Von Datenmanagement zu Data Literacy: Informatikdidaktische Aufarbeitung des Gegenstandsbereichs *Daten* für den allgemeinbildenden Schulunterricht

Dissertation

zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)
eingereicht am
Fachbereich Mathematik und Informatik
der Freien Universität Berlin

vorgelegt von
Andreas Grillenberger

Berlin, im November 2018

Erstgutachter:	Prof. Dr. Ralf Romeike, Freie Universität Berlin
Zweitgutachterin:	Prof. Dr. Ira Diethelm, Carl von Ossietzky Universität Oldenburg
Tag der Disputation:	25. Februar 2019

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig und ausschließlich unter Verwendung der angegebenen Hilfsmittel und Hilfen verfasst habe. Weiterhin versichere ich, dass diese Arbeit nicht bereits in einem früheren Promotionsverfahren eingereicht wurde.

Berlin, 06.03.2019 *Andreas Grillenberger*

Zusammenfassung

Die Thematisierung von *Daten* in der Informatik befindet sich seit über einem Jahrzehnt in einem Wandel, der nicht nur technische Neuerungen nach sich zieht, sondern auch eine umfassende Neubetrachtung der Erfassung, Speicherung und Nutzung von Daten verursachte und zur Bildung eines neuen umfassenden Fachgebiets *Datenmanagement* führte. Die Ausmaße dieser Entwicklung zeigen sich an der zunehmenden Verarbeitung komplexer Daten (*Big Data*), neuen Möglichkeiten zur Datenverarbeitung und -analyse (z. B. *Datenstromsysteme*, *Data Mining*) und nicht zuletzt an der Entstehung einer eigenen *Data Science*. Neben fachlichen Veränderungen unterliegt aber auch die gesellschaftliche Bedeutung von Daten einem Wandel: Daten stellen nicht mehr nur ein wichtiges und innovatives Thema der Informatik, sondern das zentrale Fundament der digitalen Gesellschaft dar.

Auch der Informatikunterricht konzentriert sich seit Jahren eher auf tradierte Aspekte des Fachgebiets, wie Datenbanken und Datenmodellierung, während neuere Themen allenfalls als Unterrichtskontext aufgegriffen werden. Um eine aqäquate Grundlage für den Unterricht zu diesen Themen zu schaffen, die langlebigen Aspekte der fachlichen Entwicklungen zu identifizieren und somit einen zukunftssicheren Informatikunterricht zu ermöglichen, ist eine umfassende informatikdidaktische Aufarbeitung essenziell. Somit eröffnet sich durch diese Veränderungen deutliches Potenzial, nicht nur für die Informatikdidaktik, sondern auch für die Unterrichtspraxis.

In dieser Arbeit wird daher der Gegenstandsbereich Daten und insbesondere das Fachgebiet Datenmanagement aus informatikdidaktischer Sicht umfassend aufgearbeitet, mit dem Ziel ein Fundament für die weitere Forschung und die Unterrichtspraxis zu schaffen. Dazu wird das Modell der Didaktischen Rekonstruktion als Forschungsrahmen eingesetzt und das Fachgebiet aus den Perspektiven *Fach*, *Lehrer*, *Schüler* und *Gesellschaft* untersucht. Als eines der zentralen Ergebnisse wird, basierend auf einem empirisch geprägten Ansatz, ein Modell der Schlüsselkonzepte des Datenmanagements entwickelt. Um den Bogen zu allgemeinbildenden Datenkompetenzen im Sinne einer Data Literacy zu spannen, entsteht außerdem ein Data-Literacy-Kompetenzmodell, das auf Grundlagen des Datenmanagements und der Data Science fundiert wird. Um die praktische Relevanz der Ergebnisse der Arbeit zu unterstreichen, wird auf Basis der gewonnenen Erkenntnisse die Umsetzung von Datenmanagement im Informatikunterricht skizziert. Dazu werden zwei Unterrichtswerkzeuge sowie eine Unterrichtssequenz entwickelt und erprobt.

Diese Arbeit schafft somit nicht nur eine Orientierung und Basis für die weitere Forschung im Kontext Daten, sondern sorgt durch die fachliche Klärung des Fachgebiets Datenmanagement auch dafür, dass dessen Kernaspekte greifbarer und klarer erkennbar werden. Sie zeigt exemplarisch, dass auch moderne und komplex erscheinende Themen des Datenmanagements unter Berücksichtigung der zugrundeliegenden Konzepte für den Unterricht geeignet aufbereitet werden können und betont die Relevanz dieser Themen, die in einer digitalen Gesellschaft und im Sinne der Schaffung einer Data Literacy zukünftig einen größeren Stellenwert im Informatikunterricht erlangen müssen.

Vorabveröffentlichung von Teilen dieser Arbeit

Teile dieser Arbeit wurde bereits vor Einreichung dieser Dissertation wörtlich oder sinngemäß veröffentlicht. Die entsprechenden Referenzen sind in der untenstehenden Tabelle zu dem jeweiligen Teil, Kapitel oder Abschnitt der Arbeit zugeordnet.

Mit Ausnahme einer Arbeit sind alle unten angegebenen Publikationen unter Erstautorschaft des Autors dieser Arbeit entstanden. Die Arbeit *Grillenberger, Przybylla und Romeike (2016)* entstand in gemeinsamer Erstautorschaft der beiden erstgenannten Autoren.

Kapitel/Abschnitt	**vorab veröffentlicht in**
Kapitel 2	Grillenberger und Romeike (2015): „Bringing the Innovations in Data Management to CS Education: An Educational Reconstruction Approach“ Grillenberger, Przybylla und Romeike (2016): „Bringing CS Innovations to the Classroom: a Process Model of Educational Reconstruction“
Abschnitt 3.2	Grillenberger und Romeike (2015): „Big Data im Informatikunterricht: Motivation und Umsetzung“ Grillenberger und Romeike (2018): „Datenmanagement als Thema für den Informatikunterricht“
Kapitel 4	Grillenberger und Romeike (2014): „A Comparison of the Field Data Management and its Representation in Secondary CS Curricula“
Abschnitt 5.2	Grillenberger und Romeike (2014): „Teaching Data Management: Key Competencies and Opportunities“ Grillenberger und Romeike (2015): „Big Data im Informatikunterricht: Motivation und Umsetzung“
Abschnitte 6.1, 6.2	Grillenberger und Romeike (2017): „What Teachers and Students Know about Data Management“
Teil III (insb. Kapitel 8)	Grillenberger und Romeike (2017): „Key Concepts of Data Management: An Empirical Approach“ Grillenberger und Romeike (2017): „Empirische Ermittlung der Schlüsselkonzepte des Fachgebiets Datenmanagement“
Kapitel 10	Grillenberger und Romeike (2017): „Real-Time Data Analyses in Secondary Schools Using a Block-Based Programming Language“ Grillenberger und Romeike (2018): „Datenmanagement als Thema für den Informatikunterricht“
Kapitel 11	Grillenberger und Romeike (2015): „Big Data im Informatikunterricht: Motivation und Umsetzung“ Grillenberger und Romeike (2015): „Big-Data-Analyse im Informatikunterricht mit Datenstromsystemen: Ein Unterrichtsbeispiel“ Grillenberger und Romeike (2015): „Analyzing the Twitter Data Stream Using the Snap! Learning Environment“ Grillenberger und Romeike (2017): „Real-Time Data Analyses in Secondary Schools Using a Block-Based Programming Language“
Kapitel 9	Grillenberger und Romeike (2018): „Developing a Theoretically Founded Data Literacy Competency Model“
Anhang E	Grillenberger und Romeike (2018): „Was ist Data Science? Ermittlung der informatischen Inhalte durch Analyse von Studienangeboten“

Inhaltsverzeichnis

Teil I: Einleitung und forschungsmethodische Einordnung

Teil II: Fachliche und fachdidaktische Grundlagen der Arbeit

Teil V: Abschluss

Verzeichnisse

Anhang

Teil I:

Einleitung und forschungsmethodische Einordnung

1 Einleitung

> *Our society is seriously conflicted about data. [...] But we use data every day—to choose medications or health practices, to decide on a place to live, or to make judgments about education policy and practice. The newspapers and TV news are full of data about nutrition, side effects of popular drugs, and polls for current elections. Surely there is valuable information here, but how do you judge the reliability of what you read, see, or hear?*
>
> *This is no trivial skill——and we are not preparing students to make these critical and subtle distinctions. (Rubin, 2005)*

Obwohl dieses Zitat bereits 13 Jahre alt ist, zeigt es eine auch heute noch zentrale Herausforderung: Der Begriff *Daten*, der seit jeher zentral für die Informatik ist, gewinnt zunehmend auch außerhalb dieser an Bedeutung. Daten sind nicht mehr nur ein Thema der wissenschaftlichen Auseinandersetzung in der Informatik, im Gegenteil stellen sie ein gesellschaftlich viel thematisiertes und hochbrisantes Thema dar, wie verschiedene Diskussionen, beispielsweise zu Datenschutzskandalen, einer Datenschutzgrundverordnung oder der massenhaften Erfassung von Daten über ganze Bevölkerungsgruppen zeigen. Um die in diesem Zusammenhang vorhandenen Möglichkeiten und Gefahren einschätzen zu können, aber auch um von einer passiven Rolle im Umgang mit Daten zu einer aktiven zu wechseln und Daten nicht mehr nur zu produzieren, sondern auch zu nutzen, wird heute immer häufiger gefordert, dass jeder mündige Bürger *data literate*[1] sein soll, d. h. grundlegende Kenntnisse und Kompetenzen im Umgang mit und der Verarbeitung von Daten mitbringen soll.

Aus fachlicher Sicht stellen Daten ein grundlegendes Fundament aller informationsverarbeitenden Prozesse dar und sind somit ein wichtiger Zugang zur Informatik. Während Daten bislang außerhalb der Informatik hauptsächlich als technische Notwendigkeit wahrgenommen wurden, verändert sich diese Wahrnehmung in Zusammenhang mit modernen Entwicklungen in der Datenspeicherung und -analyse in den letzten Jahren drastisch: *„The world's most valuable resource is no longer oil, but data"* titelt beispielsweise die Zeitschrift *The Economist (2017)*. Daten werden heute, im Rahmen der zunehmenden Digitalisierung aller Lebensbereiche, immer häufiger als wertvoller Rohstoff wahrgenommen. Nur wenn Informationen geeignet auf Daten abgebildet werden (hierfür wurde modern der Begriff „Datafizierung" geprägt *(vgl. Cukier und Mayer-Schönberger, 2017)*), können diese dauerhaft sicher gespeichert sowie zielführend und effizient verarbeitet werden. Die Gewinnung neuer Informationen und neuen Wissens aus einem Berg von Daten, das sogenannte „Data Mining" (vgl. auch Abschnitt 3.2.2), wird heute in verschiedenen Kontexten zunehmend als wichtig erachtet und verbreitet sich stetig weiter. Entsprechend sind Daten heute beim alltäglichen Umgang mit Informatiksystemen aber auch mit einfachen elektrischen Geräten,

[1] In dieser Arbeit wird durchgängig der englische Begriff *Data Literacy* verwendet, da sich eine deutsche Übersetzung bisher nicht etablieren konnte.

bei denen oft kaum erkennbar ist, dass sie ein Informatiksystem sind, nicht mehr wegzudenken und betreffen gleichzeitig immer stärker den eigenen Einflussbereich der Nutzer, die Daten häufig unbewusst oder unreflektiert weitergeben. Gleichzeitig ist heute jeder jederzeit und überall Produzent großer Datenmengen. Von den neuen Möglichkeiten, die sich damit eröffnen, profitieren bislang insbesondere Unternehmen, während die meisten Menschen im Kontext der Datenverarbeitung eine eher passive Rolle einnehmen und somit insbesondere Gegenstand der Analyse sind und meist nur sekundär von dieser profitieren. In verschiedenen Diskussionen *(z. B. Gillmor, 2014)* zeigt sich eine gewisse Unzufriedenheit mit dieser Situation. Jedoch sind Menschen in diesem Kontext – obwohl es teils anders erscheint – nicht prinzipiell ohnmächtig: Datenverarbeitende Prozesse und Produkte sind heute zwar häufig noch durch große Konzerne geprägt, stellen gleichzeitig aber nicht mehr deren alleinige Domäne dar. Trotz der oft komplexen Grundlagen, auf denen die Möglichkeiten und das Potenzial der Verarbeitung von Daten beruhen, kann heute prinzipiell jeder eine aktivere Rolle im Umgang mit diesen einnehmen. Immer häufiger gibt es Möglichkeiten, von Daten auch persönlich zu profitieren, beispielsweise werden heute große Mengen an Daten als „Open Data“[2] veröffentlicht. Im privaten Umfeld werden diese jedoch bisher kaum genutzt. Durch die zunehmende Relevanz dieser Thematik, nicht nur in der Wissenschaft, sondern in allen Lebensbereichen, wirken die Entwicklungen im Kontext der Datengenerierung, -verarbeitung und -analyse heute als zentrale Grundlage unserer digitalen Gesellschaft. Um die damit einhergehenden Entwicklungen und Möglichkeiten verstehen und nutzen zu können, aber auch zur Förderung eines kritischen Weltbildes, ist ein grundlegendes Verständnis der Funktionsprinzipien dieser Möglichkeiten und die Entwicklung entsprechender Kompetenzen in diesem Kontext, wie sie im Sinne einer *Data Literacy* gefordert werden, unabdingbar.

Die informatische Fundierung dieser Diskussion findet insbesondere durch die Forschung im Fachgebiet *Datenmanagement* statt: Dieses stellt seit vielen Jahren ein wichtiges und innovatives Forschungsfeld der Informatik dar, das durch diverse Entwicklungen Einfluss auf uns als Menschen und unseren Alltag nimmt, und auch außerhalb der Informatik immer stärker wahrgenommen wird. Trotzdem haben modernere Inhalte dieses Fachgebiets bis heute kaum Bedeutung im Informatikunterricht erlangt und wurden für Jahrzehnte in der informatikdidaktischen Diskussion allenfalls als Randthema betrachtet. Seit der Etablierung des Datenbankunterrichts in den 1990er Jahren *(vgl. Witten, 1994; Borg, 1987; Lück, 1990)* konzentriert sich der Unterricht in diesem Bereich überwiegend auf dieselben Themen. Dabei liegt der Schwerpunkt sowohl national als auch international typischerweise auf dem Zusammenhang von Information und Daten, der relationalen Datenmodellierung (teils unter Zuhilfenahme eines objektorientierten Modells), relationalen Datenbanksystemen und der Durchführung von Datenbankabfragen mithilfe von SQL (vgl. z. B. *Buttke und Engelmann (2007), Brichzin et al. (2007), Arbeitskreis Bildungsstandards (2008) und Hubwieser (2007)* bzw. Abschnitt 4.3). Weitere Aspekte, die in den letzten Jahren hinzukamen, wie Datenschutz oder Datensicherheit, werden je nach Schule und Schulart mehr oder weniger

[2] „*Open data and content can be freely used, modified, and shared by anyone for any purpose*“ *(Open Knowledge International, 2017)*. Unter diesem Stichwort werden heute große Datenmengen, insbesondere von öffentlichen Einrichtungen und Behörden, frei verfügbar und verwendbar veröffentlicht.

vertieft thematisiert. Die Möglichkeiten modernen Datenmanagements und die Einflüsse, die diese auf unseren Alltag haben, bleiben jedoch mit Ausnahme von Einzelfällen außen vor und werden höchstens als motivierender Kontext genutzt. Dies dürfte sicherlich auch durch die lange zurückliegende fachdidaktische Diskussion begründet sein: Für lange Zeit wurden in wenigen Artikeln allenfalls neue Unterrichtsideen oder Konzepte, welche tradierte Themen in diesem Kontext aufgreifen, vorgestellt, jedoch ohne neue Themen bzw. Entwicklungen stärker zu berücksichtigen *(vgl. Antonitsch, 2007; Bierschneider-Jakobs, 2004)*. Erst in den letzten Jahren, parallel zu dieser Arbeit, nahm dieser Themenbereich auch für den Schulunterricht Fahrt auf: Insbesondere wurden vermehrt Ansätze präsentiert, die versuchen, Aspekte modernen Datenmanagements für den Unterricht aufzubereiten *(vgl. Buffum et al., 2014; Dryer, Walia und Chattopadhyay, 2018)*. Solche werden im Informatikunterricht jedoch auch weiterhin höchstens als Randthema und meist nur fakultativ angerissen (vgl. Kapitel 6).

Es ist folglich nicht nur eine klare Lücke im derzeitigen Schulunterricht erkennbar, sondern auch ein Mangel an fachdidaktischer Diskussion zu modernen Aspekten aus dem Bereich der Daten und des Datenmanagements: Die Fragen, ob modernere Datenmanagementthemen überhaupt als Unterrichtsthema geeignet sind, was aus diesem Bereich wichtig/zentral ist, welche Kompetenzen die Schülerinnen und Schüler in diesem Bereich erwerben sollten und wie diese Themen praktisch für den Unterricht umgesetzt werden können, wurden aus informatikdidaktischer Sicht bisher nicht geklärt. Diese und weitere damit zusammenhängende Fragen werden daher in dieser Arbeit aufgegriffen.

1.1 Ziele der Arbeit

Trotz der bisher fehlenden fachdidaktischen Diskussion des Fachgebiets *Datenmanagement*, scheint die Annahme gerechtfertigt, dass dieses vielfältige informatische Konzepte und Ideen enthält, die den Informatikunterricht an Sekundarschulen[3] aus allgemeinbildender Sicht bereichern können, in diesem umsetzbar sind und zugleich eine wichtige Grundlage für die Schaffung einer Data Literacy darstellen. Aufgrund des Umfangs des Fachgebiets und seiner stetigen Weiterentwicklung, ist dieses bisher jedoch insbesondere für Lehrkräfte schwer überblickbar, sodass auch dessen allgemeinbildende Aspekte bisher im Unterricht unberücksichtigt bleiben. Um diese zu beleuchten, ist ein Ziel dieser Arbeit die Schaffung einer Grundlage für die weitere informatikdidaktische Diskussion und den Unterricht in diesem Bereich. Dazu werden die umfangreichen Einflüsse des Fachgebiets auf den Alltag herausgestellt, dieses aus informatikdidaktischer Sicht aufgearbeitet und durch zentrale Konzepte, Prinzipien und Praktiken charakterisiert. Um neben den theoretischen Aspekten dieser Arbeit auch direkt die Anwendung für den Unterricht darzustellen und Erkenntnisse für diesen zu gewinnen, werden konkrete Ideen zur Integration zentraler Aspekte des Datenmanagements in den Informatikunterricht (mit Fokus auf die Sekundarstufe I

[3] Wenn in dieser Arbeit von Sekundarschulen gesprochen wird, schließt dies alle Schulen im Bereich der Sekundarstufe I/II mit ein. Eine Beschränkung auf (ehemalige) Haupt- und Realschulen, wie in verschiedenen deutschen Bundesländern üblich, findet hier explizit nicht statt.

und II) aus der fachdidaktischen Diskussion abgeleitet, vorgestellt und ein Unterrichtskonzept im Unterrichtsversuch evaluiert. Auf diese Weise soll gleichzeitig die Umsetzbarkeit verschiedener Themen in der Schule demonstriert, aber auch Lösungsideen für typische Herausforderungen, die bei der Thematisierung im Unterricht auftreten, gegeben werden.

Die vorliegende Arbeit beschäftigt sich daher mit folgenden Forschungsfragen:

RQ1) Welche Einflüsse haben die Entwicklungen der letzten Jahre im Bereich des Datenmanagements auf den Umgang mit und die Bedeutung von Daten in Informatik, Alltag und Beruf?

