P, L2	W, L1	W, L1	-	P, L2
(4) Fortgeschrittene Informatik					
Virtualisierung & Container Management	P, L2	-	-	P, L1	P, L2
Architekturen	P, L2	-	-	P, L1	P, L2
NoSQL Data Storage	P, L2	-	-	P, L2	P, L2
Streaming & Streaming Analytics	P, L2	-	-	P, L2	P, L2
Cloud Computing	P, L2	-	-	P, L1	W, L2
Fortgeschrittene Algorithmen	W, L2	-	-	-	W, L2
(5) Kryptographie und Sicherheit					
Datensicherheit/Security by Design	W, L1	-	-	W, L1	W, L1
Kryptographie	W, L1	-	-	W, L1	W, L1
Cyber Sicherheit	W, L1	-	-	W, L1	W, L1
Blockchain in Data Science	W, L1	-	-	-	-
(6) Datenethik und Data Privacy					
Datenethik	P, L3	P, L2	W, L2	P, L3	-
Rechtlicher Rahmen (DSGVO etc.)	P, L2	P, L2	W, L2	P, L1	P, L2
Data Privacy & Data Compliance	P, L3	P, L2	W, L2	P, L1	P, L3
(7) Data Governance					
Data Policy	P, L1	P, L1	W, L1	P, L1	-
Metadatenmanagement	P, L2	P, L2	P, L2	P, L1	P, L2
Strukturen und Verantwortlichkeiten	P, L1	P, L1	W, L1	P, L1	-

ANHANG 1: AUFSCHLÜSSELUNG DER DATA-SCIENTIST-LERNINHALTE

(8) Datenintegration		ausgewählte Themen, die sich an der Domäne orientieren				
Datensammlung						
	<i>Instrumentation</i>	P, L2	P, L2	W, L1	P, L1	P, L2
	<i>Logging</i>	P, L2	P, L2	W, L1	P, L1	W, L2
	<i>Sensoren</i>	P, L2	P, L2	W, L1	P, L1	W, L2
	<i>Datenquellen</i> (Social Media, User Generated, Suchmaschinen, Kaggle etc.)	P, L2	P, L2	W, L1	P, L1	P, L2
Daten-Pipelines						
	<i>Data Flow</i>	P, L2	P, L2	P, L2	P, L2	P, L2
	<i>Infrastructure and Tools</i> (Kafka, MLFlow, Cloud Dataflow, AWS Pipelines etc.)	P, L2	P, L2	W, L1	P, L2	P, L2
	<i>ETL: Extraktion, Transformation, Laden</i>	P, L2	P, L2	W, L1	P, L2	P, L2
	<i>Strukturierte/Unstrukturierte Daten</i>	P, L2	P, L2	P, L2	P, L2	W, L2
Datenvorbereitung						
	<i>Data Quality, Data Curation etc.</i>	P, L2	P, L3	P, L2	P, L2	P, L2
	<i>Data-Wrangling/-Transformation/-Cleaning</i>	P, L2	P, L3	P, L2	P, L2	P, L2
	<i>Anomaly Detection</i>					
	<i>Basic Labelling/Aggregation, Analytics, Metrics, Segmentation</i>	P, L2	P, L3	P, L2	P, L2	P, L2
	<i>Feature Selection/Extraction, Training Sets etc.</i>	P, L2	P, L3	P, L2	P, L2	P, L2
(9) Datenvisualisierung		ausgewählte Themen, die sich an der Domäne orientieren				
	<i>Definition/Workflow</i>	P, L2	P, L1	W, L1	P, L2	P, L2
	<i>Perception Theory, Colour, Editorial Thinking</i>	P, L2	P, L1	W, L1	P, L2	P, L2
	<i>Principles, Representations</i>	P, L2	P, L1	W, L1	P, L2	P, L2
	<i>Tools</i>	P, L2	P, L1	W, L1	P, L2	P, L2
	<i>Visualization Dimensions</i>	P, L2	P, L1	W, L1	-	P, L1
	<i>Interactivity, Annotations</i>	P, L2	P, L1	W, L1	-	P, L1
(10) Data Mining (DM)		ausgewählte Themen, die sich an der Domäne orientieren				
	<i>DM KDD Process</i> (Knowledge Discovery in Databases)	P, L2	P, L2	W, L1	P, L1	P, L2
	<i>DM IR Methods</i> (Information Retrieval)	P, L2	P, L2	W, L1	-	P, L2
	<i>DM Reporting</i>	P, L2	P, L2	W, L1	-	P, L2
	<i>DM Text-, Web-, Process-Mining</i>	W, L3	P, L3	W, L2	W, L1	W, L2
	<i>Time Series Analytics</i>	W, L3	P, L3	W, L2	-	W, L2

