Was ist Data Science? Ermittlung der informatischen Inhalte durch Analyse von Studienangeboten

Andreas Grillenberger und Ralf Romeike
Friedrich-Alexander-Universität Erlangen-Nürnberg
Didaktik der Informatik
Martensstraße 3
91058 Erlangen
andreas.grillenberger@fau.de
ralf.romeike@fau.de

Abstract: In Zusammenhang mit den Entwicklungen der vergangenen Jahre, insbesondere in den Bereichen *Big Data, Datenmanagement* und *Maschinenlernen,* hat sich der Umgang mit Daten und deren Analyse wesentlich weiterentwickelt. Mittlerweile wird die Datenwissenschaft als eigene Disziplin angesehen, die auch immer stärker durch entsprechende Studiengänge an Hochschulen repräsentiert wird. Trotz dieser zunehmenden Bedeutung ist jedoch oft unklar, welche konkreten Inhalte mit ihr in Verbindung stehen, da sie in verschiedensten Ausprägungen auftritt. In diesem Beitrag werden daher die hinter der Data Science stehenden informatischen Inhalte durch eine qualitative Analyse der Modulhandbücher etablierter Studiengänge aus diesem Bereich ermittelt und so ein Beitrag zur Charakterisierung dieser Disziplin geleistet. Am Beispiel der Entwicklung eines Data-Literacy-Kompetenzmodells, die als Ausblick skizziert wird, wird die Bedeutung dieser Charakterisierung für die weitere Forschung expliziert.

Keywords: Data Science, Big Data, Inhalte, Studiengänge, Data Literacy, Kompetenzen.

1 Data Science als Gegenstand informatischer Bildung

In Zusammenhang mit der zunehmenden Verarbeitung immer komplexerer Daten in verschiedensten Kontexten werden seit kurzem vielerorts neue Studiengänge entwickelt und eingerichtet, die sich mit der *Data Science* beschäf-

tigen. Innerhalb dieser Studiengänge ist eine deutliche inhaltliche Vielfalt erkennbar. Allgemein werden jedoch üblicherweise zumindest Mathematik, Statistik und Informatik als Fundamente angesehen. Je nach konkreter Ausgestaltung spielen diese jedoch eine unterschiedlich große Rolle, sodass zum Teil deutlich unterschiedliche inhaltliche und methodische Ausrichtungen erkennbar sind. Obwohl die Datenwissenschaft sowohl in der Eigen- als auch der Fremdwahrnehmung oft als eigene Disziplin betrachtet wird, als eine Wissenschaft der (Arbeit mit und Verwaltung von) Daten, weist sie jedoch auch starke Bezüge zur Informatik auf und greift auf vielfältige informatische Grundlagen zurück.

Wie die Gesellschaft für Informatik [GI18] verdeutlicht, ist *Data Science* heute mehr als nur ein Buzzword. Im Gegenteil kommt diesem Wissenschaftsfeld eine wachsende Bedeutung zu, auch aufgrund ihrer klaren Bezüge zu der Entwicklung, die in Alltag, Gesellschaft und Politik häufig unter dem Begriff *Digitalisierung* zusammengefasst wird: Indem versucht wird, alle Bereiche des täglichen Lebens in Daten abzubilden, entsteht die Herausforderung, diese adäquat zu verwalten und zu nutzen. Diese Entwicklung macht selbst vor dem Privatleben, in dem oft Ergebnisse der kontinuierlichen Datenerfassung und -verarbeitung genutzt werden, nicht Halt: Im Sinne einer *Data Literacy* muss heute jeder einen fundierten Umgang mit Daten pflegen, solche speichern und verarbeiten. Aber auch in den meisten anderen Wissenschaftsdisziplinen gewinnt Data Science bzw. ihre Methoden an Bedeutung und wird sogar als neues viertes Wissenschaftsparadigma gehandelt (z. B. [HTT09]). Entsprechend entstehen, beispielsweise in den Geisteswissenschaften unter dem Begriff *Digital Humanities,* neue Ausrichtungen, die mindestens grundlegende Kompetenzen aus der Datenwissenschaft benötigen.

Für die informatische Bildung eröffnen sich in diesem Bereich vielfältige Forschungs- und Handlungsfelder, u. a. bei der Ermittlung und Fundierung der Kompetenzen sowohl von Data Science als auch von Data Literacy, aber auch bei der Integration dieser Bereiche in die (Hochschul-)Ausbildung und der Gestaltung und Weiterentwicklung von Data-Science-Studiengängen bzw. Data-Literacy-Modulen. Als Basis für derartige Arbeiten muss eine Klarheit darüber geschaffen werden, welche Aspekte der Informatik für die Data Science relevant bzw. notwendig sind. In diesem Beitrag erfolgt daher eine Klärung des informatischen Gegenstandsbereichs *Data Science,* indem die hinter dem Begriff stehenden informatischen Inhalte durch eine explorative Analyse bereits implementierter Data-Science-Studiengänge ermittelt werden. Somit wird ein Überblick über die Data Science gegeben und diese durch ihre informatischen Aspekte charakterisiert.