RQ2) Welche Bedeutung haben das Fachgebiet Datenmanagement bzw. mit diesem in Zusammenhang stehende Themen bereits im Informatikunterricht?

RQ3) Wie sind Vorwissen und Erfahrungen von Schülerinnen und Schülern zu dem Datenmanagement zugehörigen Themen ausgeprägt?

RQ4) Welche Unterstützung benötigen Lehrkräfte bei der Integration von Aspekten des Datenmanagements bzw. der Data Literacy in ihren Unterricht?

RQ5) Welche sind die zentralen Konzepte und Praktiken des Fachgebiets Datenmanagement, insbesondere in Hinblick auf den Informatikunterricht an Sekundarschulen?

RQ6) Welche Struktur liegt den für einen kritischen und verantwortungsbewussten Umgang mit Daten im Sinne einer Data Literacy allgemein notwendigen Kompetenzen zugrunde?

RQ7) Inwiefern ist es möglich, im Informatikunterricht grundlegende Data-Literacy-Kompetenzen herauszubilden?

1.2 Struktur der Arbeit

Die Struktur der vorliegenden Arbeit orientiert sich stark an den beschriebenen Forschungsfragen, die als roter Faden durch diese Arbeit dienen. Vorbereitend wird jedoch, nach diesen einleitenden Gedanken und der Motivation für die Arbeit, zunächst eine forschungsmethodische Einordnung vorgenommen. Außerdem wird in Kapitel 2 das *Modell der Didaktischen Rekonstruktion für den Informatikunterricht*, das dieser Arbeit als Forschungsrahmen zugrunde liegt, skizziert und dessen konkrete Umsetzung in dieser Arbeit erläutert.

Im zweiten Teil der Arbeit werden zuerst die fachlichen und fachdidaktischen Grundlagen für das weitere Vorgehen geschaffen. Dazu werden in Abschnitt 3.1 die Begriffe *Datenmanagement*, *Data Science* und *Data Literacy*, die heute häufig in ähnlichem Kontext aber mit unterschiedlicher Bedeutung verwendet werden, voneinander abgegrenzt und die in dieser Arbeit erfolgte Schwerpunktsetzung auf Datenmanagement vor diesem Hintergrund erläutert. Daraufhin wird, zur fachlichen Fundierung der weiteren Arbeit und zur

Beantwortung der ersten Forschungsfrage, die Bedeutung und Entwicklung von Datenmanagement in den letzten Jahrzehnten aus fachlicher Perspektive untersucht (Abschnitt 3.2) und ein Überblick über dessen historische Entwicklung und zentrale Themen aus diesem gegeben, die im Folgenden eine wichtige Rolle spielen werden. Obwohl es für eine Arbeit aus der Fachdidaktik ungewöhnlich ist, findet diese Grundsteinlegung aus informatischer Sicht noch vor der Einordnung in die fachdidaktische Forschung statt, die sich erst in Kapitel 4 anschließt, da auch zu deren Verständnis eine ausreichende fachliche Fundierung und ein Überblick über das Forschungsgebiet bereits zentral sind. Gemeinsam mit der Darstellung des Forschungsstandes wird auch die aktuelle Umsetzung von Themen des Datenmanagements im Informatikunterricht beschrieben und deren Bedeutung in aktuellen Curricula und Bildungsstandards untersucht (RQ2). In Kapitel 5 werden verschiedene Themen des Fachgebiets in der Erfahrungswelt der Lernenden durch konkrete informatische Phänomene verankert sowie die gesellschaftliche und berufliche Bedeutung von Datenmanagement dargestellt (RQ1). In Kapitel 6 wird anschließend die Ausgangssituation für diese Arbeit untersucht: Als Vorbereitung für die unterrichtspraktischen Untersuchungen, wird in Abschnitt 6.1 untersucht, welche Kenntnisse Lehrkräfte zu verschiedenen Datenmanagementthemen ihrer eigenen Einschätzung nach bereits haben, wie relevant sie diese für den Unterricht einschätzen und welche Hindernisse sie bei der Integration dieser in den Unterricht sehen (RQ4). Als zweite den Unterricht prägende Perspektive wird die der Schülerinnen und Schüler in Abschnitt 6.2 untersucht, indem diese im Rahmen einer Fragebogenstudie hinsichtlich ihres Vorwissens und ihrer Erfahrungen mit ausgewählten Aspekten des Datenmanagements befragt werden (RQ3).

Basierend auf diesem Fundament beschäftigt sich der dritte Teil der Arbeit mit der fachdidaktischen Aufarbeitung des Fachgebiets. Um den fachlichen Kern des Datenmanagements zu identifizieren (RQ5), wird der Fokus zuerst auf die Ermittlung der Schlüsselkonzepte des Fachgebiets gelegt. Entsprechend wird in Kapitel 7 Bezug zu Arbeiten genommen, die als Grundlage und Orientierung für eine Erforschung dieser Schlüsselkonzepte dienen können, beispielsweise die „fundamentalen Ideen der Informatik“ *(Schwill, 1993)* und die „Great Principles of Computing“ *(Denning, 2003b)*. Anschließend wird in Kapitel 8 ein Ansatz zur Erforschung der Schlüsselkonzepte entwickelt, der Aspekte unterschiedlicher Vorarbeiten und Kritikpunkte an diesen aufgreift. Durch Anwendung dieses Ansatzes wird ein Modell der Schlüsselkonzepte des Datenmanagements systematisch erstellt und dieses diskutiert. Darauf aufbauend wird in Kapitel 9 anschließend als letzter theoretischer Teil ein Kompetenzmodell der Data-Literacy entwickelt (RQ6), um den in jüngerer Zeit häufiger gestellten Forderungen nach datenbezogenen Grundkompetenzen Rechnung zu tragen.

Der vierte Teil der Arbeit greift die in den vorherigen Kapiteln gewonnenen Erkenntnisse auf und bezieht diese auf den Informatikunterricht an Sekundarschulen (RQ7). Dazu werden verschiedene Herausforderungen für den Informatikunterricht aufgegriffen und mögliche Herangehensweisen an diese vorgestellt: In Kapitel 10 werden potenzielle Datenquellen für einen adäquaten Datenmanagementunterricht diskutiert und in Kapitel 11 die Entwicklung von zwei Werkzeugen für den Informatikunterricht zu einem beispielhaft ausgewählten Thema des Datenmanagements, den Datenstromsystemen, beschrieben. Die

bei der Evaluation des Werkzeugs identifizierten Herausforderungen gehen ein in die in Kapitel 12 dargestellte Entwicklung und Erprobung eines Unterrichtskonzepts zum Thema Data Mining. In einer Synthese der in der Erprobung gewonnenen Erfahrungen werden Leitlinien für den Datenmanagementunterricht abgeleitet.

Abschließend wird die Arbeit in Kapitel 13 zusammengefasst.

2 Das Modell der Didaktischen Rekonstruktion als Forschungsframework

In der fachdidaktischen Forschung existieren verschiedene anerkannte Forschungsrahmenmodelle, die unterschiedliche Ziele verfolgen: Beispielsweise legt die *Implementationsforschung* ein besonderes Augenmerk auf die Überprüfung *„ob bzw. unter welchen Umständen sich die Ergebnisse in der Praxis realisieren lassen und welche Wirkungen und Nebenwirkungen dies hat“ (Gräsel und Parchmann, 2004)*; die *Entwicklungsforschung* beschäftigt sich mit der qualitätssteigernden Weiterentwicklung von Unterricht *(Gesellschaft für Fachdidaktik, 2015)*; und die *Didaktische Rekonstruktion* beschäftigt sich mit der Aufbereitung von Inhalten für den Unterricht unter gleichwertiger Einbeziehung der Fach- und der Schülerperspektive auf ein Thema *(Kattmann et al., 1997)*. Obwohl gerade in der Informatik eine besonders hohe Innovationsdichte und damit in der Informatikdidaktik der Bedarf für ein anerkanntes Vorgehensmodell zur didaktischen Untersuchung und Aufbereitung neuer Themen vorherrscht, deutet sich in diesem Bereich bislang kein entsprechender Konsens für ein bestimmtes Forschungsrahmenmodell an. Am geeignetsten für eine derartige Forschung erscheint derzeit jedoch das bereits erwähnte *Modell der Didaktischen Rekonstruktion*: In verschiedenen Arbeiten, insbesondere im Rahmen des Promotionsprogramms *ProDid*[4] der *Carl von Ossietzky Universität Oldenburg*, wurde dieses Modell bereits in unterschiedlichen, unter anderem naturwissenschaftlichen, Fächern vielfältig eingesetzt. Dadurch konnte gezeigt werden, dass die Didaktische Rekonstruktion für die Aufbereitung neuer Themen für den Unterricht geeignet ist und in unterschiedlich umfangreichen Kontexten und mit verschiedenen verfolgten Zielen erfolgreich genutzt werden kann. Obwohl das Modell in der Informatikdidaktik bisher eher geringe Beachtung gefunden hat (vgl. Abschnitt 2.2), scheint es vielversprechend, dieses auch in der vorliegenden Arbeit als Forschungsrahmen heranzuziehen, da es die Berücksichtigung vielfältiger Perspektiven bei der Aufarbeitung eines Themas bzw. Themenbereichs für den Unterricht nahelegt.

2.1 Ursprünge und Entwicklung der Didaktischen Rekonstruktion

Die Didaktische Rekonstruktion ist ein ursprünglich aus der Biologie- und Physikdidaktik stammendes Forschungsformat *(vgl. Kattmann et al., 1997)*. Dieses geht bei der Aufbereitung eines Themas für den Unterricht, im Gegensatz zur *Didaktischen Reduktion* oder *Didaktischen Transposition*, nicht rein von den fachlichen Inhalten aus, sondern betrachtet die Aufbereitung eines Themas für den Unterricht als Wechselspiel zwischen der fachlichen Sichtweise und den Anforderungen der Lernenden: *„Mit dem Modell der Didaktischen Rekonstruktion werden fachliche Vorstellungen, wie sie in Lehrbüchern und anderen wissenschaftlichen Quellen*

[4] `http://www.uni-oldenburg.de/diz/promotionsprogramme/prodid-didaktische-rekonstruktion/`

Ausdruck finden, mit Schülerperspektiven so in Beziehung gesetzt, daß daraus ein Unterrichtsgegenstand entwickelt werden kann.“ *(Kattmann et al., 1997)* Im fachdidaktischen Triplett (vgl. Abbildung 2.1) stellen *Kattmann et al. (1997)* das Zusammenspiel von *fachlicher Klärung* und *Schülerperspektive* auf ein Thema sowie die darauf aufbauende *didaktische Strukturierung* des Lerninhalts dar. Ein zentraler Aspekt dieses Modells ist die explizite Gleichwertigkeit von Schülerperspektive und fachlicher Perspektive auf das Thema. Von den Autoren wird angestrebt, „*die Vermittlung von Wissensbeständen und die damit verbundenen pädagogischen Aspekte in ein Gleichgewicht zu bringen*“ *(Kattmann, 2007)*. Ein Großteil der bisher basierend auf diesem Modell entstandenen Arbeiten *(z. B. Hörsch, 2007; Rutke, 2007; Schwanewedel, 2010; Kraynova, 2012)* konzentriert sich dabei auf den für dieses Modell speziellen Bereich der Schülerperspektive und dabei, genau wie die ursprüngliche Arbeit von *Gropengießer (1997)*, insbesondere auf die Ermittlung von Schülervorstellungen zum Thema. Diese Fokussierung ist vermutlich dem geschuldet, dass die jeweils betrachteten Themen relativ stabil sind und daher sowohl eine fachliche Aufarbeitung bereits stattgefunden hat, gleichzeitig aber auch wenige Veränderungen zu erwarten sind, die von Grund auf neu betrachtet werden müssten.

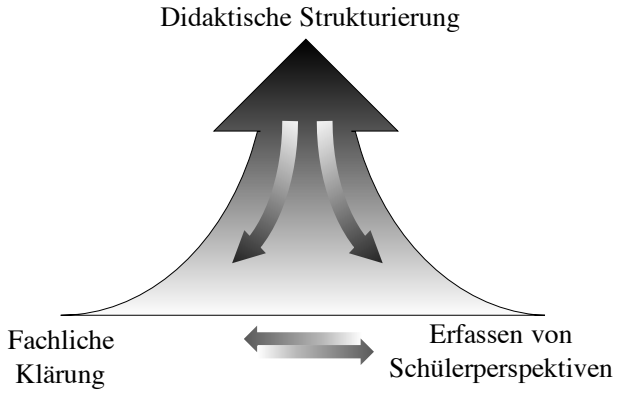


Abbildung 2.1: Didaktisches Triplett nach *Kattmann et al. (1997)*.

Trotz der weiten Verbreitung in den Naturwissenschaftsdidaktiken war das Modell der Didaktischen Rekonstruktion in der Informatikdidaktik relativ lange eher unbekannt und wurde kaum genutzt. Dies lässt sich insbesondere aufgrund des unterschiedlichen Charakters der Informatik erklären, die sich durch ihre hohe Innovationsdichte in einem steten Wandel befindet, der wesentlich stärker ist als in den traditionellen Naturwissenschaften, die auf einen relativ gefestigten Kern zurückgreifen können, der auch von innovativen Erkenntnissen kaum verändert, sondern eher ergänzt wird. Um diesem speziellen Charakter der Informatik Rechnung zu tragen, wurde das Modell der Didaktischen Rekonstruktion durch *Diethelm, Dörge et al. (2011)* für die Anwendung in der Informatikdidaktik angepasst und dabei um mehrere Perspektiven erweitert. Insbesondere wurde die Lehrerperspektive aufgenommen, die *Diethelm, Dörge et al. (2011)* als zentral erachten: *„Da Informatik eine sehr junge Tradition als Schulfach hat, immer noch keine Einigkeit über den allgemeinbildenden Anteil von Informatik (zumindest außerhalb der GI) herrscht und dadurch auch die Lehrerbildung in Informatik in Deutschland noch sehr heterogen verläuft, erachten wir gerade für die Informatik*

diesen Aspekt ebenfalls für sehr wichtig.“ Neben der Lehrerperspektive wurden in der *Didaktischen Rekonstruktion für den Informatikunterricht* noch zwei weitere Bereiche ergänzt: Bei der *Klärung gesellschaftlicher Ansprüche an das Fach* gilt es, den allgemeinbildenden Charakter der zu rekonstruierenden Inhalte herauszustellen und deren gesellschaftliche Relevanz zu evaluieren und zu beachten. Die *Auswahl informatischer Phänomene* trägt hingegen zu einem – bisher in der Informatik noch relativ wenig etablierten – phänomenorientierten Unterricht bei, dessen Ziel es ist, „*den Schülern eine Sicht auf die Wirklichkeit der Welt (vgl. [Kla91]) aus dem Blickwinkel des Faches zu erschließen um damit die Phänomene des Alltags erklären zu können*“ *(Diethelm, Dörge et al., 2011)*. Das durch diese Anpassungen entstandene *Modell der didaktischen Rekonstruktion für den Informatikunterricht* ist in Abbildung 2.2 dargestellt. Die Pfeile entsprechen dabei der gegenseitigen Beeinflussung der verschiedenen Modellbereiche, sodass dieses Modell als Beziehungsmodell aufgefasst werden kann.

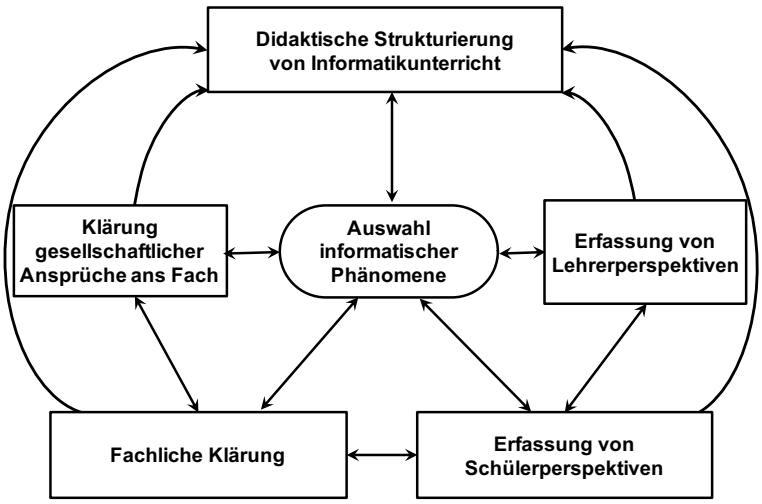


Abbildung 2.2: Modell der didaktischen Rekonstruktion für den Informatikunterricht nach *Diethelm, Dörge et al. (2011).*

2.2 Anwendung in der fachdidaktischen Forschung

In der fachdidaktischen Forschung findet eine Nutzung des Modells der Didaktischen Rekonstruktion insbesondere im Rahmen des Promotionsprogramms der Universität Oldenburg statt. Verschiedene dort entstandene Arbeiten wurden in der eng damit verbundenen Schriftenreihe *Beiträge zur Didaktischen Rekonstruktion* veröffentlicht. In dieser erschienen mittlerweile 46 Bände aus verschiedenen Disziplinen, die das Modell der Didaktischen Rekonstruktion auf unterschiedliche Weise für ihre Forschung nutzen. Eine Analyse von 25 Arbeiten[5] zeigt, dass ein Großteil davon das primäre Ziel hat, ein (neues oder tradiertes) Thema für den Unterricht aufzubereiten und, im Sinne des Modells, die verschiedenen den

[5]Es wurden die letzten 25 Bände der Schriftenreihe gewählt, von denen entweder ein Inhaltsverzeichnis oder die Arbeit selbst verfügbar war.

Unterricht prägenden Perspektiven miteinzubeziehen. Dabei ist klar erkennbar, dass die Mehrheit der Arbeiten sich bei der Schülerperspektive[6] rein auf die Ermittlung von Schülervorstellungen konzentriert (70 %). Auch allgemein ist eine klare Schwerpunktsetzung auf die Schülerperspektive erkennbar: Im Durchschnitt über die betrachteten Arbeiten befassen sich ca. 14 % der jeweiligen Arbeit mit der fachlichen Klärung, 30 % mit der Schülerperspektive und 10 % mit der didaktischen Strukturierung, während die anderen Bereiche keiner der Modellperspektiven klar zugeordnet werden konnten.

Aufgrund seiner Abstammung aus den Naturwissenschaften ist das Modell der didaktischen Rekonstruktion insbesondere in diesen verbreitet. Daher verwundert es nicht, dass sich über die gesamte Schriftenreihe hinweg 60 % der Arbeiten mit Themen aus einer Naturwissenschaft beschäftigen, dabei stammen 30 % aus der Biologie und 13 % aus der Chemie. Auch andere Bereiche sind vertreten, darunter fünf Arbeiten (≈10 %) aus der Geschichte, fünf (≈10 %) aus dem Komplex der Pädagogik, Psychologie und nicht-fachspezifischen Didaktik, sowie drei (≈6,5 %) aus dem Sachunterricht. Die Informatik ist jedoch in dieser Schriftenreihe bisher nicht vertreten.