ANHANG 1: AUFSCHLÜSSELUNG DER DATA-SCIENTIST-LERNINHALTE

(11) Maschinelles Lernen (ML): Sprachen und Werkzeuge	ausgewählte Themen, die sich an der Domäne orientieren				
Grundlegende ML-Methoden					
ML Sprachen (Python, R, Julia etc.)	P, L3	-	-	P, L2	P, L3
ML Bibliotheken (SciKitLearn, Dask etc.)	P, L3	W, L1	W, L1	P, L2	P, L3
ML Workbenches (KNIME, Weka, Rapidminer, Matlab etc.)	P, L3	W, L1	W, L1	P, L2	P, L3
Big Data Analytics Frameworks (Spark etc.)	P, L3	W, L2	W, L1	P, L2	P, L3
Grundlegende ML-Konzepte (Experiment Design, ML-Workflow, Trainingsdaten/Segmentierung, Overfitting/BIAS, Re-evaluation etc.)	P, L3	P, L3	P, L2	P, L2	P, L3
Klassische Inferenz Methoden (Regression, Support Vector Machines, Instance Based Methods/Classification, Decision Trees, Clustering, Kernels etc.)	P, L3	P, L3	P, L2	P, L2	P, L3
Regularization, Dimensionality Reduction, Ensemble Methods etc.)	P, L3	P, L3	P, L2	P, L2	P, L3
Probabilistische Modelle und Konzepte ((Hidden) Markov Modelle, Graphische Modelle etc.)	P, L3	P, L3	P, L2	P, L1	W, L2
ML-Anwendungen (Natural Language Processing etc.)	W, L1	W, L1	W, L1	W, L1	W, L1
Deep Learning/Deep Neural Networks					
Types & Concepts of Neural Networks	P, L3	P, L3	P, L2	P, L2	P, L3
Neural Network Imaging	P, L3	P, L3	P, L2	P, L2	P, L3
Neural Network Practice/Libraries (TensorFlow, Keras, Caffe, Torch, OpenNN, Theano, etc.)	P, L3	P, L3	P, L2	P, L2	P, L3
Reinforcement Learning/GAN	P, L3	P, L3	P, L2	P, L2	P, L3
(12) Klassisches Business Intelligence (BI)					
BI Phasen	W, L2	-	-	P, L1	P, L2
BI Data Warehousing/OLTP/OLAP	W, L2	-	-	W, L1	P, L2
BI ERP Systems (z.B. SAP)	W, L2	-	-	W, L1	P, L2
BI ETL Systems (z.B. Talend)	W, L2	-	-	W, L1	P, L2
(13) Domänenspezifische Anwendungen					
Domänenspezifische praktische Erfahrung (Tools/Visualisierung/Pipelining/Kommunikation bspw. in Biologie, Logistik, Handel, Urban Tech etc.)	P, L1	P, L3	P, L3	-	-
Forschungsdaten-Management	P, L1	P, L3	P, L3	-	-
(14) Data Science in der Organisation (im Kontext)	WiWi Teil der Domain, für nicht-WiWi: Entrepreneurship)				
Data Science Projektmanagement (Identifizieren, Implementieren, Einbetten etc.)	P, L2	W, L2	-	P, L1	P, L2
Organisationsspezifische Kommunikationsfähigkeiten & Soft Skills	P, L2	P, L2	-	-	-
Daten-Ökonomie/Wert in der Organisation	W, L2	W, L2	-	-	-