2 Verwandte Arbeiten

Trotz der verbreiteten Nutzung des Begriffs *Data Science* besteht keine klare Einigkeit über deren charakteristische Inhalte und Kompetenzen: Aus disziplinorientierter Sicht wird sie beispielsweise als *„data science = statistics + informatics + computing + communication + sociology + management | data + environment + thinking“*[1] definiert [Ca17]. Andere Definitionen legen den Schwerpunkt auf den praktischen Nutzen: *„Data science is the extraction of actionable knowledge directly from data through a process of discovery, or hypothesis formulation and hypothesis testing.“* [NI15] In verschiedenen Arbeiten wurde daher bereits versucht, diese neue Wissenschaftsdisziplin zu charakterisieren, beispielsweise durch ein Kompetenzmodell (vgl. [DBM17]) sowie einen Body of Knowledge (vgl. [DBW17]). Dabei wurde zum Teil empirisch vorgegangen, indem als Basis für das Kompetenzmodell die Anforderungen, die in Stellenanzeigen an Datenanalysten gestellt werden, herangezogen und Studien zu deren Kompetenzen und Fertigkeiten sowie Blogartikel und Forenbeiträge miteinbezogen wurden. Das auf dieser Basis entstandene Kompetenzmodell [DBM17] gibt somit insbesondere eine externale Sicht auf das wieder, was Data Science bzw. Data Scientists aus Sicht von Unternehmen sein bzw. können sollen. Im Gegensatz dazu wird in diesem Beitrag angestrebt, die wissenschaftlich geprägte Selbstwahrnehmung des Fachs mit Fokus auf die informatischen Aspekte abzubilden.

Auch in anderen Arbeiten wird die Bedeutung der Data Science deutlich: So konnte das HIS-Institut für Hochschulentwicklung in Deutschland bereits eine relativ große Vielfalt von rund 25 Studiengängen zur Data Science und eine steigende Tendenz ermitteln (vgl. [GI18]). In der informatikdidaktischen Forschung bleibt die Disziplin Data Science bisher jedoch nahezu unbeachtet: Zwar wurden bereits auf Bildung hin ausgerichtete Arbeiten durchgeführt und Richtlinien für Data-Science-Curricula erarbeitet (z. B. [Ve17]). Eine tiefergehende Betrachtung aus informatikdidaktischer Sicht fand bisher jedoch nicht statt. Verwandte Arbeiten aus der Informatikdidaktik stammen damit eher aus anderen Bereichen der Informatik: Insbesondere dienen Arbeiten zum Fachgebiet Datenmanagement, beispielsweise das Modell der Schlüsselkonzepte des Datenmanagements [GR17], aufgrund der starken Überschneidung mit der Data Science im Rahmen der weiteren Forschung zu Data Science und Data Literacy als wertvolle Grundlage.

---

1 Das Zeichen „|“ ist dabei als „conditional on“ zu verstehen [Ca17].

3 Ermittlung der Kerninhalte von Data-Science-Studiengängen

Zur fachlichen Fundierung der (informatikdidaktischen) Data-Science-Forschung ist es hilfreich und notwendig, die informatischen Aspekte hinter dem Begriff *Data Science* zu ermitteln. Einen wichtigen Einblick kann hier eine Untersuchung bereits etablierter Data-Science-Studiengänge geben.

Im Rahmen einer Voruntersuchung wurden Modulhandbücher weniger Data-Science-Studiengänge[2] betrachtet, die einen Einblick sowohl in die enthaltenen Inhalte als auch Kompetenzen geben. Dabei hat sich gezeigt, dass die Inhalte, die von den die Studiengänge gestaltenden Experten als zentral für die Data Science angesehen werden, aus den betrachteten Dokumenten klar ersichtlich sind. Die Untersuchung der angestrebten Kompetenzen wäre auf dieser Basis jedoch stark subjektiv geprägt: Aufgrund der unterschiedlichen Formulierung und Detaillierung der jeweils festgehaltenen Kompetenzen, beispielsweise als zum Teil nicht operationalisierte Lernziele oder auch rein inhaltsorientierte Modulbeschreibungen, würden interpretative Aspekte eine wichtige Rolle spielen. Die Objektivität der Untersuchung wäre somit stark eingeschränkt. Entsprechend wird in dieser Arbeit der Schwerpunkt auf die objektiver untersuchbaren Inhalte gelegt. Im Rahmen der Vorstudie wurden zusätzlich deutschsprachige und internationale Studiengänge[3] kontrastiert: Dabei zeigte sich, dass – vermutlich aufgrund unterschiedlicher rechtlicher Rahmenbedingungen – die Dokumente nicht nur einen sehr unterschiedlichen Aufbau haben, sondern sich auch im Detailgrad stark unterscheiden. Um auch an dieser Stelle eine möglichst nachvollziehbare Methodik anzuwenden und subjektive Einflüsse zu vermeiden, wird der Fokus daher auf Studiengänge im deutschsprachigen Raum gelegt, bei denen ein ähnlicher Data-Science-Begriff und ähnliche Grundlagen zur Ausgestaltung der charakteristischen Dokumente zu erwarten sind und somit alle Dokumente auf dieselbe Weise untersucht werden können. Um die Validität der Ergebnisse auch außerhalb des deutschsprachigen Raums zu überprüfen, werden die Ergebnisse der Untersuchung jedoch im Nachgang mit internationalen Studiengängen kontrastiert.