Doch auch in der Informatikdidaktik existieren Arbeiten, die auf das Modell der Didaktischen Rekonstruktion zurückgreifen oder Aspekte dieses Modells aufgreifen. Insbesondere sind dabei Arbeiten zu nennen, die verschiedene Bereiche des Modells (relativ isoliert) betrachten, unter anderem durch Erforschung der Schülerperspektive *(z. B. Diethelm, Wilken und Zumbrägel, 2012)*. Aber auch zu anderen Modellbereichen existieren entsprechende Arbeiten, die sich jedoch nicht zwingend selbst im Model situieren: Beispielsweise wurden durch *Diethelm, Borowski und Weber (2010)* Kontexte für den Informatikunterricht identifiziert und von *Borowski, Diethelm und Wilken (2016)* in einer Schülerbefragung Bereiche ermittelt, zu denen Schülerinnen und Schüler mehr wissen möchten und die (ohne dass dies explizit genannt wird) auf eine Reihe von Phänomenen hindeuten. Weitere Forschungsbeispiele in diesem Zusammenhang werden im Rahmen der Entwicklung des Modells der Didaktischen Rekonstruktion für den Informatikunterricht von *Diethelm, Hubwieser und Klaus (2012)* genannt. Ähnlich umfangreich wie in den anderen Fachdidaktiken wurde dieses Modell hingegen in der Informatikdidaktik bisher nur durch *Stoffers (2016)* angewandt: Um die subjektiven Theorien zu ermitteln, die Lehrerinnen und Lehrer zur fachdidaktischen Strukturierung ihres Informatikunterrichts mitbringen, wurde die von *Komorek und Prediger (2013)* vorgeschlagene Adaption der Didaktischen Rekonstruktion für die Lehrerbildung als Basis genutzt und die verschiedenen Perspektiven des Modells[7] mit einem Fokus auf die empirische Untersuchung der subjektiven Überzeugungen der Lehrkräfte betrachtet.

[6] In Arbeiten die das Modell zur Lehrerbildung anwenden und sich somit mit Lehrer- anstatt von Schülervorstellungen beschäftigen, wurden diese Lehrervorstellungen als Alternative gezählt.

[7] Im Modell der didaktischen Rekonstruktion für die Lehrerbildung *(Komorek und Prediger, 2013)* wurde das ursprüngliche Modell nach *Kattmann et al. (1997)* so angepasst, dass dessen Fokus nicht auf Schulunterricht, sondern auf der Lehrerbildung liegt. Die drei Perspektiven des Modells wurden daher wie folgt abgewandelt: *Klärung domänenspezifischer fachdidaktischer Konzeptionen*, *Empirische Untersuchung subjektiver Überzeugungen von Lehrkräften zur fachdidaktischen Strukturierung* und *Entwicklung von Leitlinien für die Unterrichtsstrukturierung als Basis für die Lehrerbildung*

2.3 Anwendung des Modells als Forschungsrahmen

Im Vergleich zu den bisherigen auf dem Modell der Didaktischen Rekonstruktion basierenden Arbeiten, werden in dieser Arbeit verschiedene Adaptionen durchgeführt, insbesondere da eine andere Zielsetzung als üblich verfolgt wird: Statt ein einzelnes Thema, zu dem eine fachliche Klärung für den Unterricht oft schon erfolgt ist, für diesen unter Berücksichtigung der weiteren Perspektive(n) des Modells neu aufzubereiten, wird hier ein kompletter Themenkomplex von Grund auf betrachtet. Eine ausgearbeitete fachdidaktische Ausgangslage ist dabei nahezu nicht existent (vgl. Kapitel 4). Da das betrachtete Fachgebiet außerdem von vielen Innovationen geprägt und noch relativ jung ist, zwar einige von dessen Begrifflichkeiten im Alltag verwendet werden, aber trotzdem eher vage bleiben und auch dessen Umfang und Grenzen eher unklar sind, muss für eine tiefgreifende und zielgerichtete Betrachtung sowohl der Schüler- als auch der Lehrerperspektive zuvor eine umfangreiche fachliche Klärung erfolgen. Da es sich um ein Themengebiet der Informatik handelt und daher die von *Diethelm, Dörge et al. (2011)* im Rahmen der Begründung des Modells der Didaktischen Rekonstruktion für den Informatikunterricht argumentierten Herausforderungen (vgl. oben) auch hier zutreffen, wird als Forschungsrahmen für diese Arbeit nicht das ursprüngliche Modell der Didaktischen Rekonstruktion nach *Kattmann et al. (1997)*, sondern das für die Informatik angepasste Modell nach *Diethelm, Dörge et al. (2011)* herangezogen. Weil sich diese Arbeit als Grundlagenarbeit zum Thema Datenmanagement im allgemeinbildenden Informatikunterricht versteht, wurde der Fokus auf die grundlegende Aufarbeitung des Themengebiets und die Schaffung einer Basis für die weitere Forschung auf diesem Gebiet gelegt. Die anderen Bereiche des Modells wurden daher weniger detailliert betrachtet. Eine solche Schwerpunktsetzung wird, trotz der im Modell betonten Gleichwertigkeit der unterschiedlichen Phasen, als zulässig erachtet, da nicht alle Perspektiven gleichermaßen in einer Arbeit bearbeitet werden können. Eine derartige Schwerpunktsetzung findet auch in anderen Arbeiten basierend auf diesem Modell üblicherweise statt.

Um den speziellen Herausforderungen dieser Arbeit zu begegnen, werden die verschiedenen Aspekte des Modells wie im Folgenden beschrieben durchlaufen:

- **Fachliche Klärung Teil I:** Zuallererst wird eine fachliche Basis geschaffen, indem ein grundsätzlicher Überblick über das Fachgebiet gewonnen wird. Dazu werden dessen Historie und zentrale Themen exploriert, aber auch die erwartete zukünftige Entwicklung betrachtet (Abschnitt 3.2).

- **Klärung gesellschaftlicher Ansprüche ans Fach:** Abschnitt 5.1 verdeutlicht grundsätzliche Anforderungen, die eine digitale Gesellschaft heute an ihre Bürgerinnen und Bürger in Zusammenhang mit dem Fachgebiet Datenmanagement stellt.

- **Auswahl informatischer Phänomene:** In Abschnitt 5.2 werden Anknüpfungspunkte an den Alltag der Schülerinnen und Schüler herausgearbeitet, bei denen diese bereits mit dem Fachgebiet Datenmanagement und dessen Phänomenen in Kontakt kommen.

- **Erfassung von Schülerperspektiven:** Die Erfassung der Schülerperspektive kann in dieser Arbeit nicht im Sinne der üblicherweise betrachteten Schülervorstellungen erfolgen, da für deren zielgerichtete Erforschung zuerst eine ausreichende fachliche Klärung erfolgt sein muss. Daher wird in Abschnitt 6.2 die Schülerperspektive stattdessen aufgegriffen, indem im Rahmen einer Fragebogenstudie das Vorwissen der Schülerinnen und Schüler zu zentralen Themen des Datenmanagements, die besonders alltagsrelevant oder Nahe am derzeitigen Informatikunterricht sind, ermittelt wird.
- **Erfassung von Lehrerperspektiven:** Die Erfassung der Lehrerperspektive beinhaltet nach *Diethelm, Dörge et al. (2011)* eine Untersuchung der Erklärungsmuster, die Lehrkräfte zu den Phänomenen haben, sowie der verfolgten Unterrichtsziele und der erwarteten Schülervorstellungen. Wie bei der Schülerperspektive kann dies jedoch erst nach einer ausreichenden fachlichen Klärung erfolgen und nur, wenn die Lehrerinnen und Lehrer schon grundlegende Erfahrungen mit den Themen im Unterricht gemacht haben, bzw. sich zumindest ausreichend eingearbeitet und auf einen Unterricht zu diesen vorbereitet haben. Da eine solche Grundlage zum Themengebiet Datenmanagement bisher nicht existiert, wird der Schwerpunkt in dieser Arbeit anders gesetzt: Es wird im Rahmen einer Fragebogenstudie (Abschnitt 6.1) untersucht, welches Wissen die Lehrerinnen und Lehrer bereits mitbringen, wie interessant sie verschiedene Themen für den Unterricht einschätzen und welche Probleme sie dabei letztlich erwarten. Auf diese Weise können schon in einem frühen Stadium wichtige Aspekte der Lehrerperspektive einbezogen werden.
- **Fachliche Klärung Teil II:** Auf Basis der gewonnenen Erkenntnisse erfolgt in Teil III, zur Schaffung eines soliden Fundaments für den Informatikunterricht, eine vertiefte fachliche Klärung des Fachgebiets unter Einbeziehung der Anforderungen bzw. Herausforderungen, die die verschiedenen zuvor betrachteten Perspektiven offenbaren. Dazu werden die Schlüsselkonzepte des Fachgebiets mithilfe eines empirischen Ansatzes identifiziert und zu einem Modell kondensiert, das u. a. für die Unterrichtsplanung, aber auch für die weitere Forschung dienlich ist.
- **Didaktische Strukturierung von Informatikunterricht:** In dieser Arbeit wird die didaktische Strukturierung (Teil IV) in zwei Teile aufgeteilt betrachtet. Um den ermittelten Anforderungen der Lehrerinnen und Lehrer Genüge zu tun, wird zuerst ein Softwarewerkzeug für den Informatikunterricht entwickelt und dessen Einsatz und Nutzen beschrieben. Im zweiten Schritt folgt die Planung einer exemplarischen Unterrichtssequenz zum Thema Data Mining, die Bezug auf zentrale Themen des Datenmanagements nimmt und im Rahmen eines Unterrichtsversuchs evaluiert wird.

Der beschriebene Ablauf wird in Abbildung 2.3 als Prozessmodell visualisiert.

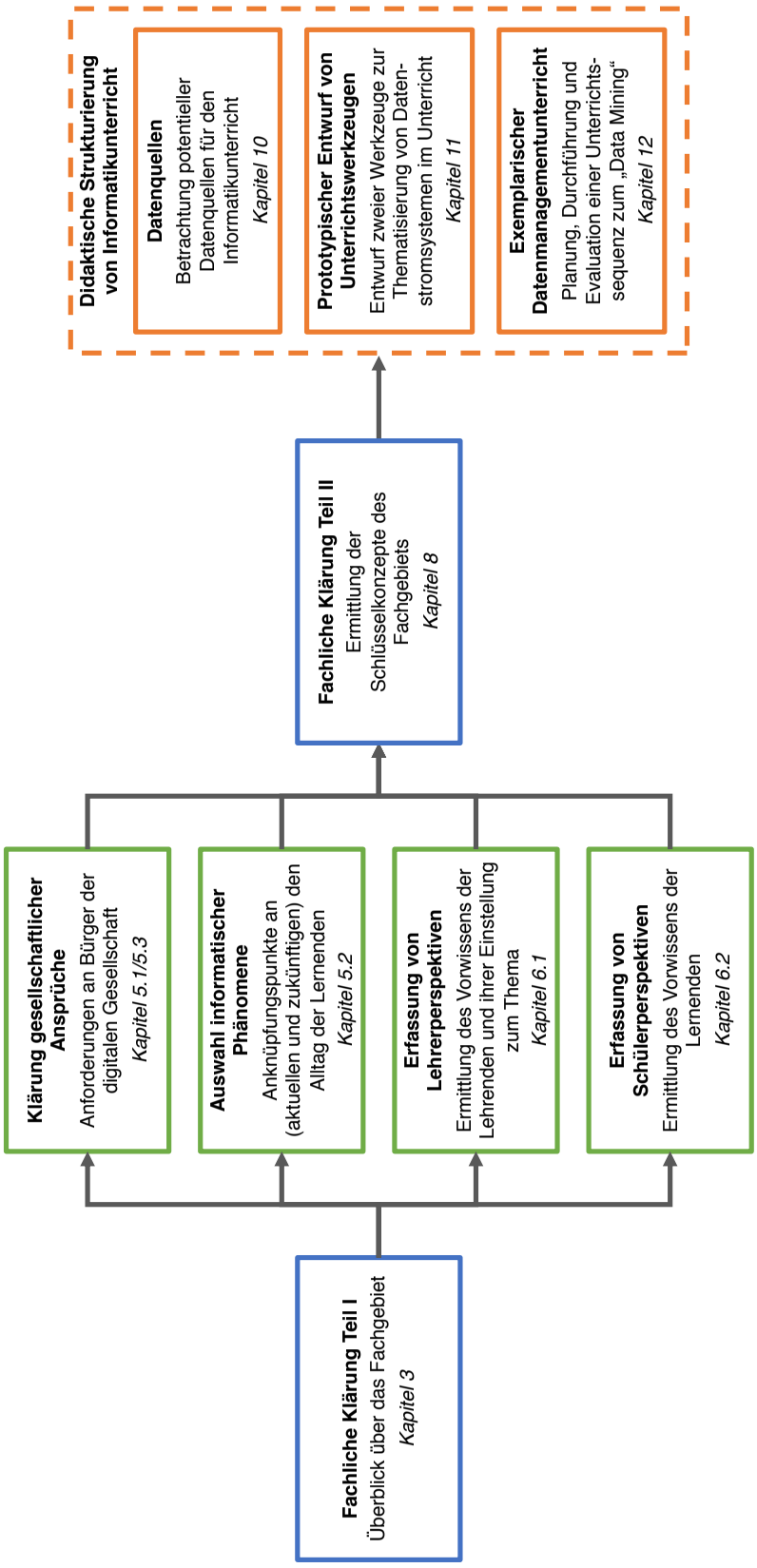


Abbildung 2.3: Ablaufmodell dieser Arbeit basierend auf den Phasen des *Modells des Didaktischen Rekonstruktion für den Informatikunterricht*.

Teil II:

Fachliche und fachdidaktische Grundlagen der Arbeit

3 Daten und Datenmanagement in der Informatik

3.1 Abgrenzung von Datenmanagement, Data Science und Data Literacy

Obwohl Daten bereits seit jeher ein in der Informatik zentrales und für diese grundlegendes Thema darstellen, ist in diesem Bereich innerhalb des letzten Jahrzehnts ein neuer und deutlicher Aufschwung erkennbar. Die Bedeutung von Daten hat – nicht nur in der Informatik – immens zugenommen und wird entsprechend in verschiedensten Kontexten viel diskutiert. Dabei treten insbesondere drei neue Begriffe, die nicht überschneidungsfrei sind, regelmäßig zutage: *Datenmanagement*, *Data Science* und *Data Literacy*. Zur Verdeutlichung des dieser Arbeit zugrundeliegenden Verständnisses werden diese Begriffe daher im Folgenden charakterisiert und definiert.

Unter *Datenmanagement* wird ein relativ neues Fachgebiet der Informatik verstanden, welches sich mit den Innovationen der letzten Jahre aus dem Fachgebiet Datenbanken herausgebildet hat. In Anlehnung an die Charakterisierung der internationalen Datenmanagementvereinigung „DAMA International“ *(vgl. DAMA International, 2010)*, wird Datenmanagement in dieser Arbeit wie folgt verstanden:

> *Datenmanagement umfasst die informatischen Grundlagen des Umgangs mit und der Verwaltung und Verarbeitung von Daten. Es bezieht alle Phasen des Datenlebenszyklus, von der Erfassung über die Strukturierung, Speicherung, Verarbeitung bis hin zur Archivierung und Löschung von Daten mit ein. Dabei steht insbesondere das Ziel im Vordergrund, den Umgang mit der Ressource Daten zu kontrollieren, diese zu schützen und sie ihrem Wert entsprechend zu nutzen.*

Mit seinen umfangreichen Schwerpunkten stellt Datenmanagement ein wichtiges und innovatives Fachgebiet der Informatik dar, das alle Facetten des Umgangs mit und der Verwaltung von Daten berücksichtigt. Klar von diesem Begriffsverständnis zu trennen ist jedoch die immer häufiger auftretende Verwendung des Begriffs Datenmanagement in anderen Kontexten: Während es beispielsweise beim Forschungsdatenmanagement eher darum geht, Forschungsdaten nachhaltig und (qualitäts-)gesichert zur Verfügung zu stellen und zu verwalten *(vgl. z. B. RatSWD, 2018)*, liegt der Fokus dieses Forschungsprojekts weniger auf der Anwendungsebene, sondern auf den informatischen Ideen des Datenmanagements.

Einen anderen Fokus als das Datenmanagement legt die *Data Science*: Diese beschäftigt sich insbesondere mit der Nutzung von Daten zur Gewinnung neuer Informationen und den dabei relevanten Methoden und Prozessen. Dabei klammert sie andere Aspekte, wie die Gewinnung der Daten oder deren Speicherung, meist aus oder betrachtet diese eher

als Randthema. Gleichzeitig kann Data Science jedoch nicht als Teilbereich des Datenmanagements aufgefasst werden, da neben verschiedenen Methoden und Konzepten des Datenmanagements auch weitere Aspekte, insbesondere auch aus Mathematik und Statistik, eine zentrale Rolle spielen. Während sich Datenmanagement speziell aus informatischer Perspektive mit Daten beschäftigt, löst sich die *Data Science* daher zum Teil von dieser, bezieht aber auch informatische Themengebiete mit ein, die im Datenmanagement weniger relevant sind. Insbesondere stellt maschinelles Lernen einen wichtigen Teilaspekt der Data Science dar. Gleichzeitig wird, in Ergänzung zu den informatischen, mathematischen bzw. statistischen Aspekten, auch der Perspektive des jeweiligen Anwendungsfeldes besonderes Gewicht beigemessen. Data Science kann somit, wie in Abbildung 3.1 dargestellt, als Schnittmenge dieser Perspektiven aufgefasst werden. Bei einer noch detaillierteren Betrachtung kann diese Datenwissenschaft aus disziplinorientierter Sicht wie folgt beschrieben werden: „$datascience = statistics + informatics + computing + communication + sociology + management|data + environment + thinking$“[8] *(Cao, 2017)*. In Anlehnung an diese Charakterisierungen und die Definition durch das US-Amerikanische National Institute of Standards and Technology *(vgl. NIST Big Data Public Working Group, 2015)* wird Data Science in dieser Arbeit wie folgt verstanden:

> *Data Science befasst sich mit der Gewinnung neuen Wissens aus Daten. Dazu werden oft umfangreiche Datenmengen mithilfe statistischer und algorithmischer Methoden unter Berücksichtigung der Expertise aus dem jeweiligen Anwendungsfeld untersucht. Data Science muss somit interdisziplinär betrachtet werden und weist nicht nur Schnittpunkte mit Informatik (insbesondere dem Datenmanagement und dem Maschinenlernen), sondern auch mit Mathematik und Statistik sowie vielfältigen Anwendungsfeldern auf.*

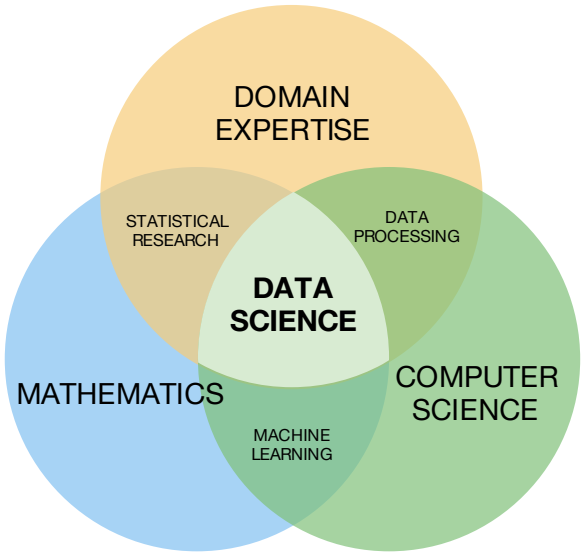


Abbildung 3.1: Charakterisierung der Data Science nach *Palmer (2015)*.

[8] Das Zeichen „|“ wird durch den Autor des Zitats als „conditional on“ verstanden.