ANHANG 2: NOTWENDIGE VORAUSSETZUNGEN PERSONA B UND C

PERSONA	Voraussetzungen für A	Voraussetzungen für B.1 oder B.2	Voraussetzungen für B.1
KOMPETENZFELD			
LERNINHALT			
(1) Grundlagen Mathematik			
Lineare Algebra	P, L2	P, L2	P, L2
Statistik	P, L2	P, L2	P, L2
Stochastik	P, L2	P, L2	P, L2
Analysis/Calculus	P, L2	P, L2	P, L2
(2) Fortgeschrittene Mathematik			
Fortgeschrittene Lineare Algebra	-	-	-
Fortgeschrittene Statistik	-	-	-
Fortgeschrittene Stochastik	-	-	-
Fortgeschrittene Analysis/Calculus	-	-	-
Simulationsverfahren	-	-	-
(3) Grundlagen Informatik			
Programmierung	P, L2	P, L2	P, L2
Software-Engineering	W, L1	-	W, L1
Datenbanken/SQL	W, L1	-	W, L1
Komplexität von Algorithmen	W, L1	-	W, L1
(4) Fortgeschrittene Informatik			
Virtualisierung & Container Management	W, L1	-	-
Architekturen	W, L1	-	-
NoSQL Data Storage	W, L1	-	-
Streaming & Streaming Analytics	-	-	-
Cloud Computing	-	-	-
Fortgeschrittene Algorithmen	W, L1	-	-
(5) Kryptographie und Sicherheit			
Datensicherheit/Security by Design	-	-	-
Kryptographie	-	-	-
Cyber Sicherheit	-	-	-
Blockchain in Data Science	-	-	-
(6) Datenethik und Data Privacy			
Datenethik	W, L1	-	W, L1
Rechtlicher Rahmen (DSGVO etc.)	W, L1	-	W, L1
Data Privacy & Data Compliance	W, L1	-	W, L1

ANHANG 2: NOTWENDIGE VORAUSSETZUNGEN PERSONA B UND C

(7) Data Governance			
<i>Data Policy</i>	-	-	-
<i>Metadatenmanagement</i>	-	-	-
<i>Strukturen und Verantwortlichkeiten</i>	-	-	-
(8) Datenintegration			
Datensammlung			
<i>Instrumentation</i>	-	-	-
<i>Logging</i>	-	-	-
<i>Sensoren</i>	-	-	-
<i>Datenquellen (Social Media, User Generated, Suchmaschinen, Kaggle etc.)</i>	-	-	-
Daten-Pipelines			
<i>Data Flow</i>	-	-	-
<i>Infrastructure and Tools</i>	-	-	-
<i>ETL: Extraktion, Transformation, Laden</i>	-	-	-
<i>Strukturierte/Unstrukturierte Daten</i>	-	-	-
Datenvorbereitung			
<i>Data Quality, Data Curation etc.</i>	-	-	-
<i>Data-Wrangling/-Transformation/-Cleaning</i>	-	-	-
<i>Anomaly Detection</i>	-	-	-
<i>Basic Labelling/Aggregation, Analytics, Metrics, Segmentation</i>	-	-	-
<i>Feature Selection/Extraction, Training Sets etc.</i>	-	-	-
(9) Datenvisualisierung			
<i>Definition/Workflow</i>	-	-	-
<i>Perception Theory, Colour, Editorial Thinking</i>	-	-	-
<i>Principles, Representations</i>	-	-	-
<i>Tools</i>	-	-	-
<i>Visualization Dimensions</i>	-	-	-
<i>Interactivity, Annotations</i>	-	-	-
(10) Data Mining (DM)			
<i>DM KDD Process (Knowledge Discovery in Databases)</i>	-	-	-
<i>DM IR Methods (Information Retrieval)</i>	-	-	-
<i>DM Reporting</i>	-	-	-
<i>DM Text-, Web-, Process-Mining</i>	-	-	-
<i>Time Series Analytics</i>	-	-	-

ANHANG 2: NOTWENDIGE VORAUSSETZUNGEN PERSONA B UND C

(11) Maschinelles Lernen (ML): Sprachen und Werkzeuge			
Grundlegende ML-Methoden			
<i>ML Sprachen</i>	-	P, L1	-
<i>ML Bibliotheken</i>	-	-	-
<i>ML Workbenches</i>	-	-	-
<i>Big Data Analytics Frameworks</i>	-	-	-
<i>Grundlegende ML-Konzepte</i>	-	-	-
<i>Klassische Inferenz Methoden</i>	-	-	-
<i>Regularization, Dimensionality Reduction,</i>	-	-	-
<i>Probabilistische Modelle und Konzepte</i>	-	-	-
<i>ML-Anwendungen</i>	-	-	-
Deep Learning/Deep Neural Networks			
<i>Types & Concepts of Neural Networks</i>	-	-	-
<i>Neural Network Imaging</i>	-	-	-
<i>Neural Network Practice/Libraries</i> (TensorFlow, Keras, Caffee, Torch, OpenNN, Theano, etc.)	-	-	-
<i>Reinforcement Learning/GAN</i>	-	-	-
(12) Klassisches Business Intelligence (BI)			
<i>BI Phasen</i>	-	-	-
<i>BI Data Warehousing/OLTP/OLAP</i>	-	-	-
<i>BI ERP Systems (z.B. SAP)</i>	-	-	-
<i>BI ETL Systems (z.B. Talend)</i>	-	-	-
(13) Domänenspezifische Anwendungen			
<i>Domainspezifische praktische Erfahrung</i> (Tools/Visualisierung/Pipelining/Kommunikation n bspw. in Biologie, Logistik, Handel, Urban Tech etc.)	-	-	-
<i>Forschungsdaten-Management</i>	-	-	-
(14) Data Science in der Organisation (im Kontext)			
<i>Data Science Projektmanagement</i> (Identifizieren, Implementieren, Einbetten etc.)	-	-	-
<i>Organisationsspezifische Kommunikationsfähigkeiten & Soft Skills</i>	-	-	-
<i>Daten-Ökonomie/Wert in der Organisation</i>	-	-	-