---

2 In der Vorabuntersuchung wurden Data-Science-Studiengänge der Universitäten Marburg und Stuttgart sowie der Hochschule der Medien Stuttgart und der Hochschule Albstadt-Sigmaringen betrachtet.

3 Zur Kontrastierung wurden die Data-Science-Studiengänge der Universität Berkeley und der IT-Universität Copenhagen betrachtet.

Die zentrale Forschungsfrage dieser Arbeit lautet daher: *Welche zentralen Inhalte charakterisieren Data-Science-Studiengänge im deutschsprachigen Raum?* Um dieser nachzugehen, wurde ein empirischer Ansatz basierend auf einer qualitativen Inhaltsanalyse nach Mayring [Ma10] gewählt. Dieser erlaubt es unter anderem, einen Literaturkanon systematisch zu explorieren und darauf basierend ein Kategoriensystem aufzubauen, das hier der inhaltlichen Charakterisierung der Data Science entsprechen wird. Im Allgemeinen kann dieses Kategoriensystem induktiv, aus dem Material heraus, oder deduktiv, auf Basis existierender Arbeiten, aufgebaut werden. Das angestrebte Ziel dieser Analyse legt klar einen induktiven Ansatz nahe, da damit das analysierte Material adäquat repräsentiert werden kann, ohne dass die Einordnung in ein existierendes Kategoriensystem zu einer Beeinflussung und zu Einschränkungen bei der Exploration kommen kann.

Im Sinne der Methodik nach Mayring wird die Analyse in mehrere Schritte zerlegt: Zuerst wird der Literaturkanon und, zur Erhöhung der Genauigkeit und Objektivität, auch die Kodiereinheit und Analysekriterien festgelegt. Darauf basierend erfolgt die Analyse der Dokumente, die typischerweise in die Erstellung eines hierarchisch organisierten Kategoriensystems mündet. In dieser Arbeit wird jedoch vorerst auf die Hierarchisierung verzichtet und stattdessen eine flache Kategorienmenge aufgebaut und erst im Nachgang in einer expliziten Strukturierungsphase hierarchisiert. Dies wird für das angestrebte Ziel als vorteilhaft erachtet, da auf diese Weise frühe Beeinflussungen der Analyse durch die Hierarchisierung und den Versuch, neue Aspekte eher in das bereits existierende System einzuordnen anstatt die Hierarchie zu ergänzen, vermieden werden.

3.1 Festlegung von Literaturkanon, Kodiereinheit und Analysekriterien

Als Basis für die Analyse werden Modulhandbücher verschiedener Data-Science-Studiengänge aus dem deutschsprachigen Raum ausgewählt. Um einen Überblick über diese zu gewinnen, wurde der Hochschulkompass[4] herangezogen. Unter den dort auffindbaren Studiengängen sind jedoch verschiedene, die zwar Themen der Data Science anschneiden, aber eher am Rande betrachten und andere Schwerpunkte setzen: beispielsweise Informatikstudiengänge,

4 http://www.hochschulkompass.de, zugegriffen am 25. 06. 2018

die lediglich wenige Data-Science-Module verpflichtend beinhalten oder nur die Möglichkeit bieten, sich in Wahlpflichtmodulen darauf zu spezialisieren. Um eine Beeinflussung der Ergebnisse durch diese zu vermeiden, wurden solche Studiengänge nicht miteinbezogen, da in diesen nicht klar erkennbar ist, welche Inhalte speziell für den Bereich Data Science als zentral erachtet werden.

Innerhalb der auf diese Weise vorselektierten Studiengänge wurde eine weitere Filterung der betrachteten Module vorgenommen: Da das Ziel die Ermittlung zentraler Inhalte der Data-Science-Studiengänge war, wurden nur Module betrachtet, die von den Hochschulen als Pflicht – und somit als Mindestanforderungsprofil für Absolventen des jeweiligen Studiengangs – erachtet werden. Wahlpflicht- und Wahlmodule blieben somit unberücksichtigt, genauso wie konsekutive Masterstudiengänge, die als Vertiefung gegenüber dem vorherigen Data-Science-Bachelorstudiengang betrachtet wurden. Nicht-konsekutive Master- wurden jedoch genauso wie Bachelorstudiengänge miteinbezogen, da sie gleichermaßen die Grundlagen der Data Science thematisieren müssen.