Die *Data Literacy* löst sich noch stärker von den fachlichen Perspektiven, die dem Umgang mit Daten zugrunde liegen. Stattdessen wird das Thema *Daten* aus einer stark fächerübergreifenden Perspektive unter Berücksichtigung allgemein relevanter Aspekte betrachtet. Wie durch das Hochschulforum Digitalisierung definiert, wird Data Literacy daher wie folgt verstanden:

> *Data literacy ist die Fähigkeit, planvoll mit Daten umzugehen und sie im jeweiligen Kontext bewusst einsetzen und hinterfragen zu können. Dazu gehören die Kompetenzen, Daten zu erfassen, erkunden, managen, kuratieren, analysieren, visualisieren, interpretieren, kontextualisieren, beurteilen und anzuwenden. Data literacy ist eine zentrale Kompetenz für die Digitalisierung und die globale Wissensgesellschaft in allen Sektoren und Disziplinen. (Hochschulforum Digitalisierung, 2017)*

Diese Definition greift die zunehmende Bedeutung von Daten, insbesondere in allen Wissenschaftsbereichen, auf: Immer häufiger wird datengetriebene Forschung, beispielsweise unter dem Begriff *data-intensive scientific discovery*, als neues viertes Wissenschaftsparadigma neben Empirie, Theorie und Simulation angesehen *(vgl. Hey, Tansley und Tolle, 2009)*. Entsprechend wird Data Literacy heute als zentraler Kompetenzbereich jedes wissenschaftlichen Studiums diskutiert und von einer steigenden Anzahl an Hochschulen aufgegriffen. Doch auch außerhalb der Wissenschaften gewinnen Daten heute in vielfältigen Kontexten an Bedeutung, wie im Laufe dieser Arbeit an verschiedenen Stellen gezeigt wird. Entsprechend verdeutlicht diese Arbeit die Notwendigkeit, Data Literacy nicht nur in der Hochschulbildung zu verorten, sondern als Teil des allgemeinbildenden Schulunterrichts zu betrachten, da diese einen wesentlichen Beitrag zum mündigen und verantwortungsbewussten Leben in der heutigen digitalisierten Gesellschaft liefern kann.

Von der Data Literacy müssen weitere ähnliche Konzepte unterschieden werden: In einer von der Gesellschaft für Informatik und dem Fraunhofer IESE gemeinsam durchgeführten Studie *(Heidrich, Bauer und Krupka, 2018)* wird insbesondere der *Information Literacy* eine große Überschneidung mit der Data Literacy attestiert. Diese befasst sich jedoch aus einer anderen Perspektive mit der Thematik: Während in der Data Literacy die Gewinnung neuer Informationen zentraler Bestandteil ist, stehen bei der Information Literacy der Umgang mit und die Aufbereitung von bereits vorhandenen Informationen im Vordergrund *(vgl. bspw. ACRL Board, 2016)*. Von *Schield (2018)* werden Data Literacy und Information Literacy sogar so differenziert, dass sich erstere nur mit Daten beschäftigt, während alle mit Information in Zusammenhang stehenden Aspekte in den Bereich der Information Literacy fallen und die *Statistical Literacy* zwischen diesen beiden Bereichen vermittelt. Diese Betrachtung von Data Literacy widerspricht jedoch den weit verbreiteten Definitionen der Data Literacy durch das *Hochschulforum Digitalisierung (2017)* und von *Ridsdale et al. (2015)*. Weitere von *Heidrich, Bauer und Krupka (2018)* als eng verwandt zur Data Literacy angesehene Konzepte sind die *Data Information Literacy* und die *Science Data Literacy*, die als Spezialisierungen von Information Literacy bzw. Data Literacy im Kontext der wissenschaftlichen Forschung verstanden werden. Auch die *Digital Literacy* weist nach vielen Definitionen gewisse Überschneidungen mit der Data Literacy auf, wird aber meist eher technologieorientiert betrachtet (z. B. „*create information using digital technology*“, vgl. *Hoadley und Favaro*

(2015)). Diese Vielfalt an verschiedenen, oft eng verwandten, Literacy-Begriffen verdeutlicht die über die Jahre anhaltende Diskussion in diesem Bereich, aber auch verschiedene dabei mögliche Schwerpunkte. In Tabelle 3.1 werden die verschiedenen dargestellten Begriffe in Anlehnung an *Qin und D'ignazio (2010)* zusammengefasst bzw. zueinander kontrastiert. In dieser Arbeit wird der Schwerpunkt, wie in der aktuellen Forschung im informatiknahen Bereich derzeit üblich, auf das Konzept der Data Literacy gelegt.

	Data Literacy	**Science Data Literacy**	**Digital Literacy**	**Information Literacy**	**Data Information Literacy**
Fokus	Daten in verschiedenen Kontexten	Daten im wissenschaftlichen Umfeld	Information in verschiedenen Kontexten	Information in verschiedenen Kontexten	Information im wissenschaftlichen Umfeld
angestrebte Kompetenzen	Daten erfassen, erkunden, managen, kuratieren, analysieren, visualisieren, interpretieren, kontextualisieren, beurteilen und anwenden	sammeln, verarbeiten, verwalten, evaluieren und nutzen von Daten im wissenschaftlichen Erkenntnisprozess	finden, organisieren, evaluieren und erzeugen von Information mithilfe digitaler Technologien	entdecken von Information, verstehen wie diese erzeugt werden, erkennen ihres Wertes und des Nutzens für die Schaffung neuen Wissens	Anwendung von Information Literacy im Forschungskontext
Quelle	*Hochschulforum Digitalisierung (2017), Ridsdale et al. (2015)*	*Qin und D'ignazio (2010)*	*Hoadley und Favaro (2015)*	*ACRL Board (2016)*	*Carlson und Johnston (2015)*

Tabelle 3.1: Unterschiede verschiedener Literacy-Begriffe in Anlehnung an *Qin und D'ignazio (2010)*.

Obwohl gerade der Bereich Data Literacy sicherlich auch für die informatische Allgemeinbildung spannend ist und in dieser aufgegriffen werden sollte, wird der primäre Schwerpunkt dieser Arbeit auf das Fachgebiet Datenmanagement gelegt: Dieses berücksichtigt die informatische Perspektive auf die aktuellen Entwicklungen im Bereich der Datenverwaltung und kann somit als Zugang zu den zentralen Grundlagen des gesamten Themenkomplexes dienen; gleichzeitig stellt es eine wichtige Grundlage für die Data Literacy dar. Durch Betrachtung dieses Themengebiets kann somit eine fachliche Fundierung für die weitere fachdidaktische Forschung im Kontext der Daten geschaffen werden. Auf Basis dieser Grundlage werden durch Verschiebung des Schwerpunkts auf die Data Literacy in Kapitel 9 zentrale Kompetenzen, die jeder heute für einen fundierten Umgang mit Daten erwerben sollte, strukturiert und somit ein Perspektivwechsel vom Fachgebiet Datenmanagement hin zu den allgemeinbildenden Aspekten, die dieses beinhaltet, vorgenommen.

3.2 Datenmanagement als Fachgebiet der Informatik

Aus fachlicher Perspektive werden alle Aspekte der Verwaltung und Verarbeitung von Daten heute insbesondere im Fachgebiet *Datenmanagement* thematisiert. Trotz der eigentlich langen Historie, kann *Datenmanagement* als relativ junges Fachgebiet der Informatik angesehen werden, das sich erst mit den Innovationen der letzten Jahre herausgebildet hat. Der bekannteste und für Datenmanagementsysteme prototypische Vertreter sind Datenbanksysteme. Diese seit langem bewährten und optimierten Systeme stellen den Ursprung der Entwicklung des Fachgebiets dar (vgl. Abschnitt 3.2.1). Neben Datenbanken gibt es jedoch

eine Vielzahl weiterer für unterschiedliche Zwecke optimierter Datenmanagementsysteme, die sich in den letzten Jahren und Jahrzehnten oft drastisch weiterentwickelt haben. Bekannte Beispiele stellen u. a. *dateibasierte Datenspeicher*, *dokumentenbasierte Datenspeicher*, *Data Warehouses* und *Datenstromsysteme* dar. All diese Systeme existieren heute wiederum in verschiedenen Ausprägungen, die für unterschiedliche Anwendungszwecke optimiert sind. Neben diesen Systemen wird das Fachgebiet heute außerdem noch von verschiedenen weiteren Themen, wie beispielsweise *Big Data*, *Datenanalyse* und speziell *Data Mining*, *Datenqualität* und *Metadaten* geprägt. Gleichzeitig befasst sich Datenmanagement natürlich auch mit Querschnittsthemen der Informatik wie beispielsweise der *Datensicherheit*[9], dem ethisch korrekten Umgang mit Daten[10] sowie dem Datenschutz[11]. Auch in anderen Bereichen der Informatik, wie beispielsweise der Softwareentwicklung, ist Datenmanagement heute integraler Bestandteil, der jederzeit miteinbezogen bzw. berücksichtigt werden muss.

Zur genaueren Charakterisierung des Fachgebiets wird im Folgenden kurz dessen Genese nachgezeichnet, sowie an einer Auswahl aktueller Themen die Breite sowie aktuelle Herausforderungen dieses Fachgebiets dargestellt. Die professionelle Perspektive wird daraufhin eingenommen, indem zentrale Aspekte einer professionellen Charakterisierung von Datenmanagement, des „Data Management Body of Knowledge“, beschrieben werden. Zur Verdeutlichung der zukünftigen Relevanz wird außerdem die prognostizierte Entwicklung des Fachgebiets und seiner Bedeutung skizziert.

3.2.1 Historische Entwicklung

In den letzten zehn bis fünfzehn Jahren war im Datenmanagement eine stetige Weiterentwicklung zu verzeichnen. Mit dieser ging auch ein deutlicher Anstieg an wissenschaftlichen Beiträgen aus und in diesem Fachgebiet einher. Nicht zuletzt hat sich das gesamte Datenmanagement erst in dieser Zeit aus dem bisherigen Forschungsfeld Datenbanken entwickelt. Zentral für diese Entwicklung ist der exponentielle Anstieg der Menge an Daten, die die Menschheit heute vorhält und verwaltet und die heute nach übereinstimmenden Schätzungen bereits die Zettabyte-Marke überschritten hat (vgl. Abbildung 3.2 sowie Abbildung 3.3). Die mit diesem Wachstum einhergehenden steigenden Anforderungen an Datenmanagement stellen eine der zentralen Herausforderungen des Fachgebiets dar und haben wesentlich zu dessen Weiterentwicklung beigetragen.

[9] Sowohl im Sinne der technischen Sicherheit bzw. Ausfallsicherheit (engl. „safety“), als auch dem Schutz vor Angriffen (engl. „security“)

[10] Hierzu zählt beispielsweise die Wahrung der Persönlichkeitsrechte der Personen, deren Daten in Data-Mining-Analysen analysiert werden.

[11] Der Begriff „Datenschutz“ ist aus informatischer Sicht klar vom ethisch korrekten Umgang mit Daten zu unterscheiden, da es beim Datenschutz insbesondere um den (technischen bzw. organisatorischen) Schutz personenbezogener bzw. -beziehbarer Daten vor Diebstahl, Manipulation und Missbrauch sowie die damit einhergehenden Methoden und Maßnahmen geht. Beispielsweise versteht die *Europäische Union (2016)* in der Datenschutz-Grundverordnung den Begriff *Datenschutz* als den *„Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten“*

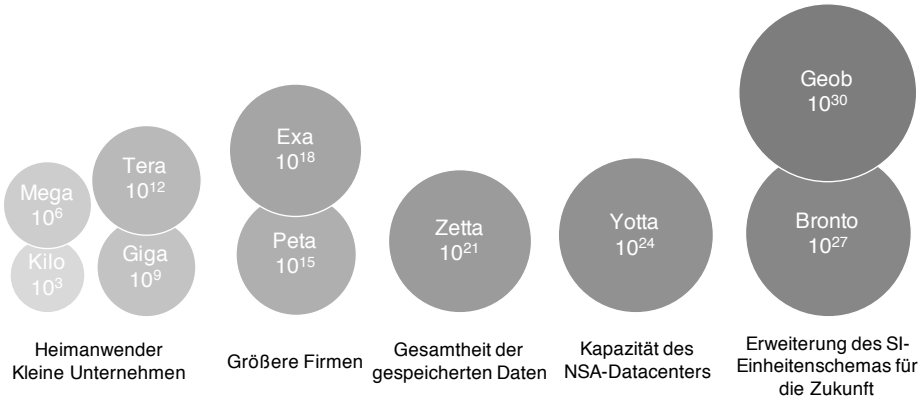


Abbildung 3.2: Überblick über die Größenordnungen der heute in verschiedenen Bereichen gespeicherten Datenmengen.

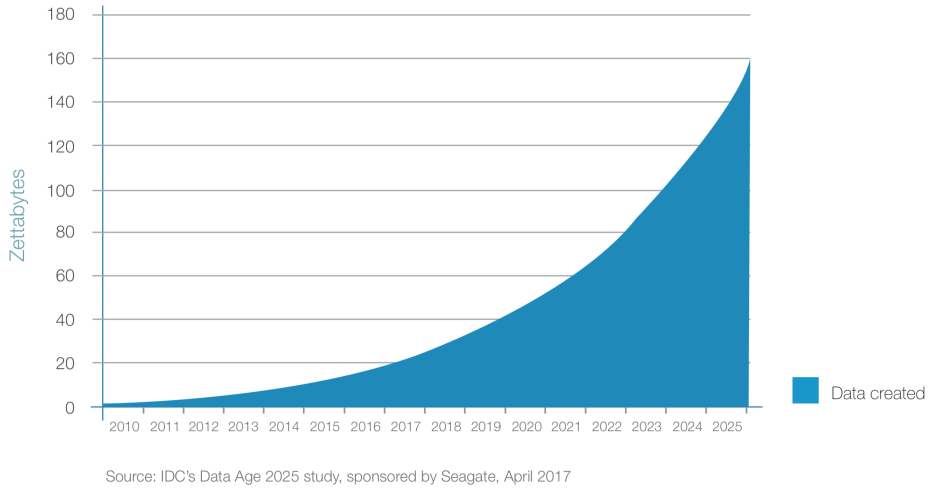


Abbildung 3.3: Wachstum der Datenmenge der Menschheit *(Quelle: Reinsel, Gantz und Rydning, 2017)*.

Jedoch zeigt sich auch im historischen Rückblick, dass die Gewinnung, Speicherung und Analyse großer Datenmengen mit dem Ziel der Informationsgewinnung, heute zumeist als *Big Data* bezeichnet, nichts komplett Neues ist. Die erste dokumentierte Arbeit, die bereits dem heutigen *Data Mining* ähnelt, wurde schon 1855 durchgeführt: Der US-Marineoffizier Matthew Fontaine Maury erfasste und analysierte händisch 1,2 Millionen Datenpunkte aus verschiedenen Datenquellen. Dabei handelte es sich einerseits um solche Daten, die explizit und strukturiert für diesen Zweck erfasst wurden (z. B. von ihm erstellte und von Schiffsbesatzungen für ihn ausgefüllte Berichtsformulare), aber auch um völlig andere und weniger strukturierte Daten (beispielsweise aus Seelogbüchern und Karten stammend), die schon lange vorher und zu völlig anderem Zweck erfasst worden sind. Durch Kombination dieser Daten versuchte Maury, Seerouten durch Ausnutzung von Strömungen und Winden zu optimieren. Mit seinen Ergebnissen *(vgl. Maury, 1855)* konnte die Reisezeit der Schiffe auf den betrachteten Routen durchschnittlich um ein Drittel reduziert werden *(Mayer-Schönberger und Cukier, 2013)*. Obwohl die Datenmenge für heutige Verhältnisse si-

cherlich klein ist, kann sie für die händische Datenverarbeitung als durchaus beträchtlich angesehen werden. Zugleich zeigen sich zentrale Eigenschaften moderner Datenanalysen: Es werden nicht nur verschiedenste Datenquellen miteinander verknüpft, sondern dabei insbesondere auch solche Daten miteinbezogen, die sowieso schon vorhanden sind und für völlig andere Zwecke erfasst wurden. Nicht nur die Datenanalyse von Maury, sondern auch ihre Einflüsse können sich bei zeitgeschichtlicher Betrachtung durchaus mit denen heutiger Datenanalysen messen.

Ein weiteres ähnlich gelagertes Beispiel, das die Bedeutung korrelationsbasierter Datenanalysen verdeutlicht, stellt die Entdeckung der Ursachen der Cholera-Epidemie von Hamburg in den Jahren 1892/93 durch Robert Koch dar. Aus den ihm zur Verfügung stehenden Daten über Cholerafälle in den damals noch politisch getrennten Städten Altona, Hamburg und Wandsbek konnte Koch eine klare Korrelation zwischen den Merkmalen „an Cholera erkrankt“ und „in Hamburg lebend“ ablesen: Abbildung 3.4 zeigt, dass die politische Grenze zwischen Hamburg und Altona zugleich (mit wenigen Ausnahmen) die Grenze der Choleraepidemie darstellte. Insbesondere *„vor einer Straße, welche auf einer längeren Strecke die Grenze bildet, wurde die Hamburger Seite von Cholera befallen, die Altona blieb frei“ (Koch, 1893)*. Obwohl natürlich diese Korrelation nicht in eine Kausalität überführt werden kann – dies würde bedeuten, dass sich die Krankheit an einer politischen Grenze orientiert – konnte durch weitere Untersuchungen eine zugrundeliegende Kausalität aufgedeckt werden: ein an den politischen Grenzen getrenntes, unterschiedlich weit entwickeltes Wasserversorgungssystem. Auch dieses Beispiel demonstriert, dass die Analyse von Daten, auch basierend auf Korrelationen, kein neuartiger Ansatz ist, aber klare Erfolge verspricht.

Abbildung 3.4: Cholerafälle an der Grenze von Hamburg (südlich) und Altona. Aus: *Exner (2009)*.

Obwohl die Verbreitung solcher Datenanalysen erst in den letzten Jahren enorm zugenommen hat, ist es bereits seit der Erfindung von relationalen Datenbanken *(vgl. Codd, 1970)* in den 1970er/80er Jahren möglich, größere Datenmengen strukturiert zu speichern und zu aggregieren. Innerhalb der letzten zehn bis fünfzehn Jahre kam es zu einer deutlichen Aufweitung dieses Themengebiets – das Fachgebiet Datenmanagement entstand. Der Kristallisationskeim dieser Entwicklung ist meist unter dem Stichwort *Big Data* bekannt, das heute häufig durch die großen Mengen verschiedenster Arten von Daten, die in hoher Geschwindigkeit gespeichert und verarbeitet werden müssen, charakterisiert wird *(Kemper*

und Eickler, 2015). Auch der Wandel von kausalitätsbasierten zu immer häufiger korrelationsbasierten Datenanalysen (oft unter dem Stichwort *Data Mining*) ist sicherlich eine der Hauptursachen dafür, dass heute oft von einem Paradigmenwechsel im Bereich der Datenverwaltung und -analyse gesprochen wird (vgl. z. B. *Fischer (2014)*). Ein bekanntes Beispiel für korrelationsbasierte Datenanalysen sind Produktempfehlungen, beispielsweise in Onlineshops *(Sommer, 2013)*: Hier werden immer häufiger die Einkäufe der Kunden auf scheinbare Zusammenhänge hin analysiert und oft weitere Daten, beispielsweise solche aus sozialen Medien, miteinbezogen, um auf diese Weise persönlichere Empfehlungen zu geben.

Heute durchdringt Datenmanagement unser gesamtes Leben: Neben seiner Bedeutung in der Informatik ist dieses Fachgebiet oft Auslöser oder Thema von gesellschaftlichen Diskursen, beispielsweise in Zusammenhang mit der Speicherung und Analyse großer Datenmengen durch Geheimdienste oder im Rahmen von Vorratsdatenspeicherung, bei Datenschutzthemen oder der Erfassung und Auswertung von Kundendaten durch Webportale und immer öfter auch im traditionellen Handel, beispielsweise durch Nutzung von Bonuskartensystemen. Gleichzeitig nutzt heute auch Jeder verschiedene Produkte, die ohne die Innovationen im Datenmanagement, wenn überhaupt, nur eingeschränkt möglich wären, wie zum Beispiel Cloud-Datenspeicher, Möglichkeiten zur Datensynchronisierung, moderne Suchmaschinen oder soziale Medien.