AUTORINNEN UND AUTOREN

Folgende Autorinnen und Autoren (in alphabetischer Reihenfolge) haben an den Empfehlungen mitgearbeitet:

- Abedjan, Ziawasch (TU Berlin / GI-FG DB)
- Brefeld, Ulf (Leuphana Universität Lüneburg / Plattform Lernende Systeme)
- Bürkle, Joachim (DB Systel / Deutsche Bahn)
- Desel, Jörg (FernUniversität Hagen / GI-FG ISH / Studienkommission Fakultätentag Informatik)
- Edlich, Stefan (Beuth Hochschule, Berlin)
- Eppler, Thomas (Hochschule Albstadt-Sigmaringen)
- Goedicke, Michael (Universität Duisburg-Essen / GI-Vizepräsident)
- Heidrich, Jens (Fraunhofer-Institut für Experimentelles Software Engineering IESE / GI-FG Software-Messung und -Bewertung)
- Höppner, Stephan (Atos)
- Kast, Stefan M. (TU Dortmund / Gesellschaft Deutscher Chemiker)
- Krupka, Daniel (Gesellschaft für Informatik)
- Lang, Klaus (TH Bingen / Vorsitzender Fachbereichstag Informatik)
- Liggesmeyer, Peter (Fraunhofer IESE / Sprechende GI-Task Force „Data Science“)
- Tropmann-Frick, Marina (HAW Hamburg)

IMPRESSUM

HERAUSGABE

Gesellschaft für Informatik e.V.
Spreepalais am Dom, Anna-Louisa-Karsch-Str. 2, 10178 Berlin

REDAKTION

Ziawasch Abedjan, Ulf Brefeld, Joachim Bürkle, Jörg Desel, Stefan Edlich, Thomas Eppler,
Michael Goedicke, Jens Heidrich, Stephan Höppner, Stefan M. Kast, Daniel Krupka, Klaus Lang,
Peter Liggesmeyer, Marina Tropmann-Frick, Klaus Lang

GESTALTUNG

Sarah Bauer

STAND

Dezember 2019

COPYRIGHT

Diese Publikation steht unter der Lizenz CC BY-SA 4.0.

ÜBER DIE GESELLSCHAFT FÜR INFORMATIK E. V.

Die Gesellschaft für Informatik e.V. (GI) ist mit rund 20.000 persönlichen und 250 korporativen Mitgliedern die größte und wichtigste Fachgesellschaft für Informatik im deutschsprachigen Raum. 2019 feiert die GI ihr 50-jähriges Gründungsjubiläum. Seit 1969 vertritt sie die Interessen der Informatikerinnen und Informatiker in Wissenschaft, Wirtschaft, öffentlicher Verwaltung, Gesellschaft und Politik. Mit 14 Fachbereichen, über 30 aktiven Regionalgruppen und unzähligen Fachgruppen ist die GI Plattform und Sprachrohr für alle Disziplinen in der Informatik. Die GI-Mitglieder binden sich an die Ethischen Leitlinien für Informatikerinnen und Informatiker der Gesellschaft für Informatik e.V.. Weitere Informationen finden Sie unter www.gi.de.

GESELLSCHAFT FÜR INFORMATIK E. V. (GI)

Geschäftsstelle Bonn
Wissenschaftszentrum
Ahrstr. 45
53175 Bonn
Tel.: +49 228 302-145
Fax: +49 228 302-167
E-Mail: bonn@gi.de

Geschäftsstelle Berlin
Spreepalais am Dom
Anna-Louisa-Karsch-Str. 2
10178 Berlin
Tel.: +49 30 7261 566-15
Fax: +49 30 7261 566-19
E-Mail: berlin@gi.de

gs@gi.de
www.gi.de

 [/informatikradar](https://twitter.com/informatikradar)
 [/company/gesellschaft-fuer-informatik](https://www.linkedin.com/company/gesellschaft-fuer-informatik)
 [/net/gi](https://x.com/net_gi)