Diesen Kriterien folgend wurden Studiengänge der folgenden Hochschulen, unabhängig von ihrer Ausgestaltung (bspw. mit Fokus auf Informatik oder Mathematik), als Basis für die Analyse ausgewählt: *Beuth Hochschule Berlin, Hochschule Darmstadt, Technische Universität Dortmund, Universität Jena, Universität München/LMU, Universität Mannheim*[5]*, Hochschule Albstadt-Sigmaringen, Universität Salzburg, Hochschule der Medien Stuttgart* (alle M. Sc.), *Universität Marburg, Universität Stuttgart* (beide B. Sc.).

Die Kodiereinheit wurde auf semantische Weise so festlegt, dass jede Kodierung, unabhängig von ihrer Länge im Text, sich jeweils auf genau einen inhaltlichen Aspekt beziehen soll. Als Auswahlkriterium wurde festgelegt, dass alle in den Dokumenten genannten Inhalte betrachtet werden, jedoch wurden aufgrund des Analyseziels Inhalte, die nicht charakteristisch für Data Science sind, sondern allgemeine informatische oder mathematische Grundlagen repräsentieren, nur in geringem Detailgrad erfasst. Somit wurden beispielsweise *endliche Automaten* und *elementare Statistik* nicht im Detail erfasst, sondern nur die Kategorien *Statistik* bzw. *Theoretische Informatik* eingeführt. Hingegen wurde beispielsweise die Methode *Klassifikation,* die in der Datenanalyse eine wichtige Rolle spielt und in der Betrachtung in den Curricula

5 Der *Mannheim Master in Data Science* ist im Folgenden nicht erkennbar, da für diesen keine Pflichtmodule definiert sind und somit im Sinne der Auswahlkriterien keine Module berücksichtigt werden konnten.

über die rein mathematische Sichtweise hinausgeht, zunächst explizit in das Kategoriensystem aufgenommen.

3.2 Analyse der Dokumente und Strukturierung der Ergebnisse

Als nächstes wurden die ausgewählten Dokumente unter Berücksichtigung der festgelegten Kriterien systematisch analysiert. Statt direkt ein hierarchisches Kategoriensystem aufzubauen, wurde eine Liste aller genannten Inhalte erstellt. Dabei wurden Kodierungen vermieden, die zu detailliert sind, soweit dies bereits an dieser Stelle klar ersichtlich war: Beispielsweise wurden bei gemeinsamer Nennung einer großen Anzahl verschiedener Methoden zum überwachten Lernen nicht alle Methoden einzeln aufgenommen, sondern gesammelt unter dem geeigneten Überbegriff, da diese im nächsten Schritt sowieso zusammengefasst worden wären. Nicht in allen Fällen war dies jedoch bereits zu diesem Zeitpunkt ersichtlich, insbesondere wenn verwandte Begriffe nicht gemeinsam, sondern in unterschiedlichen Bereichen oder Dokumenten genannt wurden. Diese potentiellen Zusammenfassungen wurden vorerst unterlassen. Somit resultierte die Analyse in einer Liste von 106 inhaltlichen Aspekten der Data Science, die jedoch nicht überschneidungsfrei und auf unterschiedlichem Detailgrad angesiedelt waren. Daher wurde diese Liste im folgenden Schritt strukturiert und zusammengefasst.

Dazu wurde diese Liste hierarchisiert: Es wurde entschieden, dass auf der obersten Ebene nur eher abstrakte Begriffe genannt werden sollen, die die großen Themenbereiche der Data Science repräsentieren. Entsprechend wurden möglichst wenige Codierungen auf oberster Ebene angestrebt. Zusätzlich wurden die beiden Bereiche *informatische Grundlagen* sowie *mathematische Grundlagen* eingeführt, unter denen alle Aspekte zusammengefasst werden, die eher übergreifend bzw. nicht speziell der Data Science zugehörig sind, der Vollständigkeit halber aber nicht unerwähnt bleiben sollen. Trotzdem werden diese Grundlagen bewusst weniger tiefgehend betrachtet, sodass die darunter angeordneten Begriffe auf einem anderen Niveau angesiedelt sind als die im Fokus dieser Arbeit stehenden Inhalte der Data Science. Auf den tieferen Ebenen wurden Begriffe unter dem jeweiligen Oberbegriff zusammengefasst, die von ihrer Bedeutung zusammengehörig sind und häufig gemeinsam genannt wurden, jedoch nur wenn durch die Zusammenfassung keine relevanten Details verloren gehen. Entsprechend wurden die Methoden der Datenanalyse in vier Kategorien subsumiert: *Methoden des unüberwachten Lernens* (insbesondere *Clustering, Assoziation*), *Methoden des überwachten*

Lernens (insbesondere *Klassifikation, Entscheidungsbäume, Regression*), *Komplexere Methoden* (beispielsweise *Neuronale Netze*) sowie *Verknüpfung von Methoden/Ensemble-Learning.* Auf eine prinzipiell mögliche weitere Zusammenfassung, beispielsweise unter dem Begriff *Methoden der Datenanalyse,* wurde bewusst verzichtet, da dadurch die Unterscheidung und die unterschiedlichen Zwecke und Charakteristika in den Hintergrund rücken würden und so insbesondere auch der Einblick in die unterschiedliche Abdeckung dieser Methoden in den verschiedenen Studiengängen verloren gehen würde.