3.2.2 Zentrale Themen des Datenmanagements

Um eine Grundlage für die weitere Arbeit zu schaffen, werden im Folgenden zentrale Themen des Datenmanagements expliziert und auf das Wesentliche reduziert dargestellt. Diese Themen wurden anhand der Forschungsschwerpunkte des Fachgebiets so ausgewählt, dass sie dieses auf einem aktuellen Stand repräsentieren und dessen Breite aufzeigen. Während die fünf Themen *Big Data*, *verteilte Datenspeicher*, *Data Mining*, *Datenstromsysteme* und *Metadaten* detaillierter beschrieben werden, wird zur weitergehenden FUndierung daraufhin noch ein Überblick über den *Data Management Body of Knowledge* gegeben, der das Fachgebiet aus beruflich-angewandter Sicht charakterisiert.

Verwaltung und Nutzung großer Datenmengen: Big Data

Big Data ist sicherlich der bekannteste Begriff in Zusammenhang mit den aktuellen Entwicklungen im Datenmanagement. Trotz der Relevanz dieses Begriffs existiert jedoch keine klare und weithin anerkannte Definition. Unter Bezug auf *McBurney (2013)* schreibt beispielsweise die Gesellschaft für Informatik in ihrem Informatiklexikon: *„Der Ursprung und die erstmalige Verwendung des Begriffes Big Data im aktuellen Kontext sind nicht ganz eindeutig und es werden unterschiedliche Quellen genannt, die den Begriff in der aktuellen Verwendung geprägt haben könnten.“ (Klein, Tran-Gia und Hartmann, 2013)* Es besteht lediglich Einigkeit darüber, dass Big Data insbesondere durch die sog. *drei V* (vgl. Abbildung 3.5; *Kemper und*

Eickler (2015) und Laney (2001)) charakterisiert wird: große Datenmengen (*volume*), hohe Geschwindigkeit (*velocity*) und unterschiedlichste Arten von Daten (*variety*). Diese Eigenschaften sind oft zentral für Datenanalysen in der heutigen digitalen Gesellschaft. Neben diesen *drei V* werden häufig noch weitere genannt, insbesondere die Vertrauenswürdigkeit (*veracity*) und der Wert (*value*) der Daten (vgl. z. B. *Ali-ud-din Khan, Fahim Uddin und Gupta (2014)*).

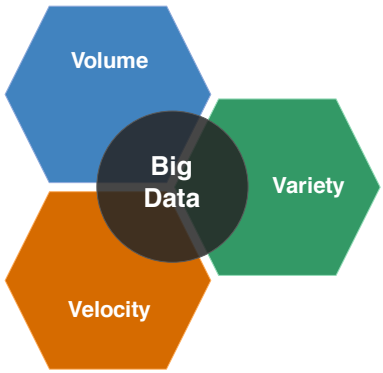


Abbildung 3.5: Die drei zentralen Eigenschaften von Big Data nach *Laney (2001)*.

Im Zusammenhang mit Big Data entstanden verschiedene neuartige Technologien, wie parallelisierte Analysealgorithmen (beispielsweise Googles Map-Reduce-Algorithmus, vgl. *Dean und Ghemawat (2008)*) oder Datenbanksysteme, die klassische Datenbanken auf den Prüfstand stellen und versuchen, den neuen Anforderungen gerecht zu werden: Diese nicht-relationalen, oft *NoSQL-Datenbanken* genannten, Ansätze ermöglichen typischerweise eine verteilte und schnelle Datenspeicherung. Die Bezeichnung *NoSQL* hat sich zwar durchgesetzt, ist aber aufgrund der Verwechslungsgefahr mit der relationalen Datenbank „noSQL“ *(Strozzi, o.D.)*, die explizit kein SQL unterstützt, und aufgrund der fehlleitenden intuitiven Interpretation ungünstig gewählt: Es geht bei diesen Datenbankenkonzepten nicht darum, dass diese kein SQL unterstützen, sondern um eine Abkehr vom relationalen Modell der bisher üblichen Datenbanken. Es handelt sich dabei also um nicht-relationale Ansätze, die teilweise sogar eine Unterstützung von SQL bzw. davon abgeleiteten Sprachen mitbringen. In diesem Sinn wird die Abkürzung daher heute üblicherweise als „***N**ot **o**nly **SQL***“ interpretiert *(Edlich et al., 2011)*. Diese und weitere im Zusammenhang mit Big Data entstandene Datenmanagementsysteme erlauben, insbesondere durch ihre oft stark verteilten und weniger strukturierten Ansätze, eine höhere Komplexität der Datenanalysen. Gerade die sogenannten dokumentenorientierten Datenbanken lassen dem Nutzer sehr starke Freiräume, da sie kein definiertes Datenschema benötigen: Beispielsweise erlauben dokumentenorientierte NoSQL-Datenbanken die Speicherung unterschiedlich strukturierter Dokumente in derselben *Kollektion*[12]. Die Einhaltung vereinbarter Strukturen oder Datenmodelle obliegt daher dem Anwender beziehungsweise der Anwendung anstatt dem Datenbanksystem selbst. Gleichzeitig wird zur Formulierung von Anfragen an die meisten

[12] In dokumentenorientierten Datenbanken wird der Begriff *Kollektion* meist auf konzeptionell gleicher Ebene eingesetzt, wie die Tabelle bei relationalen Datenbanken. Je nach konkretem System werden jedoch auch andere Begriffe genutzt.

nichtrelationalen Datenbanken nicht die Anfragesprache SQL verwendet, sondern eigene vom jeweiligen System abhängige Sprachen, die sich häufig an bekannter Syntax, beispielsweise der objektorientierten Syntax von Java, orientieren. Je nach Intention können solche Datenbanken im Informatikunterricht daher einen Blick über den Tellerrand ermöglichen, neue Sichtweisen vermitteln oder als alternatives Werkzeug für die Vermittlung klassischer Konzepte dienen.

Handhabung immer größerer Datenmengen: Verteilte Datenspeicher und Cloud-Speicherung

Nicht nur um Daten besonders einfach mit anderen teilen zu können, sondern auch um einen gewissen Schutz vor Datenverlust zu erlangen und flexibel auf Daten zugreifen zu können, setzen sich im Alltag heute immer stärker Cloud-Datenspeicher durch, die die Speicherung in eine nebulöse Cloud-Infrastruktur auslagern, die durch verschiedene Anbieter unterhalten oder ggf. in Firmen selbst eingerichtet wird. Ein Großteil der Smartphone-Nutzer nutzt beispielsweise den Clouddienst des jeweiligen Betriebssystemherstellers um Daten wie Kontakte, E-Mails, Passwörter usw. dort zu speichern. Während im Alltag der Begriff *Cloud* oft als Synonym für diese Form der Datenspeicherung verwendet wird, ist er aus informatischer Sichtweise jedoch umfangreicher und berücksichtigt unter anderen die Auslagerung nicht nur von Daten, sondern beispielsweise auch kompletter Infrastrukturen. Der aus Perspektive des Datenmanagements relevante Aspekt des Cloud Computing, die Datenspeicherung in der Cloud, stellt entsprechend nur eine spezielle Form der verteilten Datenspeicherung dar. Diese gewinnt auch mit anderen Zielen als im privaten Umfeld an Bedeutung, insbesondere aufgrund der zunehmenden Datenmengen, die heute gespeichert werden: Mehrere Exabyte an Daten können nicht mehr auf einem einzelnen Datenspeicher gespeichert werden, alleine aus Gründen der eingeschränkten Speicherdichte heutiger Festplatten und der physischen Größe, die diese daher annehmen müssten. Gleichzeitig ist zu erwarten, dass die gesamte Datenmenge der Menschheit auch weiterhin stärker ansteigt als die Kapazität der zur Verfügung stehenden Speichermedien, sodass die verteilte Speicherung immer wichtiger wird. Dabei tritt jedoch ein Konflikt mit den klassischen Anforderungen an Datenspeicherung auf: Daten sollen üblicherweise konsistent und schnell zugreifbar vorgehalten werden. Um die Konsistenz der Daten bei verteilter Datenspeicherung wahren zu können, muss jedoch sichergestellt werden, dass beispielsweise Änderungen an den gespeicherten Daten auf allen beteiligten Servern vollzogen wurden, bevor eine neue Transaktion zugelassen wird – es ist offensichtlich, dass diese Prüfung die Geschwindigkeit dieser Aktion reduziert (vgl. Abbildung 3.6) und somit durch verteilte Datenspeicherung die Sicherstellung von Konsistenz die Geschwindigkeit einschränkt. Dies wird durch das *CAP-Theorem (Edlich et al., 2011; Brewer, 2012)* beschrieben, laut dem die drei Eigenschaften Konsistenz (***c****onsistency*), Verfügbarkeit (***a****vailability*) und Partitionstoleranz (***p****artition tolerance*) unvereinbar sind (vgl. Abbildung 3.7). Nur zwei dieser Eigenschaften können gleichzeitig sichergestellt werden. Das CAP-Theorem verdeutlicht daher eine der

[13] Das ACID-Paradigma beschreibt die vier zentralen Eigenschaften relationaler Datenbanken: **a**tomicity, **c**onsistency, **i**solation, **d**urability *(vgl. Kemper und Eickler, 2015)*.

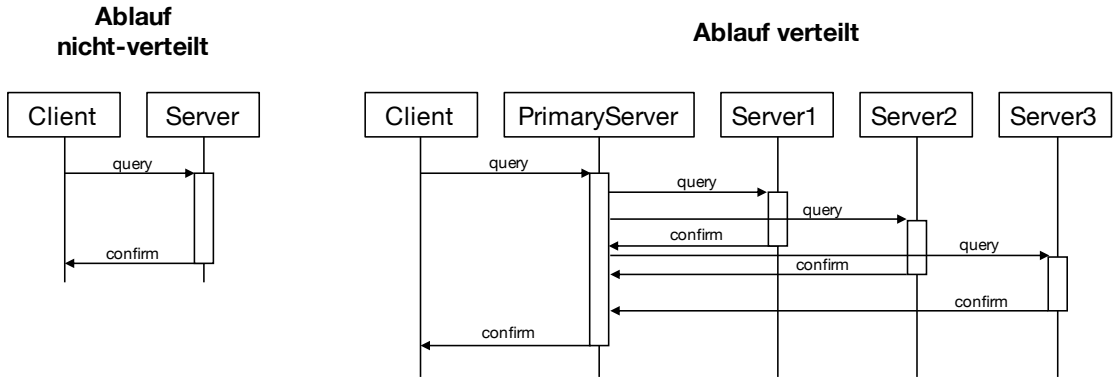


Abbildung 3.6: Vergleich der Komplexität von normalen und verteilten Transaktionen.

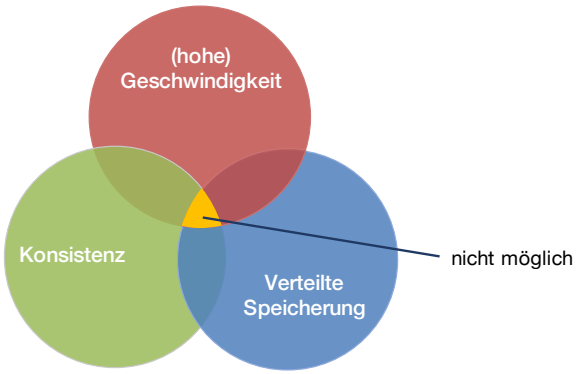


Abbildung 3.7: Veranschaulichung des CAP-Theorems *(vgl. Brewer, 2012)*.

zentralen Herausforderungen, denen Anwender und Entwickler bei der Auswahl bzw. dem Entwurf von modernen Datenmanagementsystemen heute gegenüberstehen.

Während relationale Datenbanken typischerweise für Konsistenz und Verfügbarkeit optimiert sind und dem ACID[13]-Paradigma gehorchen, aber nur eingeschränkt für verteilte Datenspeicherung geeignet sind, vernachlässigen nicht-relationale Datenbanken typischerweise die Konsistenz des Datenbestandes. Sie genügen dem BASE[14]-Paradigma, daher sind sie grundsätzlich immer verfügbar, in einem dynamisch veränderlichen Zustand, der (z. B. zur Erhöhung der Konsistenz) durch das Datenbanksystem jederzeit (ohne externe Auslöser in Form einer Transaktion) verändert werden kann, und sie werden, früher oder später, konsistent vorliegen. Es wird daher keine ständige Konsistenz garantiert, obwohl die Herstellung eines konsistenten Zustands weiterhin ein Ziel ist, weswegen das Datenbankmanagementsystem regelmäßig versucht, diesen zu erreichen. Solche Datenbankvarianten werden daher typischerweise in Situationen eingesetzt, in denen die dauerhafte Konsistenz der Daten weniger kritisch ist, beispielsweise bei verteilten Webanwendungen wie sozialen Netzwerken oder Suchmaschinen. Bekannte Vertreter dieser Datenbanken stellen die verteilte und auf Hochleistung optimierte Datenbank *Google BigTable (Chang et al., 2008)*, die

[14]Nicht-relationale Datenbanken gehorchen oft dem BASE-Konsistenzmodell: **b**asically **a**vailable, **s**oft-state, **e**ventually consistent *(vgl. Edlich et al., 2011)*.

unter anderem für Google Books genutzt wird, die graphbasierte Datenbank *Neo4J*[15] oder die heute weit verbreitet eingesetzte dokumentenorientierte Datenbank *MongoDB*[16] dar.

Mit der zunehmenden Verbreitung verteilter Datenspeicherung steigt auch die Bedeutung von verteilten Datenanalysen: Um eine hohe Analysegeschwindigkeit zu erzielen, ist es heute in vielen Fällen essenziell, die Analyseprozesse auf verschiedenen Rechenknoten parallel durchzuführen. Dabei hat beispielsweise der *Map-Reduce*-Algorithmus (vgl. Abbildung 3.8; *Dean und Ghemawat (2008)*) zentrale Bedeutung: Dieser teilt die Verarbeitung der Daten in vier Verarbeitungsschritte[17] auf, von denen die zwei den Namen des Algorithmus bestimmenden durch den Nutzer spezifiziert und jeweils parallel auf verschiedenen Rechenknoten ausgeführt werden. Dies ermöglicht eine hochgradig parallele Verarbeitung von Daten, die auf denselben oder anderen Rechenknoten vorliegen, und somit eine sehr hohe Flexibilität der Datenanalyse. Das Map-Reduce-Verfahren wird daher heute in einer Vielzahl von Anwendungsfällen erfolgreich eingesetzt und stellt ein wichtiges Beispiel für die Datenverarbeitung im Big-Data-Zeitalter dar.

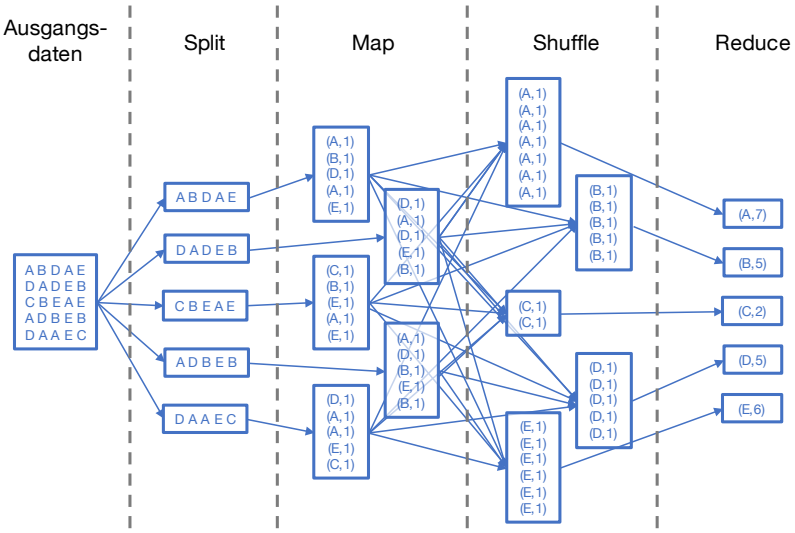


Abbildung 3.8: Veranschaulichung des Map-Reduce-Algorithmus am Beispiel einer Wortzählung.

Strukturierung und Beschreibung von Daten: Metadaten

Ein weiteres wichtiges Thema des Datenmanagements, das zugleich auf einer eher technischen aber auch auf einer eher nutzerorientierten Ebene angesiedelt sein kann, stellen Metadaten dar: Einerseits werden diese oft zur internen Datenorganisation und -strukturierung innerhalb von Datenmanagementsystemen eingesetzt, beispielsweise in Form von Prüfsummen oder (Such-)Indizes, andererseits treten sie auch auf Anwendungsebene und damit im

[15] `https://neo4j.com`
[16] `https://www.mongodb.com`
[17] Je nach Darstellung wird teils auf den ersten Schritt, *split*, verzichtet, sodass der Algorithmus auch häufig mit drei Schritten dargestellt ist.

Wahrnehmungsbereich des Nutzers zutage. Heute spielen sie selbst im gesellschaftlichen Diskurs eine zunehmend wichtige Rolle: Die Möglichkeiten und Gefahren der Erfassung von Metadaten und insbesondere der Einfluss auf unsere Privatsphäre wird immer häufiger diskutiert.

Aus fachlicher Sicht können Metadaten drei verschiedene Funktionen erfüllen *(vgl. Riley, 2017)*:

- *Deskriptive Metadaten* beschreiben oder identifizieren Informationen und Informationsquellen und machen diese damit (einfacher) zugänglich bzw. auffindbar. Es handelt sich dabei beispielsweise um den Namen einer Datei, den Ort einer Fotoaufnahme, die Autoreninformationen von Dokumenten, bei einer HTTP-Anfrage mitgelieferte Informationen über den Client oder den Primärschlüssel eines Datensatzes in einer Datenbank.
- *Administrative Metadaten* werden beispielsweise zur Verwaltung von Informationen und zur Rechtekontrolle eingesetzt. In einer Datenbank können dies beispielsweise die Rechte der Nutzer auf einzelne Tabellen sein, bei der Verwendung von Cloudspeichern die Zugriffsrechte verschiedener Nutzer auf dieselben Dateien oder in Informationsverwaltungssystemen ein Löschdatum für Informationen.
- *Strukturelle Metadaten* stellen insbesondere Beziehungen zwischen Daten und Datenquellen dar, die zur Navigation und Präsentation eingesetzt werden können. Es handelt sich beispielsweise um Kategorisierungen von E-Mails, die Einsortierung von Dateien in eine Ordnerstruktur, aber auch um Fremdschlüssel in Datenbanken.

Diese drei Funktionen schließen sich jedoch nicht gegenseitig aus: So erfüllt beispielsweise ein einem Foto als Schlagwort mitgegebenes Metadatum einerseits eine deskriptive, oft aber auch eine strukturelle Funktion. Allgemein kann also gesagt werden, dass Metadaten die von ihnen beschriebenen Daten anreichern, ohne die Originaldaten zu verändern. Auf diese Weise machen sie viele Funktionen von Informatiksysteme möglich: Ohne Metadaten wäre die Synchronisation von Daten genauso wenig denkbar, wie beispielsweise der Zugriff auf Fotos durch Suche nach deren Entstehungsort oder die Organisation von Dateien in Ordnerstrukturen. Gleichzeitig schafft die eher nebenbei und oft im Hintergrund stattfindende Erzeugung von Metadaten auch verschiedene Gefahrenpotentiale: So wurden beispielsweise Fälle bekannt, in denen durch Nutzung von Metadaten Kommunikations- oder Bewegungsprofile erstellt oder durch ein fehlendes Bewusstsein über Metadaten Ausschnitte geheimer Dokumente veröffentlicht worden sind. Für einen verantwortungsbewussten und selbstbestimmten Umgang mit Daten ist daher heute ein Bewusstsein für die Allgegenwärtigkeit von Metadaten unabdingbar.

Ein weiteres aktuelles Beispiel für die Aussagekraft von Metadaten, die nahezu unbemerkt von den Nutzern erfasst wurden, stellt die *Strava Heat Map (Strava, 2017)* dar: Auf dieser Karte stellt die Firma Strava, Hersteller einer Fitness-Tracking-App, die Aktivitäten seiner Nutzerinnen und Nutzer dar. Die Erfassung der Positionsdaten war dabei sicherlich nicht der primäre Zweck der eingesetzten Fitnesstracker, sondern wurde den eigentlich erfassten

Aktivitätsdaten als vom Nutzer kaum wahrgenommenes Metadatum mitgeliefert. Während durch diese Daten auf den ersten Blick eher harmlos erscheinenende Informationen darüber gewonnen werden können, wo Personen besonders aktiv sind, erlauben diese auch wesentlich sensiblere Einblicke: Beispielsweise können durch diese Aktivitätsdaten klare Einblicke in die Strukturen von Militärbasen gewonnen werden *(The Verge, 2018)*, die sogar detaillierter sind als die Satellitenbilder verschiedener Kartendienste. Dieses Beispiel demonstriert damit die Sekundärnutzung von Daten, die in diesem Fall nicht durch Nutzung klassischer Datenanalysemethoden, sondern durch eine rein visuelle Auswertung geschieht.

Korrelationsbasierte Datenanalyse: Data Mining

In engem Zusammenhang mit den beiden vorherigen Themen steht das „Data Mining", eine Vorgehensweise bei Datenanalysen, die durch Big Data besondere Bedeutung erlangte: Begrifflich angelehnt an den Goldbergbau, geht es dabei um die Suche nach neuen, wertvollen und nicht-trivialen Informationen in großen Datenmengen *(nach Kantardzic, 2011)*. Der Analogie entsprechend sind moderne Datenanalysen oft weniger zielgerichtet als klassische: Statt direkt eine Goldader abzubauen, werden willkürlich Tunnel in den Datenberg getrieben und nach dem Informationsgold gesucht. Bezogen auf Datenanalysen bedeutet das: Um neue Zusammenhänge zwischen Daten zu entdecken oder Trends in diesen zu erkennen, werden die Daten durch (meist statistische Verfahren) schon fast willkürlich miteinander in Beziehung gesetzt. Dabei geht es zumeist nicht um die Ermittlung von Kausalzusammenhängen, sondern um eine rein korrelationsbasierte Betrachtung (selbst wenn die entdeckten Korrelationen logisch nicht erklärbar sind). Um der Gefahr zu entgehen, rein zufällige Zusammenhänge als gegeben hinzunehmen, müssen für solche Analysen möglichst große Datenmengen vorliegen – idealerweise ein vollständiger Datenbestand (vgl. z. B. *Mayer-Schönberger und Cukier (2013)*).

Ein typisches Beispiel für solche Data-Mining-Analysen stellt die Analyse von Suchanfragen durch Suchmaschinenbetreiber dar. Beispielsweise entwickelte Google unter dem Namen *Flu Trends*[18] ein Modell zur Vorhersage von Grippewellen. Die Herangehensweise an das Problem ist typisch für Data Mining: Statt sich auf rein logische Begriffe wie „Grippe", verschiedene Symptome oder Medikamentennamen zu fokussieren, wurden willkürlich wirkende Suchwörter mit den Statistiken der Gesundheitsbehörden abgeglichen und dabei nach Korrelationen gesucht *(Ginsberg et al., 2009)*. Ergebnis war ein Katalog von Suchbegriffen und darauf basierend ein Modell, das trotz zeitweiser Fehler in der absoluten Zahl an erwarteten Grippefällen die Verläufe der Grippewellen gut vorhersagen konnte *(Valdivia et al., 2010)*. Während knapp 80% der 100 Begriffe mit der höchsten Korrelation erklärbar waren (es handelte sich beispielsweise um Symptome der Grippe oder Medikamentennamen), wiesen 19 zwar eine hohe Korrelation auf, konnten aber nicht logisch erklärt werden. Auch wenn Flu Trends mittlerweile eingestellt wurde, bietet Google mit

[18] `https://www.google.org/flutrends`

Trends[19] auch weiterhin ein Werkzeug zur Suchdatenanalyse an, mit dem jeder die Häufigkeit verschiedener Suchbegriffe auswerten und so eigene korrelationsbasierte Analysen durchführen kann – die wie bei Google Flu Trends zur Entdeckung von Kausalitäten oder auch zu rein korrelationsbasierten Aussagen führen kann. Beispielsweise lassen sich Vermutungen über die Verbreitung von bestimmten Geräten bzw. Diensten oder die Beliebtheit verschiedener Personen in verschiedenen Regionen aufstellen, die zeitliche Entwicklung der Popularität von Suchbegriffen nachvollziehen und damit im Zusammenhang stehende Themen finden. Eine Demonstration der Möglichkeiten von Google Trends wurde beispielsweise zur Bundestagswahl 2017 veröffentlicht[20], basierend auf diesen Daten wurden auch verschiedenste Analysen durchgeführt: Abbildung 3.9 zeigt eine Auswertung dieser Daten, bei denen spannenderweise die relativen Suchhäufigkeiten nach Kandidaten der verschiedenen Parteien am Vortrag der Wahl die Ergebnisse der Bundestagswahl relativ gut widerspiegeln, ohne dass ein Blick auf die dahinterstehenden Kausalitäten geworfen wurde.

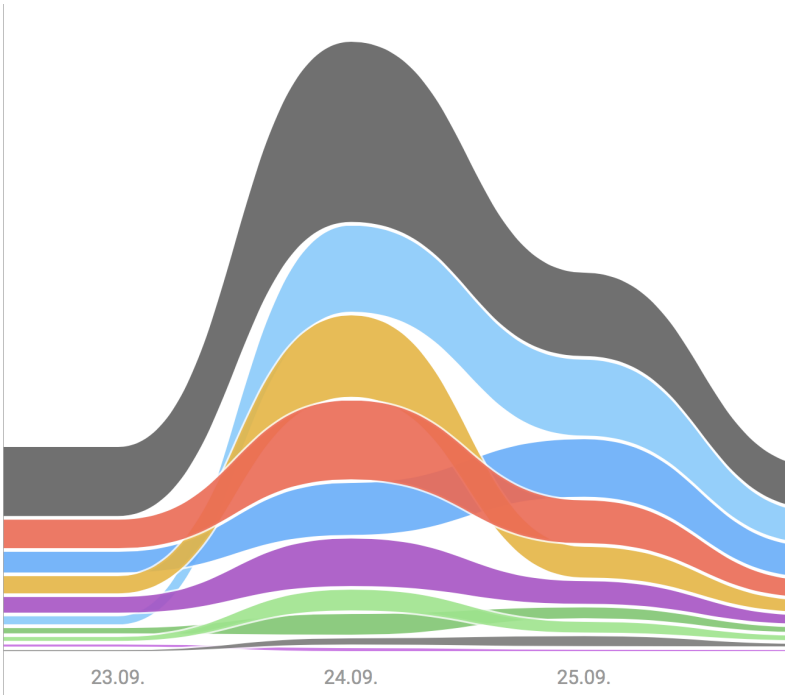


Abbildung 3.9: Visualisierung von Daten aus Google Trends zur Bundestagswahl 2017. Die Farben entsprechen den üblichen Parteifarben, Farbabstufungen visualisieren mehrere Kandidaten derselben Partei. *[Quelle: Google News Lab/2Q17.de Grafik von* `http://www.2q17.de/last-7.html`*].*

[19] `https://trends.google.de`
[20] `https://trends.google.de/trends/story/DE_cu_mmQVZFkBAADWgM_en`

Datenauswertung in Echtzeit: Datenstromsysteme

Eine gänzlich neue Möglichkeit der Datennutzung bzw. -auswertung stellen heute Datenstromsysteme dar: Im Gegensatz zu Datenbanken, in denen Daten dauerhaft vorgehalten und daher mehrfach für Analysen herangezogen werden können, verarbeiten Datenstromsysteme Daten in Echtzeit. Dazu werden vorher definierte Analysen sofort durchgeführt und die Daten nicht längerfristig gespeichert, sondern höchstens kurzzeitig zwischengespeichert. Dies steht klar im Gegensatz zum Grundprinzip einer Datenbank, bei der die Dauerhaftigkeit des Datenbestands essenziell ist. Diese Vorgehensweise führt insbesondere dazu, dass Analysen schneller und speichereffizienter als bei Datenbanken durchführbar sind. Gleichzeitig büßen sie jedoch Flexibilität ein, da einmal analysierte Daten nicht für weitere Analysen zur Verfügung stehen, weswegen insbesondere klassische Data-Mining-Beispiele mit diesen Systemen nicht durchführbar sind. Datenstromsysteme verarbeiten einen Eingabedatenstrom[21] und erzeugen einen aus den Analyseergebnissen bestehenden und dem Eingabedatenstrom gleichgetakteten Ergebnisdatenstrom. In Abbildung 3.10 wird die Funktionsweise von Datenstromsystemen im Vergleich zu Datenbanken dargestellt. Wie dort erkennbar ist, können Datenstromsysteme auf ein einfaches Prinzip reduziert werden, obwohl Echtzeitanalysen im Allgemein ein eher komplexes Thema sind: Sie fungieren als Filter für den Datenstrom, bei dem die herausgefilterten Daten weiterverarbeitet und ggf. in aggregierter Form zwischengespeichert werden. Durch dieses einfache Prinzip ist es möglich, Schülerinnen und Schülern im Informatikunterricht die Möglichkeit zu geben, selbst Datenanalysen in Echtzeit durchzuführen und dahinterstehende Prinzip zu verstehen.

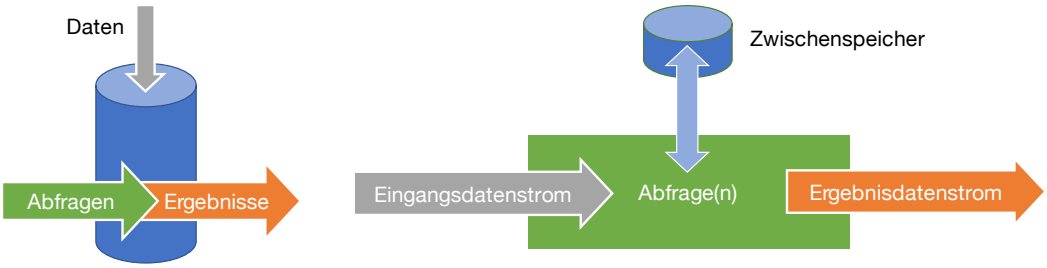


Abbildung 3.10: Vergleich des Funktionsprinzips von Datenbanksystemen (links) und Datenstromsystemen (rechts).

Ein Anwendungsgebiet von Datenstromsystemen sind Trendanalysen, wie sie unter anderem bei Twitter stattfinden. Auf den ersten Blick erscheint die von diesem Dienst verarbeitete Datenmenge aufgrund der maximalen Länge eines Tweets (ursprünglich 140 bzw. mittlerweile 280 Zeichen) eher gering. Doch durch die große Anzahl (über 6.000 Tweets pro Sekunde) und die umfangreichen enthaltenen Metadaten, fallen derzeit täglich etwa 260 GB (ca. 500 Byte pro Tweet) an Daten an. Die zeitnahe Analyse scheitert bei Nutzung einer herkömmlichen Datenbank schon an der in kurzer Zeit anfallenden großen Datenmenge, da die Speicherung größerer Datenmengen verteilt stattfinden muss, wodurch die Analyse ausgebremst wird (vgl. CAP-Theorem). Insbesondere ist solche datenbankbasierte

[21] *Golab und Özsu (2003)* definieren einen Datenstrom wie folgt: *„A data stream is a real-time, continuous, ordered (implicitly by arrival time or explicitly by timestamp) sequence of items.“*

Analyse hier jedoch auch aufgrund der hohen Dynamik der Daten kaum möglich: Um aktuelle Analyseergebnisse bereitzustellen, ist es nötig, Analysen in sehr schneller Taktfolge durchzuführen, wodurch die Datenbank stark belastet wird. Insbesondere bei hoher (schreibender) Aktivität der Nutzer müsste daher eine Abwägung getroffen werden zwischen der weiterhin performanten Nutzung des Dienstes und der Aktualität der – gerade zu diesen Zeitpunkten besonders interessanten – Datenanalysen. Mit Datenstromsystemen sind solche Analysen jedoch relativ einfach und effizient möglich, da sie kontinuierliche Analysen erlauben, wie sie in Datenbanken nicht möglich sind. Zwei Unterrichtswerkzeuge zu diesem Thema sowie eine beispielhafte Thematisierung im Unterricht werden in Kapitel 11 beschrieben.

3.2.3 Datenmanagement aus professioneller Sicht: Der Data Management Body of Knowledge

Die beschriebenen Themen des Datenmanagements offenbaren die Vielfalt dieses Fachgebiets. Zur weitergehenden fachlichen Fundierung und da in der weiteren Arbeit nochmals auf diese zurückgegriffen wird, wird im Folgenden ein Überblick über die derzeit einzige Charakterisierung des Datenmanagements aus fachlicher bzw. professioneller Sicht gegeben. Dieser *Data Mangement Body of Knowledge* (DAMA-DMBoK, *DAMA International (2017)*) wurde von der *Data Management Assocation International*[22], kurz DAMA, entwickelt und beschreibt das Fachgebiet durch verschiedene Konzepte, Wissensbereiche und Aktivitäten aus der Perspektive von *Data Management Professionals*[23].

Der DAMA-DMBoK hebt insbesondere 13 Prinzipien von Datenmanagement hervor *(DAMA International, 2017)*:

1. *Data is an asset with unique properties*: Daten sind ein Gut, dass sich von anderen insbesondere dadurch unterscheidet, dass es bei der Nutzung nicht verbraucht wird.
2. *The value of data can and should be expressed in economic terms*: Daten haben einen (ökonomischen) Wert, der stark von ihrer Qualität abhängt.
3. *Managing data means managing the quality of data*: Es ist im Datenmanagement zwingend nötig, eine hohe Datenqualität sicherzustellen.
4. *It takes Metadata to manage data*: Um Daten zu verwalten, müssen Metadaten genutzt werden.
5. *It takes planning to manage data*: Zur koordinierten Arbeit mit Daten ist eine Planung sowohl der Architektur als auch des Prozesses nötig.

[22] `http://dama.org`

[23] *„A Data Management Professional is any person who works in any facet of data management (from technical management of data throughout its lifecycle to ensuring that data is properly utilized and leveraged) to meet strategic organizational goals. Data management professionals fill numerous roles, from the highly technical (e. g. database administrators, network administrators, programmers) to strategic business (e. g., Data Stewards, Data Strategists, Chief Data Officers).“ (DAMA International, 2017)*

6. *Data management is cross-functional; it requires a range of skills and expertise*: Es werden sowohl technische als auch nicht-technische Fähigkeiten zum Umgang mit Daten benötigt.
7. *Data management requires an enterprise perspective*: Für eine möglichst hohe Effektivität, muss Datenmanagement unternehmensweit anstatt nur lokal angewendet werden.
8. *Data management must account for a range of perspectives*: Datenmanagement muss sich kontinuierlich weiterentwickeln und verschiedene Perspektiven auf Datenerzeugung und -nutzung berücksichtigen.
9. *Data management is lifecycle management*: Es muss der gesamte Lebenszyklus von Daten betrachtet und miteinbezogen werden.
10. *Different types of data have different lifecycle characteristics*: Die unterschiedlichen Anforderungen verschiedener Arten von Daten und ihre unterschiedlichen Charakteristika müssen berücksichtigt werden.
11. *Managing data includes managing the risks associated with data*: Da Daten ein Gut von hohem Wert darstellen, entstehen im Zusammenhang mit diesen auch Risiken, die in allen Fällen mit betrachtet werden müssen.
12. *Data management requirements must drive Information Technology decisions*: Die im Unternehmen eingesetzte Informationstechnologie muss so gewählt werden, dass die Anforderungen des Datenmanagements erfüllt werden.
13. *Effective data management requires leadership commitment*: Datenmanagement kann nur durch eine engagierte Führung vorangebracht werden.

Diese Prinzipien lassen einen klaren Blick auf die Vielfalt von Datenmanagement aus unternehmerischer Sicht zu. Gleichzeitig zeigt sich aber, dass diese Prinzipien aufgrund ihrer Sichtweise sicherlich nur sehr eingeschränkt auf allgemeinbildenden Schulunterricht übertragen werden können: Sie sind insbesondere mit Blick auf den Einsatz von Daten in (größeren) Unternehmen ausgewählt, ohne Aspekte wie die Sinnhaftigkeit für den Alltag von Schülerinnen und Schülern bzw. auch jeder anderen Person miteinzubeziehen. Die Betrachtung des DAMA-DMBoK kann daher zwar wichtige Aspekte des Fachgebiets charakterisieren, aber keine didaktische Aufarbeitung des Fachgebiets ersetzen. Dies zeigt sich auch an den im *DAMA Wheel* (vgl. Abbildung 3.11) dargestellten Wissensbereichen, für die dasselbe gilt. Eine detailliere Diskussion dieses Aspekts erfolgt im Rahmen der Klärung der Bedeutung von Datenmanagement im derzeitigen Informatikunterricht in Kapitel 6.

3.2.4 Ausblick auf die erwartete zukünftige Entwicklung

Wie die vorherigen Abschnitte zeigten, hat die Bedeutung von Daten und Datenmanagement in den letzten 15 Jahren deutlich zugenommen, ein Nachlassen oder Rückgang dieser Entwicklung ist aber auch in den nächsten Jahren nicht zu erwarten: Nach verschiedenen Prognosen (z. B. *Gantz und Reinsel (2012)*) wird sich die gesamte Datenmenge der Menschheit auch weiterhin etwa alle zwei Jahre verdoppeln. Entsprechend soll sich bis 2025 das Datenvolumen der Menschheit gegenüber 2016 auf 163 Zettabyte verzehnfachen *(Reinsel,*

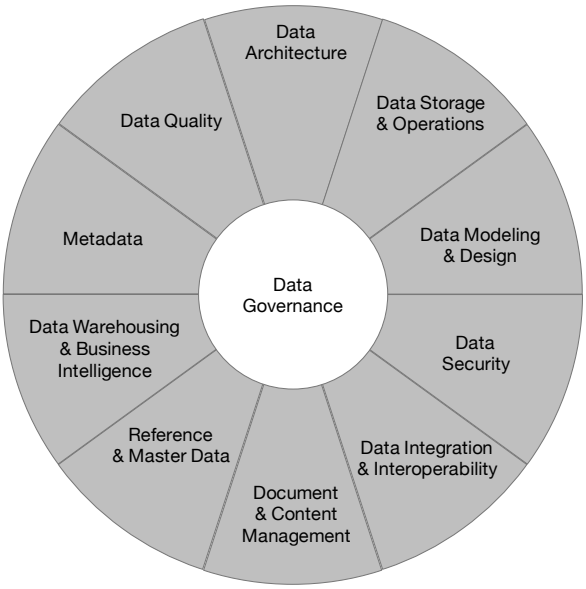


Abbildung 3.11: Data Management Functional Framework (*DAMA Wheel*; *DAMA International (2017)*).