Zusätzlich wurden bei der Strukturierung alle Aspekte aus der Charakterisierung ausgefiltert, die nicht mit anderen zusammenfassbar waren, aber durch ihre geringe Repräsentation in weniger als einem Fünftel der analysierten Dokumente kaum geeignet sind, um den Kern der Data Science zu charakterisieren. Als Ergebnis der Untersuchung steht daher ein Kategoriensystem, das 31 inhaltliche Aspekte aus sechs großen Themenbereichen berücksichtigt. Diese Charakterisierung der Data Science wird in Tab. 1 zusammen mit der Abdeckung in den verschiedenen Studiengängen dargestellt.

3.3 Diskussion der Ergebnisse

Das entstandene Kategoriensystem beschreibt die Data Science aus informatischer Sicht durch vier große Bereiche: *Datenanalyse und Maschinenlernen, Big Data, Datenschutz, Ethik* und *Datenspeicher.* Hinzu kommen *informatische* und *mathematische Grundlagen*.

Ohne detaillierter auf die eigentlichen Inhalte einzugehen, kann bereits ein wesentlicher Einblick in die Data Science gewonnen werden, indem diese Bereiche diskutiert werden:

Notwendige Grundlagen

Obwohl die meisten der analysierten Data-Science-Studiengänge zu einem Masterabschluss hinführen und daher als Aufbaustudium konzipiert sind, ist klar erkennbar, dass sie ein unterschiedlich ausgeprägtes Fundament an informatischen und mathematischen Grundlagen voraussetzen und diese in entsprechenden Modulen thematisieren.

Obwohl die mathematischen Grundlagen nur oberflächlich untersucht wurden, kann ein erster Einblick gewonnen werden: Während alle untersuchten Studiengänge Kenntnisse in *Statistik* als notwendig erachten, werden *Lineare Algebra* und *Analysis* nur in jeweils 40 % der Studiengänge (beide

gemeinsam in 30 %) genannt. Eine detaillierte Untersuchung der mathematischen Aspekte würde jedoch den Rahmen dieser Analyse sprengen.

Auch der Blick auf die *informatischen Grundlagen* zeigt ein ähnliches Bild: Es ist erkennbar, dass insbesondere *Programmierung* sowie *Algorithmen und Datenstrukturen* als besonders zentral erachtet werden und in 90 % bzw. 50 % der analysierten Dokumente genannt wurden. Außerhalb dieser beiden Bereiche besteht jedoch nur eine geringe Übereinstimmung der Studiengänge, beispielsweise werden in 20 % Grundlagen in *Betriebssystemen* und *Theoretischer Informatik* ausgebildet, während nur in einem auch *Rechnerkommunikation* thematisiert wird. Im Bereich der Grundlagen besteht daher zwar ein grundlegendes Fundament, über das große Einigkeit herrscht, aber gleichzeitig eine breite Vielfalt an potentiellen Inhalten die, je nach Studienort und Ausrichtung des Studiengangs, von einem Data-Science-Absolventen beherrscht werden sollen.

Inhaltliche Charakterisierung der Data Science

Neben den notwendigen Grundlagen konnten vier für die Data Science spezifischere Inhaltsbereiche ermittelt werden. Diese umfassen *Datenspeicher,* insbesondere Aspekte des Fachgebiets Datenbanken/Datenmanagement wie (relationale) Datenbanken, Datenmodellierung und Abfragesprachen, den Bereich *Datenschutz und Ethik,* der bei Datenanalysen aus gesellschaftlicher aber auch informatischer Sicht eine wichtige Rolle spielt, sowie mit *Datenanalyse und Maschinenlernen* und *Big Data* Aspekte, die im Kontext der Data Science eine neue Bedeutung in der Informatik erlangen.

Insbesondere *Datenanalyse und Maschinenlernen* sind in der Datenwissenschaft zentral und werden in allen betrachteten Dokumenten umfangreich thematisiert. Dabei liegt ein Schwerpunkt auf der *Auswahl, Beurteilung und Anpassung von Analysemodellen* (in 100 % der Studiengänge), dem *Prozess der Datenanalyse* (70 %), sowie verschiedenen Methoden, insbesondere des *überwachten Lernens* (80 %), wie Klassifikation, Entscheidungsbäume und Regression. Durch die stark analytisch geprägte Sichtweise dieses Bereichs stellt er gleichzeitig den Bezug zur zentralen Aufgabe von Datenwissenschaftlern dar, die sich weniger um die konkrete Datensammlung bzw. langfristige Speicherung kümmern, sondern eher um die Gewinnung neuer Informationen aus Daten durch Nutzung entsprechender Methoden, sowie um die Aufbereitung der Analyseergebnisse.