Gantz und Rydning, 2017). Davon werden rund 60 % im Unternehmensumfeld erzeugt und verarbeitet. Der Anteil lebenskritischer Daten wird sich dabei bis 2025 gegenüber 2016 auf 20 % verdoppeln *(Reinsel, Gantz und Rydning, 2017)*. Die Menge der privat erzeugten und vorgehaltenen Daten wird insbesondere mit dem *Internet der Dinge* und der zunehmenden Anzahl vernetzter Informatiksysteme weiter ansteigen. Auch aus technologischer Sichtweise bahnen sich daher auch zukünftig weitere Neuerungen an: Beispielsweise bereiten sich Festplattenhersteller auf immer größere Speicherdichten vor und erwarten ein Anwachsen der Speicherkapazität und der benötigten Zugriffsgeschwindigkeiten. Gleichzeitig ist unklar, welche weiteren Einflüsse Entwicklungen wie z. B. Quantencomputing haben werden; es wird jedoch erwartet, dass höhere Verarbeitungsgeschwindigkeiten insbesondere zu einer Revolution im Maschinenlernen führen werden – was zu größeren verarbeitbaren Datenmengen führt und in der Folge auch eine weiter zunehmende Bedeutung von Datenmanagement nach sich zieht. Es bleibt daher zu erwarten, dass die Entwicklungen im Datenmanagement weiterhin ähnlich rasant fortschreiten wie in den letzten 10–15 Jahren. Es werden auch zukünftig, bedingt durch die enormen Datenmengen, die verwaltet werden müssen, immer höhere Anforderungen an Datenmanagementmethoden und -technologien gestellt werden, die auch mit heutigen Systemen bisher nur eingeschränkt erfüllt werden können. Durch die zunehmende Verbreitung des Internets der Dinge, maschinellen Lernens und von Mustererkennung, aber auch durch eine weiter zunehmende Nutzung mobiler Geräte ergeben sich vielfältige neue Herausforderungen und Chancen.

Datenmanagement wird sich daher auch in den kommenden Jahren stetig weiterentwickeln und bleibt ein innovationsträchtiges Fach- und Forschungsgebiet der Informatik, das zu vielen technologischen Neuerungen führen wird. Die heutige Breite und Relevanz dieses Fachgebiets wird durch den in diesem Bereich derzeit stark auf Datenmodellierung und Datenbanken ausgerichteten Informatikunterricht nur eingeschränkt adressiert. Um auch

weiterhin einen zukunftssicheren Informatikunterricht zu ermöglichen, der im Sinne einer Wissenschaftspropädeutik einerseits die zentralen Aspekte des Faches aufgreift, aber diese gleichzeitig unter Berücksichtigung zentraler Kriterien eines allgemeinbildenden Unterrichts auswählt, findet in dieser Arbeit eine ausführliche fachdidaktische Untersuchung des Datenmanagements statt. In deren Rahmen werden insbesondere zentrale Begriffe geklärt, die als langlebig erwarteten und im Fachgebiet zentralen Konzepte extrahiert und diese anhand von unterrichtspraktischen Beispielen unterlegt.

4 Daten und Datenmanagement in Informatikdidaktik und -unterricht

Während die fachlichen Veränderungen im Datenmanagement bereits über ein Jahrzehnt andauern, waren diese in der informatikdidaktischen Forschung, aber auch im Schulunterricht, bislang eher von geringer Bedeutung. Da sich der informatikdidaktische Forschungsstand in diesem Bereich insbesondere auf eine stark praxisorientierte Betrachtung des Themengebiets Datenbanken stützt und somit eine klare Trennung der informatikdidaktischen und unterrichtlichen Perspektive kaum möglich ist, werden beide in diesem Kapitel teils gemeinsam betrachtet. Im ersten Teil des Kapitels wird daher die gemeinsame Entwicklung von Informatikunterricht und -didaktik seit Etablierung des informatikdidaktischen Diskurses zu Datenbanken charakterisiert. Im zweiten Teil wird der als Fundament für diese Arbeit dienende informatikdidaktische Forschungstand beschrieben und als Basis für die weitere Arbeit im dritten Abschnitt des Kapitels in einer qualitativen Analyse von Lehrplänen und Curricula die derzeitige Bedeutung von Themen aus dem Umfeld des Datenmanagements untersucht.

4.1 Gegenstand informatischer Bildung

Im Informatikunterricht wird der Gegenstandsbereich Daten üblicherweise aus zwei Perspektiven betrachtet: Einerseits stellen Daten ein wichtiges Thema in der Programmierung dar, in der es u. a. um die konkrete algorithmische Verarbeitung von Daten geht, aber beispielsweise auch Datenstrukturen eine zentrale Rolle spielen. Andererseits werden Daten aber auch im Bereich der Datenverwaltung mit Hilfe von Datenbanken aufgegriffen. Im Sinne der Schwerpunktsetzung dieser Arbeit, wird hier auf den zweiteren Bereich fokussiert: Das Thema Datenbanken, als ursprünglicher Vertreter des heutigen Fachgebiets Datenmanagement, gewann in Informatikunterricht und Informatikdidaktik seit Mitte der 1980er Jahre an Bedeutung: 15 Jahre nach der Vorstellung des Konzepts der *relationalen Datenbanken* durch *Codd (1970)* wurde dieses Thema sukzessive in verschiedenen stark praxisorientierten Arbeiten aufgegriffen und in vielen Bundesländern in den Informatikunterricht miteinbezogen.

In den frühen Jahren des Datenbankunterrichts stand oft die Nutzung von Datenbanken als prototypisches Informatiksystem im Vordergrund: So war es nicht unüblich, dass zentrale Lernziele unter anderen die Nennung und Beschreibung der *„wesentlichen Funktionseinheiten eines Datenbanksystems“ (Käberich und Steigerwald, 1986)* und das Kennenlernen und Bedienen eines exemplarischen Datenbanksystems waren. Gleichzeitig wird bis heute auch traditionell häufig das Thema Datenschutz und dessen Grenzen im Datenbankunterricht mitbetrachtet (z. B. *„die Aussagen des Datenschutzgesetzes [...] kennen“* sowie *„erkennen, dass es möglich ist, mittels einfacher Abgleichtechniken scheinbar anonyme Daten zu personalisieren“ (Käberich und Steigerwald, 1986)*). *Witten (1994)* fasste die Situation zu Anfang der

1990er Jahre wie folgt zusammen: „*Es gibt bekanntlich zahlreiche Unterrichtsvorschläge zu den Themen Dateiverwaltung, (PC-)Datenbanken und Datenschutz für die informationstechnische Grundbildung und den Informatikunterricht in der Sekundarstufe I.*“ Trotz dieser vielfältigen Unterrichtsvorschläge, die zum Teil versuchen, anstatt einer Orientierung an den technischen Systemen die informatischen Ideen hinter Datenbanken in den Vordergrund zu stellen (z. B. Transaktionalität, Abstraktion und Zugriffskontrolle bei *Modrow (1996)*), wird weiterhin oft stark auf die Nutzung von Datenbanksystemen fokussiert: Spätestens 1995, mit Einzug des Programms *Access* in das *Microsoft-Office*-Programmpaket, schien der Umgang mit dieser Anwendung als ähnlich wichtig erachtet zu werden, wie die im Rahmen von Grundlagen der Informatik oft umfangreich thematisierten Textverarbeitungs- und Tabellenkalkulationswerkzeuge. Beispielsweise wurde dabei – stark an der Bedienoberfläche von *Microsoft Access* orientiert – die Erstellung, Bearbeitung und Nutzung von Tabellen in einer Datenbank thematisiert (vgl. bspw. *Schuh et al. (2002)*). Informatische Ideen, wie Primärschlüssel oder Indizierung, wurden dabei nur rudimentär miteinbezogen, soweit und sobald sie für die Nutzung der Systeme notwendig waren.

Seit der immer stärkeren Abkehr von einer reinen informationstechnischen Grundbildung und einer damit einhergehenden Orientierung an der Nutzung von Informatiksystemen, sind auch im datenbankorientierten Unterricht die informatischen Konzepte weiter in den Vordergrund gerückt. Je nach konkreter Ausprägung des Informatikunterrichts stellt daher heute die Datenmodellierung einen stärkeren Schwerpunkt dar, obwohl diese beispielsweise auch bereits bei *Modrow (1996)* ein wichtiger Ausgangspunkt für ein tiefergehendes Verständnis von Datenbanken ist. Gleichzeitig wird üblicherweise versucht, informatische Prinzipien hinter Datenbankmanagementsystemen, insbesondere am Beispiel von Konsistenz, zu thematisieren. Gleichzeitig spielt aber auch die Sprache SQL eine zentrale Rolle, die meist die einzige vom imperativen Programmierparadigma abweichende Programmiersprache im Informatikunterricht ist. Diese Schwerpunkte stellen sowohl national als auch international einen klaren Konsens dar, wie sich auch in Abschnitt 4.3 in einer Analyse von Curricula und Bildungsstandards zeigt. Eine umfassende informatikdidaktische Fundierung des Datenbankunterrichts fand jedoch weder national noch international statt.

Zur Erreichung der Unterrichtsziele werden oft professionelle Datenbanken eingesetzt, z. B. MySQL in Kombination mit einer Verwaltungsoberfläche wie HeidiSQL oder phpMyAdmin. Es kommen jedoch auch integrierte Lösungen wie Microsoft Access oder OpenOffice.org Base zum Einsatz. Während gerade bei den zuerst erwähnten Kombinationen zwangsläufig ein eher an einer Abfragesprache orientierter Zugang im Vordergrund steht, rückt diese bei Nutzung einer der eher grafisch orientierten Anwendungen stärker in den Hintergrund. Aufgrund der in beiden Fällen vorhandenen optischen Ähnlichkeiten mit Tabellenkalkulationswerkzeugen und der häufigen curricularen Nähe der beiden Themen werden diese Systemtypen im Unterricht häufig in Abgrenzung voneinander betrachtet. Der Unterricht wird dabei oft an existierenden datenbankbasierten Systemen orientiert und dadurch mehr oder weniger stark in der Lebenswelt der Schülerinnen und Schüler verankert: So schlägt ein Schulbuch den Aufbau eines Musikkatalogs oder einer eigenen relativ umfangreichen Versandhandel-Datenbank vor *(Engelmann, 2006)*, ein anderes greift

das Thema der Schulbibliothek auf *(Fischer, Knapp und Neupert, 2006)*. In der Unterrichtsrealität spielen jedoch auch andere moderne Kontexte, insbesondere soziale Netzwerke, eine wichtige Rolle. Beispielsweise stellt *Dorn (2017)* unter dem Motto „*durch Datenbanken Möglichkeiten und Risiken in sozialen Netzwerken verstehen*“ ein für die Schule konzipiertes Werkzeug zur Simulation eines sozialen Netzwerks vor, das Schülerinnen und Schülern den Blick hinter die Kulissen und die Nutzung der dort gespeicherten Daten erlaubt. Ähnlich konzipierte Unterrichtsansätze existieren auch beispielsweise im Rahmen von Informatik im Kontext (z. B. VideoCenter *(Penon, 2013)*, FitnessCenter *(Penon, 2017)*).

Neben Datenmodellierung und der Arbeit mit Datenbanken, stellen auch gesellschaftlich relevante Aspekte des Umgangs mit Daten ein im Unterricht zentrales Thema im betrachteten Bereich dar: Wie auch die in Abschnitt 4.3 beschriebene Curriculumsanalyse zeigt, wird insbesondere das Thema Datenschutz in vielen Lehrplänen und Curricula berücksichtigt. Entsprechend gibt es bereits verschiedene Ansätze, dieses Thema im Informatikunterricht aufzugreifen. Ein besonders bekannter Vertreter ist das in seiner ursprünglichen Version bereits vor 30 Jahren erschienene *Planspiel Datenschutz (Hammer und Prodesch, 1987)*, das über die Jahre vielfältig angepasst wurde (z. B. *Dietz und Oppermann (2011)*). Dieses ist, wie auch die meisten anderen Ansätze in diesem Bereich, insbesondere darauf ausgelegt, die Schülerinnen und Schüler mit lebensnahen Szenarien zu konfrontieren und ihnen anhand dieser die Möglichkeit zu geben, exemplarisch zu erfahren, *„welche Datenspuren sie wo hinterlassen und wer auf diese Daten zugreifen kann“ (Dietz und Oppermann, 2011)*. Entsprechend tragen diese Ansätze zu einem bewussten Umgang mit Daten wesentlich bei, können jedoch noch keinen umfassenden Kompetenzerwerb fördern, der eine tiefgreifende Beurteilung, Beeinflussung und Nutzung heute entstehender Möglichkeiten erlaubt.

4.2 Gegenstand informatikdidaktischer Forschung

Trotz der enormen Veränderungen im Datenmanagement und der Bedeutung dieses Fachgebiets für das tägliche Leben, wurden seit den 1990er Jahren weder Datenmanagement im Allgemeinen, noch spezielle Themen aus diesem Gebiet in der informatikdidaktischen Forschung ausführlicher aufgegriffen: Stattdessen werden in stark unterrichtspraktisch orientierten Arbeiten verschiedene Zugänge zum Themenbereich Datenmodellierung und Datenbanken vorgeschlagen und erprobt (vgl. Abschnitt 4.1).

Auch aktuellere Forschungsarbeiten aus dem betrachteten Umfeld konzentrieren sich insbesondere auf neue bzw. andere Zugänge zum Thema Datenbanken und auf die Konzeption entsprechender Kurse: Beispielsweise wurde durch *Antonitsch (2007)* ein eher forschender Zugang zu Datenbanken vorgeschlagen und erfolgreich erprobt, der die Rückkopplung zwischen der Modellierung der Datenbankstruktur und deren Implementierung in den Fokus nimmt. Dazu werden beide Bereiche eng verschränkt im Unterricht thematisiert und so die Arbeit mit dem Datenbanksystem und die Entwicklung von Datenmodellen stärker in Zusammenhang gebracht, indem die Lernenden durch eigene Anfragen lernen, wie Da-

tenbanken strukturiert werden müssen, um die gewünschten Informationen abspeichern und abfragen zu können.

Andere Arbeiten aus dem Umfeld des Datenmanagements fokussieren weniger auf konzeptuelle Veränderungen des Unterrichts, sondern versuchen als Best-Practice-Beispiele neue Inhalte oder Beispiele aufzugreifen und diese im Unterricht zu vermitteln, oft jedoch ohne eine fundierte informatikdidaktische Basis. Beispielsweise diskutieren *Buffum et al. (2014)* die Einführung eines Unterrichtsmoduls zum Thema Big Data für die U. S.-amerikanische Middle School. Dazu wurde ein Curriculumsdesign vorgeschlagen und mit einer kleinen Gruppe von Lehrerinnen und Lehrern bzw. Schülerinnen und Schülern erprobt. Dieses Beispiel zeigt zwar, dass grundlegende Aspekte der Themen, die in Zusammenhang mit dem Schlagwort Big Data stehen, im Unterricht umsetzbar scheinen.

Über die fachlichen Inhalte und Kompetenzen hinaus, existieren auch verschiedene Ansätze zur Förderung eines kritischen Weltbildes der Schülerinnen und Schüler im Kontext Daten: Beispielsweise befassen sich *Acker und Bowler (2017)* mit der Konzeption eines Workshops zur Förderung eines kritischen und aufmerksamen Umgangs mit Datenspuren insbesondere im Kontext sozialer Medien.

Neben der konkreten unterrichtlichen Umsetzung wird derzeit in verschiedenen Arbeiten auch eine curriculare Integration datenbezogener Inhalte und Kompetenzen diskutiert und dafür eine Basis geschaffen, indem beispielsweise *Heinemann et al. (2018)* ein Data-Science-Curriculum für Sekundarschulen entwickeln und erproben, das auch verschiedene Aspekte des Datenmanagements berücksichtigt. Da Data Science üblicherweise nicht nur eine informatische Perspektive einschließt, sondern auch die eines konkreten Anwendungsfachs sowie die der Mathematik, existiert hier auch ein thematischer Bezug zur mathematikdidaktischen Forschung: Aus mathematischer Perspektive geht es jedoch weniger um die Datenspeicherung und das Datenmanagement, sondern eher um mathematische Algorithmen, Korrelation und Kausalität, die Bestimmung der Aussagekraft von Stichproben und Ähnliches *(vgl. Ridgway, Nicholson und Gal, 2018)*. Die aus dieser Perspektive unter dem Begriff *Statistical Literacy* zusammengefassten daten- und statistikbezogenen Inhalte und Kompetenzen wurden zwar zum Teil schon ausführlicher betrachtet *(vgl. Sharma, 2017)*, können aus informatikdidaktischer Perspektive jedoch allenfalls zur Abgrenzung herangezogen werden.

Aus informatikdidaktischer Sicht existiert somit eine klare Forschungslücke: Trotz der enormen Entwicklungen der letzten Jahre wurde bisher keine fachdidaktische Diskussion des Feldes Datenmanagement bzw. der darin enthaltenen Themen angestoßen, sondern der tradierte Stand beibehalten. Eine Beurteilung der Relevanz und Eignung sowohl von neuen als auch tradierten Themen des Datenmanagements für den allgemeinbildenden Informatikunterricht ist daher auf Basis des bisherigen nur wenig umfangreichen Forschungsstands nicht möglich.

Über den thematischen Forschungsstand hinaus, existieren jedoch verschiedene Arbeiten, die ähnliche Forschungsdesiderate verfolgen: Eine ähnliche *thematische Forschungslücke* existierte beispielsweise auch, als die *objektorientierte Programmierung* Einzug in die Informatik

hielt und andere Programmierparadigmen ablöste *(vgl. z. B. Schulte, 2003; Brinda, 2004; Diethelm, 2007; Schwill, 1995)*, aber auch in den Bereichen *agile Methoden (vgl. Kastl und Romeike, 2015)*, *Debugging (vgl. Michaeli und Romeike, 2017)* oder *Eingebettete Systeme und Physical Computing (vgl. Przybylla, 2018)*. Zur Adressierung einer solchen Forschungslücke wird häufig sowohl ein theoretisch orientierter Beitrag geleistet, aber auch unterrichtspraktische Untersuchungen durchgeführt. Diese in fachdidaktischen Arbeiten übliche Zielsetzung wird auch in dieser Arbeit verfolgt.

4.3 Datenmanagement in Bildungsstandards und Curricula

Trotz der eher geringen Bedeutung von Datenmanagement in der aktuellen fachdidaktischen Diskussion, ist das tradierte Unterrichtsthema *Datenbanken* in Sekundarschulen auch heute noch weit verbreitet und hat in den letzten Jahren sogar noch an Bedeutung gewonnen. Die Relevanz dieses Themas zeigt sich sowohl auf nationaler als auch internationaler Ebene bei einem Blick in Bildungsstandards und Curricula: Beispielsweise legen die Empfehlungen für Bildungsstandards der Gesellschaft für Informatik e. V. für die Sekundarstufe I nahe, dass Schülerinnen und Schüler *„Bei der Verwendung eines Datenbanksystems [...] beim Anlegen von Tabellen über die Wertebereiche der Attribute [entscheiden], wobei sie nun ausdrücklich zwischen Zahlen und Texten unterscheiden müssen" (Arbeitskreis Bildungsstandards, 2008)*. In der Sekundarstufe II sollen die Schülerinnen und Schüler beispielsweise *„zu einem Realitätsausschnitt ein Datenmodell [erstellen] und [...] es als Datenbank implementieren" (Arbeitskreis Bildungsstandards SII, 2016)* können. Auf internationaler Ebene wird das Thema unter anderem in den K–12 Informatikstandards der ACM/CSTA aufgegriffen: So lautet ein Ziel für die Jahrgangsstufen 9–10, dass die Schülerinnen und Schüler *„Techniken diskutieren, die zum Speichern, Verarbeiten und Abrufen von Daten genutzt werden (z. B. Dateien, Datenbanken, Data Warehouses)" (CSTA Standards Taskforce, 2016)*[24]; in den Jahrgangsstufen 11–12 wird dieses weiter vertieft, indem die Schülerinnen und Schüler *„Datenanalysen nutzen, um signifikante Muster in komplexem Systemen zu identifizieren (z. B. Nutzen von vorhandenen Datensätzen und Erschließen ihrer Bedeutung)" (CSTA Standards Taskforce, 2016)*[24]. Bereits diese Beispiele zeigen unterschiedliche Dimensionen des Datenmanagementunterrichts. Erweitert man das Themenfeld von *Datenbanken und Datenmodellierung* hin zu *Datenmanagement* im Allgemeinen, ist durch die zunehmende Auswahl an Unterrichtsgegenständen eine noch deutlichere Varianz zu erwarten. Es ist daher bisher unklar, welche Bedeutung das Fachgebiet Datenmanagement heute im Informatikunterricht einnimmt und welche Themen in diesem vertreten sind. Eine Hypothese kann jedoch die historische Betrachtung des aus dem Themengebiet *Datenbanken* hervorgegangenen Fachgebiets *Datenmanagement* sowie der informatikdidaktischen Forschung der letzten beiden Jahrzehnte liefern: Diese legt den Schluss nahe, dass zwar datenbanknahe Themen im Unterricht (mehr oder weniger ausführlich) thematisiert werden, dieser aufgrund der fehlenden fachdidaktischen Forschung und Diskussion zu moderneren Themen des Datenmanagements jedoch vermutlich nur in Einzelfällen klar über

[24] Eigene Übersetzung des Autors

klassische Datenbanken hinausgeht. Entsprechend ist zu erwarten, dass neuere Themen diese in den entsprechenden Bildungsstandards und Curricula kaum erkennbar sind.