Der zweite besonders zentrale Bereich ist *Big Data:* Die Erfassung, Verarbeitung und Analyse großer und vielfältiger Datenmengen innerhalb kurzer

Tab. 1: Kategoriensystem zur Beschreibung der Inhalte der Data Science zusammen mit ihrer jeweiligen Abdeckung in den analysierten Studiengängen

	HdM Stuttgart	Univ. Salzburg	HS Albst.-Sigmar.	Univ. Stuttgart	LMU München	Univ. Marburg	Univ. Jena	TU Dortmund	HS Darmstadt	Beuth HS Berlin	Anz. Nennungen	% der Studiengänge
Mathematik												
Analysis				×		×	×			×	4	40 %
Lineare Algebra				×		×			×	×	4	40 %
Statistik	×	×	×	×	×	×	×	×	×	×	10	100 %
Informatische Grundlagen												
Algorithmen und Datenstrukturen		×		×		×	×	×			5	50 %
Betriebssysteme							×			×	2	20 %
IT-Security					×	×				×	3	30 %
Programmierung	×	×	×	×		×	×	×	×	×	9	90 %
Rechnerarchitektur						×					1	10 %
Rechnerkommunikation							×			×	2	20 %
Software Engineering				×		×				×	3	30 %
Theoretische Informatik				×		×					2	20 %
Verteilte Systeme										×	1	10 %
Datenanalyse und Maschinenlernen												
Analyseprozess	×	×	×		×		×		×	×	7	70 %
Datenvorverarbeitung			×			×				×	3	30 %
Ensemblelernen	×		×	×	×				×	×	6	60 %
Komplexere Methoden	×		×						×		3	30 %

Methoden überwachtes Lernen	×	×	×	×	×	×			×	×	8	80 %
Methoden unüberwachtes Lernen	×		×		×	×				×	5	50 %
Modellauswahl, -beurteilung, -anpassung	×	×	×	×	×	×	×	×	×	×	10	100 %
Text Mining	×		×	×						×	4	40 %
Visualisierung	×	×					×			×	4	40 %
Big Data												
Algorithmen und Methoden der Big-Data-Verarbeitung	×		×		×		×			×	5	50 %
Big-Data-Architekturen und -Systeme	×		×	×	×		×			×	6	60 %
Prinzipien der Big-Data-Analyse	×		×	×						×	4	40 %
Webdaten	×		×							×	3	30 %
Datenschutz, Ethik												
Datenschutz	×	×			×				×	×	5	50 %
Ethische Aspekte	×	×			×				×	×	5	50 %
Sicherheit	×			×	×				×	×	5	50 %
Datenspeicher												
Daten(bank)modellierung	×	×	×	×		×	×				6	60 %
Datenbanken (insb. relational)	×	×	×			×	×		×		6	60 %
Datenbanksprachen	×	×	×			×	×				5	50 %

Zeiträume stellt eine wesentliche Herausforderung für die Data Science dar. Daher ist es nicht verwunderlich, dass insbesondere *Systeme* (in 60 % der Studiengänge) und *Methoden* (50 %) zur Beherrschung großer Datenmengen eine wichtige Rolle spielen. Je nach Interpretation kann *Big Data* jedoch auch anderen Bereichen Fachgebieten, insbesondere dem Datenmanagement, zugeordnet werden. Dies liefert eine mögliche Erklärung für die vergleichsweise geringe Repräsentation in vielen Studiengängen und teils deutlichen Unterschiede.

Internationaler Vergleich

Zum internationalen Vergleich der Ergebnisse wurden drei Studiengänge ausgewählt, die klar definierte Inhalte vorweisen und somit für den Vergleich geeignet sind: Der an der University of Berkeley angebotene *Master of Information and Data Science* sowie die *Master of Science in Data Science* der University of Washington und der Columbia University. Auch hier wurden entsprechende Modulhandbücher bzw. andere Dokumente, die die Studiengänge und deren Inhalte klar beschreiben, in einer qualitativen Inhaltsanalyse untersucht. Dieser lagen dieselben Grundsätze wie zuvor zugrunde, jedoch mit dem Unterschied, dass kein neues Kategoriensystem induktiv aufgebaut, sondern, die zuvor aufgebaute Charakterisierung deduktiv an die Materialien herangetragen wurde, um einen Vergleich zu ermöglichen. Die Ergebnisse der Analyse sind in Tab. 2 abgebildet.

Bei Auswertung der drei Studiengänge wurden keine Begriffe ermittelt, die nicht sinnvoll in die zuvor entwickelte Charakterisierung eingeordnet werden konnten. Es kann daher eine hohe Vollständigkeit dieser Charakterisierung angenommen werden. Gleichzeitig zeigt sich, dass diese Studiengänge zwar nicht die identischen Schwerpunkte setzen, Abweichungen aber relativ gering sind und nur geringfügig unterschiedlichen Schwerpunktsetzungen entsprechen sowie dem unterschiedlichen Bildungssystem geschuldet sind. Gerade dieser letzte Aspekt zeigt, dass die getrennte Analyse sinnvoll war, da so unterschiedliche Charakteristika und Voraussetzungen klar erkennbar bleiben.

Es ist auch erkennbar, dass ähnliche mathematische und informatische Vorkenntnisse vorausgesetzt werden. Außerdem stellt auch in diesen Studiengängen der Bereich *Datenanalyse und Maschinenlernen* den zentralen Schwerpunkt dar.

Tab. 2: Einordnung dreier US-amerikanischer Curricula in das entwickelte Kategoriensystem

	Berkeley Univ.	Columbia Univ.	Univ. of Washington	Anz. der Nennungen	% der Studiengänge
Mathematik					
Analysis					
Lineare Algebra					
Statistik	×	×		2	67%
Informatische Grundlagen					
Algorithmen und Datenstrukturen		×		1	33%
Betriebssysteme					
IT-Security					
Programmierung	×	×		2	67%
Rechnerarchitektur					
Rechnerkommunikation					
Software Engineering					
Theoretische Informatik					
Verteilte Systeme					
Datenanalyse und Maschinenlernen					
Analyseprozess	×			1	33%
Datenvorverarbeitung		×	×	2	67%
Ensemblelernen	×	×	×	3	100%
Komplexere Methoden	×	×		2	67%
Methoden überwachtes Lernen	o	×	×	2	67%
Methoden unüberwachtes Lernen	o		×	1	33%
Modellauswahl, -beurteilung, -anpassung		×		1	33%
Text Mining					
Visualisierung			×	1	33%
Big Data					
Algorithmen und Methoden der Big-Data-Verarbeitung		×		1	33%
Big-Data-Architekturen und -Systeme	×	×	×	3	100%
Prinzipien der Big-Data-Analyse	×		×	2	67%
Webdaten		×		1	33%
Datenschutz, Ethik					
Datenschutz			×	1	33%
Ethische Aspekte			×	1	33%
Sicherheit					
Datenspeicher					
Daten(bank)modellierung		×	×	2	67%
Datenbanken (insb. relational)	×	×	×	3	100%
Datenbanksprachen		×	×	2	67%

4 Ausblick: Entwicklung eines Data-Literacy-Kompetenzmodells

Die beschriebene Arbeit schafft durch eine fundierte Charakterisierung der Data Science einen wichtigen Beitrag für zukünftige Forschungsarbeiten: Neben einer Fundierung der Studiengangs- und Curriculums(weiter)entwicklung, können die ermittelten Inhalte der Data Science auch zur Fundierung der fachdidaktischen Forschung beitragen. In den nächsten Jahren wird das heute noch relativ neue Thema *Data Literacy* vermutlich ein wichtiges Thema der fachdidaktischen Forschung werden. Dieser insbesondere aus dem Hochschulkontext stammende neue Begriff zur Beschreibung einer Sammlung an Grundkompetenzen zum fundierten und zielgerichteten Umgang mit Daten ist insbesondere durch Aspekte des Datenmanagements, das eine eher statische Sichtweise auf Daten beinhaltet, und der Data Science, die die dynamischen Aspekte berücksichtigt, geprägt.

Zur Charakterisierung dieser Kompetenzen ist ein fachlich fundiertes Kompetenzmodell unabdingbar, welches bis dato noch nicht existiert. Bisher werden hingegen entsprechende Kompetenzen oft aus eher bedarfsorientierter Sichtweise betrachtet (vgl. [Ri15]). Ein solches Modell kann insbesondere unter Rückgriff auf die hier beschriebenen Inhalte der Data Science und weitere existierende Arbeiten zum Datenmanagement entwickelt werden, da diese Bereiche die zentralen Ursprünge von Data-Literacy-Kompetenzen darstellen. Um sowohl praktische Kompetenzen, die dieser Bereich beinhaltet, als auch deren theoretische fachliche Fundierung zu berücksichtigen, kann die im Bereich der informatischen Bildung in Deutschland bereits bewährte Zweiteilung in Inhalts- und Prozessbereiche (vgl. [GI08]) übernommen werden. Zur Ableitung der Inhaltsbereiche dienen dabei die hier ermittelten Inhalte der Data Science gemeinsam mit den Schlüsselkonzepten des Datenmanagements (vgl. [GR17]), während die Prozessbereiche aus dem Datenlebenszyklus (vgl. [GR17]) bestimmt werden können, der die prozessorientierten Aspekte des Umgangs mit Daten widerspiegelt. Ein Prototyp für ein solches Kompetenzmodell der Data Literacy, der in der weiteren Forschung noch stärker zu fundieren und zu evaluieren ist, kann auf dieser Basis wie in Abbildung 1 abgebildet aussehen.

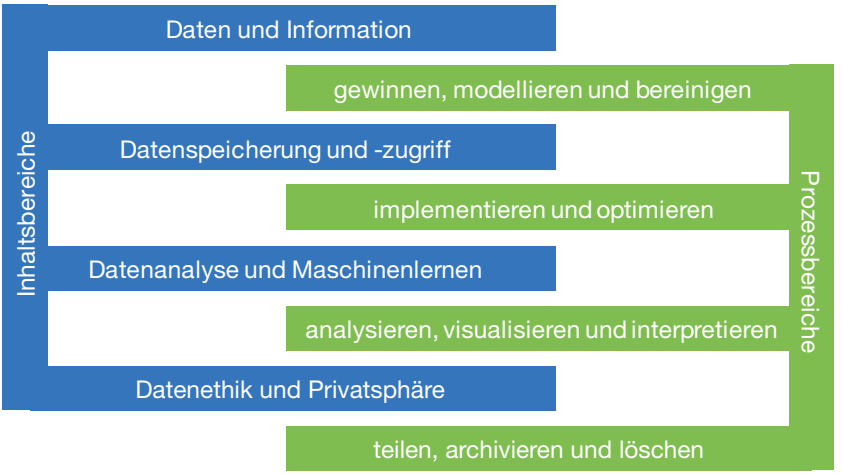

Abb. 1: Entwurf eines Data-Literacy-Kompetenzmodells

5 Zusammenfassung und Ausblick

Die durchgeführte Analyse gibt einen klaren Überblick über die Inhalte der Data Science und trägt zu ihrer Charakterisierung bei. Es zeigt sich, dass Data Science durch verschiedene informatische Aspekte geprägt ist, aber auch mathematische Aspekte nicht zu vernachlässigen sind. Aus informatischer Sicht kann die Data Science, basierend auf den in dieser Arbeit gewonnenen Ergebnissen, als Schnittmenge der beiden Themenbereiche *Datenbanken/Datenmanagement* sowie *Maschinenlernen/Datenanalyse* betrachtet werden: Das erste Gebiet berücksichtigt die eher statischen Aspekte, das zweite konzentriert sich auf die dynamischen. Zusätzlich zu diesen beiden Bereichen müssen jedoch auch Aspekte von *Big Data,* welches den speziellen Umgang mit großen und vielfältigen Datenmengen thematisiert, sowie die *Datenethik,* die gesellschaftliche und individuelle Wirkungen berücksichtigt, miteinbezogen werden. Durch die starke Verwurzlung der Data Science in der Informatik, die durch die hier beschriebene Arbeit expliziert wird, zeigt sich, dass die Datenwissenschaft ein Thema für die Forschung im Kontext der informatischen Bildung darstellt. Die hier beschriebene Explikation der Inhalte der Data Science kann einen Beitrag zu einer einheitlicheren Betrachtung dieser neuen Wissenschaft liefern, die bisher oft stark unterschiedlich geprägt ist, gleichzeitig aber auch zur gezielten Curriculumsentwicklung und zum Vergleich verschiedener Studiengänge in diesem Bereich eingesetzt werden.

Zusätzlich liefert sie Impulse für die weitere Forschung und kann, wie durch die skizzierte Entwicklung eines Data-Literacy-Kompetenzmodells gezeigt, als Basis für diese eingesetzt werden.

Literaturverzeichnis

[Ca17] Cao, L.: Data Science: A Comprehensive Overview. ACM Comput. Surv. 50/3, 43:1–43:42, Juni 2017.

[DBW17] Demchenko, Y.; Belloum, A.; Wiktorski, T.: EDISON Data Science Framework: Part 2. Data Science Body of Knowledge (DS-BoK) Release 2, 2017.

[DMB17] Demchenko, Y.; Manieri, A.; Belloum, A.: EDISON Data Science Framework: Part 1. Data Science Competence Framework (CF-DS) Release 2, 2017.

[GI08] Gesellschaft für Informatik e. V., AK Bildungsstandards: Grundsätze und Standards für die Informatik in der Schule: Bildungsstandards Informatik für die Sekundarstufe I. LOG IN 150/151, 2008.

[GI18] Gesellschaft für Informatik e. V., PAK Data Science: Data Literacy und Data Science Education: Digitale Kompetenzen in der Hochschulausbildung, 2018.

[GR17] Grillenberger, A.; Romeike, R.: Key Concepts of Data Management: An Empirical Approach. In (Suero Montero, C.; Joy, M.; Eds.): Proceedings of the 17th Koli Calling International Conference on Computing Education Research, Koli, Finland, November 16–19, 2017, ACM, New York, NY, USA, 2017.

[HTT09] Hey, T.; Tansley, S.; Tolle, K.: The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, 2009.

[Ma10] Mayring, P.: Qualitative Inhaltsanalyse: Grundlagen und Techniken. Beltz, 2010.

[NI15] NIST Big Data Public Working Group: NIST Big Data Interoperability Framework: Volume 1, Definitions, 2015.

[Ri15] Ridsdale, C.; Rothwell, J.; Smit, M. et al.: Strategies and Best Practices for Data Literacy Education: Knowledge Synthesis Report, 2015.

[Ve17] Veaux, R. D. D.; Agarwal, M.; Averett, M. et al.: Curriculum Guidelines for Undergraduate Programs in Data Science. Annual Review of Statistics and Its Application 4/1, S. 15–30, 2017.