4.3.1 Ziele der Untersuchung

Als Basis für diese Arbeit und die informatikdidaktische Diskussion zum Fachgebiet Datenmanagement allgemein, ist es daher essenziell, die zweite Forschungsfrage dieser Arbeit zu beantworten, indem untersucht wird, welche Bedeutung Datenmanagement bereits heute in verschiedenen Lehrplänen, Curricula und Bildungsstandards hat, um – unter der Annahme, dass sich Unterricht stark an diesen Vorgaben orientiert – Rückschlüsse auf diesen zu ziehen. Um die Fragestellung weiter zu konkretisieren, wird sie in folgende Unterfragen aufgeteilt:

F1 *„Welche Themen des Datenmanagements sind bereits heute in Lehrplänen, Curricula und Bildungsstandards vertreten?“*

F2 *„Wie unterscheiden sich die aktuelle schulische und wissenschaftliche Sichtweise auf das Feld Datenmanagement?“*

Bei dieser Untersuchung müssen daher zwei Perspektiven betrachtet werden: Einerseits kann aus Sicht der betrachteten Dokumente eine Bestandsaufnahme der aktuellen Themen in diesen stattfinden, andererseits ist es auch wichtig, aus fachlicher Sicht auf diese Vorgaben für den Unterricht zu blicken, um Themen zu erkennen, die dort bisher nicht oder nur am Rande vertreten sind.

4.3.2 Untersuchungsmethode: Qualitative Inhaltsanalyse

Insbesondere zur Beantwortung der ersten Fragestellung ist eine explorative Analyse anerkannter Bildungsstandards sowie verschiedener Curricula, beides sowohl auf nationaler als auch internationaler Ebene, hilfreich. Um auch die zweite Fragestellung miteinzubeziehen, kann gleichzeitig ein Abgleich mit den Themen des Datenmanagements aus fachlicher Sicht stattfinden, indem die in den Bildungsstandards und Curricula gefundenen Themen in eine wissenschaftliche Charakterisierung des Fachgebiets eingeordnet werden. Auf diese Weise kann gleichermaßen sowohl die Extension von Datenmanagement im Informatikunterricht als auch die Abdeckung des Fachgebiets durch diesen erfasst werden.

Ähnlich gelagerte Analysen wurden bereits in verschiedenen Kontexten durchgeführt, insbesondere die aus der evidenzbasierten Medizin stammenden und heute auch im Softwareengineering häufig eingesetzten *Mapping Studies* haben ein ähnliches Ziel. Im Vergleich mit den hier angestrebten Zielen, sind diese jedoch auf einer höheren Ebene angesiedelt: Es geht dabei insbesondere um die systematische und objektive Erforschung der Art und des Umfangs der Forschung in einem Fachgebiet, insbesondere mit dem Ziel Forschungslücken zu identifizieren *(vgl. Budgen et al., 2008)*. Dieser Ansatz kann auf die hier angestrebten Ziele eingeschränkt übertragen werden: Da eine systematische und objektive Erforschung der

Bedeutung und des Umfangs von Datenmanagement im Informatikunterricht angestrebt wird, kann durch Nutzung dieses Ansatzes prinzipiell eine Beantwortung der ersten Fragestellung dieser Studie erfolgen. Um die Zweite zu beantworten, müsste jedoch eine weitere Studie durchgeführt bzw. ein Vergleich der Ergebnisse der Mapping-Studie mit einer Charakterisierung des Fachgebiets aus fachlicher Sicht stattfinden. Es wurde sich daher an dieser Stelle gegen einen solchen Ansatz entschieden und stattdessen ein Weg gewählt, der zur Erforschung beider Fragestellungen gleichermaßen dienlich ist: Aufgrund des Ziels, auf explorative Weise eine umfangreiche Charakterisierung derzeitiger Unterrichtsthemen zu erstellen, wurde die *Qualitative Inhaltsanalyse* nach *Mayring (2010)* mit deduktiver Kategorienbildung und induktiver Erweiterung des Kategoriensystems als methodische Grundlage für diese Analyse gewählt. Obwohl sich gerade bei der Anwendung im Kontext dieser Studie verschiedene Gemeinsamkeiten der Qualitativen Inhaltsanalyse mit den Mapping Studies zeigen, ermöglicht die deduktive Ableitung des Kategoriensystems und die Einordnung der gefundenen Begriffe in dieses nicht nur die Beantwortung der ersten Frage, sondern adressiert gleichzeitig auch die Zweite, da als Basis für das Kategoriensystem eine fachliche Charakterisierung des Fachgebiets herangezogen werden kann. Orientiert an der Methodik nach Mayring wurde der Analyseprozess in mehrere Schritte geteilt, die im Folgenden kurz charakterisiert werden (vgl. auch Abbildung 4.1).

Abbildung 4.1: Ablauf der Analyse der Qualitativen Inhaltsanalyse zur Bedeutung von Datenmanagement in Bildungsstandards und Curricula.

- **Auswahl der Materialien:** Ein zentraler Schritt der qualitativen Inhaltsanalyse nach *Mayring (2010)* ist die Auswahl geeigneter Literatur bzw. Materialien für die Analyse. Da dieses Material nur selten mehr als eine Stichprobe sein kann, wird hier eine möglichst repräsentative Auswahl angestrebt, die valide Rückschlüsse auf die betrachtete Grundgesamtheit erlaubt. In dem hier betrachteten Fall der Analyse von Curricula und Bildungsstandards sollen daher verschiedene Ausrichtungen des Informatikunterrichts miteinbezogen und eine Auswahl getroffen werden, die sowohl einen Blick in die Tiefe als auch in die Breite erlaubt. Dazu werden sowohl nationale als auch internationale Dokumente einbezogen werden.

- **Entwicklung des Kategoriensystems:** Nach Auswahl des Materialkanons gilt es als Nächstes, zu entscheiden, wie das Kategoriensystem aufgebaut wird: Die qualitative Inhaltsanalyse erlaubt generell sowohl die induktive Erstellung eines Kategoriensystems auf Grundlage der analysierten Dokumente während der eigentlichen Durchführung der Analyse, als auch eine deduktive Ableitung des Kategoriensystems aus bereits existierenden Arbeiten. Auch eine Kombination beider Varianten ist zulässig und häufig anzutreffen. Bei Wahl eines (teils) deduktiven Verfahrens wird in dieser Phase auch das (ggf. vorläufige) Kategoriensystem entwickelt.

- **Festlegung von Kodiereinheit und Auswahlkriterien:** Als letzter Schritt vor der eigentlichen Analyse müssen die Kodiereinheit, d. h. minimale/maximale Länge bzw. Umfang zu kodierender Textstücke, sowie Auswahlkriterien, die bestimmen, wann eine Textpassage kodiert wird, festgelegt werden.
- **Durchführung:** Nach den vorbereitenden Schritten folgt die eigentliche Analyse der Dokumente. Diese kann, falls die analysierten Dokumente nicht aufeinander aufbauen, in willkürlicher Reihenfolge erfolgen. Die gefundenen Vorkommen von Begriffen/Textstellen, die den Auswahlkriterien genügen, werden in das Kategoriensystem eingeordnet (ggf. unter Nutzung induktiver Ergänzungen).
- **Auswertung:** Abschließend werden die Ergebnisse unter Berücksichtigung der zugrundeliegenden Fragestellungen, aber auch des unterschiedlichen Charakters und Detailgrads der analysierten Dokumente sowie ihrer Verbindlichkeit, ausgewertet und zusammengefasst.

4.3.3 Durchführung und Auswertung

Auswahl des Materials

Um der Analyse einen repräsentativen Literaturkanon zugrunde zu legen, wurden verschiedene zum Analysezeitpunkt (Juni/Juli 2014) aktuellen Lehrpläne der Gymnasien in Deutschland, sowie eine Auswahl verschiedener internationaler Curricula, aber auch nationale sowie internationale Bildungsstandards für Informatikunterricht in den Sekundarstufen ausgewählt. Dadurch werden einerseits verschiedene Orientierungen, Charakteristika und Möglichkeiten des Informatikunterrichts in verschiedenen (Bundes-)Ländern berücksichtigt, gleichzeitig sind aber auch, durch Schwerpunktsetzung auf die deutschen Curricula, Rückschlüsse auf die spezielle Curriculumsentwicklung in Deutschland möglich. Die als Literaturkorpus gewählten Dokumente sind, zusammen mit Kürzeln zur Referenzierung in der Analyse, in Tabelle 4.1 aufgelistet.

Entwicklung des Kategoriensystems

Zur Ermittlung der aktuellen Themen (*F1*), wäre eine induktive Erstellung des Kategoriensystems geeignet, was einem Ansatz wie bei den zuvor erwähnten Mapping Studies sehr nahekommt. Eine deduktive Erstellung des Kategoriensystems wäre hingegen insbesondere zur Analyse der Lücke zwischen Fachwissenschaft und Unterricht (*F2*) dienlich, da dabei ein Vergleich zwischen den den Unterricht prägenden Dokumenten und dem derzeitigen Stand im Fachgebiet möglich wird. Eine alleinige deduktive Herleitung des Kategoriensystems und Anwendung auf die Dokumente schränkt jedoch die Analyse insofern ein, als Aspekte, die zwar in diesen erwähnt werden, im deduktiv aus einer fachlichen Quelle abgeleiteten Kategoriensystem jedoch nicht vorkommen, nicht kodiert werden könnten. Dies würde der ersten Fragestellung entgegenstehen. Eine Mischung beider Varianten, d. h.

Kürzel	Dokument	Quellenangabe
[EPA]	Einheitliche Prüfungsanforderungen Informatik	*(Kultusministerkonferenz, 2004)*
[GI]	Grundsätze und Standards für die Informatik in der Schule: Bildungsstandards Informatik für die Sekundarstufe I	*(Arbeitskreis Bildungsstandards, 2008)*
[K12]	ACM/CSTA K–12 Computer Science Standards	*(Seehorn et al., 2011)*
[BY]	Lehrplan für das Fach Natur & Technik bzw. Informatik des achtjährigen Gymnasiums in Bayern	*(Staatsinstitut für Schulqualität und Bildungsforschung, 2009)*
[HE]	Lehrplan für das Fach Informatik des Bildungsgangs Gymnasium in Hessen	*(Hessisches Kultusministerium, 2010)*
[HH]	Bildungsplan Gymnasium Sekundarstufe I für das Informatik Wahlpflichtfach sowie Rahmenplan Informatik für die Gymnasiale Oberstufe in Hamburg	*(Behörde für Schule und Berufsbildung, Hamburg, 2009; Behörde für Schule und Berufsbildung, Hamburg, 2011)*
[NRW]	Kernlehrplan Informatik für die Sekunderstufe II Gymnasium/Gesamtschule in Nordrhein-Westfalen	*(Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen, 2013)*
[RLP]	Lehrplan Informatik an Gymnasien und Integrierten Gesamtschulen (Sekundarstufe I und gymnasiale Oberstufe) in Rheinland-Pfalz	*(Ministerium für Bildung, Wissenschaft, Jugend und Kultur RLP, 2011b; Ministerium für Bildung, Wissenschaft, Jugend und Kultur RLP, 2011a)*
[AT]	Lehrplan für das Fach Informatik in der Oberstufe der allgemeinbildenden höheren Schule in Österreich	*(Bundesministerium für Bildung, 2012)*
[CA]	Ontario Curriculum Science & Technology, Ontario, Canada	*(Ontario Ministry of Education, 2007)*
[CAS]	Computing at School Curriculum, Vereinigtes Königreich	*(Computing At School, 2012)*
[IS]	Curriculum for a High-School Program in CS, Israel	*(Gal-Ezer und Harel, 1999)*

Tabelle 4.1: Analysierte Curricula und Lehrpläne zur Untersuchung des aktuellen Stands von Datenmanagement im Informatikunterricht.

eine deduktive Erstellung des initialen Kategoriensystems, das aber im Laufe der Analyse – sofern nötig – induktiv ergänzt wird, berücksichtigt hingegen beide Analyseziele gleichermaßen, sodass die Entscheidung auf diese Möglichkeit fiel.

Als Grundlage für das initiale deduktiv erstellte Basissystem wurde der zuvor schon beschriebene Data Management Body of Knowledge *DAMA International (2010)* ausgewählt, da dieser derzeit die einzige umfangreiche Charakterisierung des Fachgebiets darstellt und von diversen Experten aus dem Fachgebiet erstellt wurde. Obwohl diese Charakterisierung aus professioneller Perspektive erstellt wurde und daher nicht direkt auf die Schule übertragbar ist, bietet das DAMA-DMBoK trotzdem eine breite Übersicht über das Feld und ist daher als Grundlage für die Untersuchung geeignet. Es muss jedoch berücksichtigt werden, dass nicht alle dort beschriebenen Aspekte auch sinnvoll im allgemeinbildenden Schulunterricht thematisiert werden können, sodass aus entsprechenden Lücken nicht direkt eine Notwendigkeit abgeleitet werden kann, diese auch zu schließen. Als Basis für das Kategoriensystem bieten sich insbesondere die im DAMA-DMBoK beschriebenen Funktionen des Datenmanagements (vgl. auch Abbildung 3.11) an, da diese nach der Beschreibung der Autoren die zentralen Konzepte des Datenmanagements darstellen. Somit kann erwartet werden, dass auf dieser Ebene die größten Überschneidungen mit den Themen des allgemeinbildenden Schulunterrichts existieren, während sie gleichzeitig auch das Fachgebiet gut charakterisieren. Die folgenden Funktionen von Datenmanagement werden im DAMA-DMBoK beschrieben[25]:

- **Data Governance:** Strategische Planung und Steuerung des Datenmanagementprozesses auf Unternehmensebene.
- **Data Architecture Management:** Planung und Verwaltung komplexer Datenstrukturen.
- **Data Development:** Verwaltung strukturierter Daten, vom Design bis hin zur Wartung (insbesondere unter Nutzung von Datenbanksystemen).
- **Data Operations Management:** Wartung, Support und Administration von Datenmanagementsystemen.
- **Data Security Management:** Datensicherheits- und Datenschutzaspekte in Bezug auf Datenmanagement
- **Reference & Master Data Management:** (organisatorische) Sicherstellung einer konsistenten Datennutzung in verschiedenen Systemen.
- **Data Warehousing & Business Intelligence Management:** Nutzung von Datenanalysen zur Entscheidungsfindung, insbesondere unter Nutzung von „Data Warehouse"-Systemen und „Business Intelligence"-Methoden.

[25] Die hier beschriebenen Funktionen des Datenmanagements weichen geringfügig von den in Abschnitt 3.2.3 beschriebenen ab, was dadurch begründet ist, dass zur Zeit der Studie die erste Auflage des DAMA-DMBoK aktuell war und daher als Basis herangezogen wurde, für die Erläuterungen des Fachgebiets in dieser Arbeit aber schon die neue Auflage zur Verfügung stand. Die Unterschiede liegen jedoch eher im Detail, sodass keine relevanten Einflüsse auf die Ergebnisse anzunehmen sind.

- **Document & Content Management:** Speicherung, Strukturierung und Verwaltung unstrukturierter Daten (hauptsächlich unter Nutzung anderer Speicher als Datenbanken).
- **Meta-data Management:** Nutzung von Metadaten sowie Steuerung wo und wie diese eingesetzt werden.
- **Data Quality Management:** Maßnahmen zur Sicherstellung, Erhöhung und Kontrolle der Qualität der eingesetzten Daten.

Bei Betrachtung dieser Funktionen kann schnell festgestellt werden, dass diese zum Teil relativ gut zum Schulunterricht passen, z. B. sind Aspekte des *Data Security Management* in diesem durchaus vorstellbar, u. a. indem Zugriffsrechte thematisiert werden. Andere können jedoch nicht direkt übertragen werden: In der dargestellten Ausprägung stellen sie eher speziell in Unternehmenskontexten wichtige Aspekte dar, die nur eingeschränkt auf den allgemeinbildenden Unterricht übertragbar sind. Dabei handelt es sich insbesondere um die Thematisierung komplexer Datenarchitekturen, Prozesse und verschiedenster zusammenspielender Systeme und Akteure: Dies gilt beispielsweise für *Data Governance*, da es sich hier insbesondere um die Entwicklung und Überwachung von Richtlinien für den unternehmensweit einheitlichen Umgang mit Daten handelt und das *Data Operations Management*, welches sich auf Bereiche und Kontexte bezieht, die für den Schulunterricht nicht relevant sind, da schulische Projekte bzw. der Umgang mit Daten im Unterricht im Allgemeinen solche Tätigkeiten kaum als notwendig erscheinen lässt. Auch das *Reference & Master Data Management* betont mit der Replikation von Daten über verschiedenste Systeme und die systemübergreifende Nutzung von Daten eher Aspekte, wie sie in der Schule kaum direkt thematisierbar sind. Entsprechend sind solche Bereiche in der dargestellten Ausprägung kaum für die Schule und damit auch nicht für das zu erstellende Kategoriensystem relevant. Obwohl manche Funktionen damit auf den ersten Blick keine Verbindung zum Informatikunterricht zu haben scheinen, enthalten sie aber bei offenerer Betrachtung auch relevante Grundlagen und Ideen, die auch im Schulunterricht thematisiert werden könnten: Ein zentraler Aspekt des *Data Architecture Management* ist beispielsweise die Strukturierung von Daten, bei der erwartet werden kann, dass sie in verschiedenen Curricula einen zentralen Aspekt des Umgangs mit Daten darstellt. Das *Reference & Meta Data Management* kann beispielsweise auf Datensynchronisation bezogen werden, gleichzeitig stellen Metadaten an sich ein potenziell wichtiges Thema dar. Auch *Data Warehousing & Business Intelligence Management* ist zwar ein relativ spezielles Thema, das an sich wenig relevant scheint. Dabei handelt es sich jedoch insbesondere um zwei typische Beispiele der Datennutzung zur Entscheidungsfindung im Unternehmenskontext. Dieser Aspekt kann daher entsprechend als *Data Usage* verallgemeinert und somit potenziell doch im Schulkontext betrachtet werden. Bei der Ableitung des Kategoriensystems aus dem DAMA-DMBoK wurden daher verschiedene Anpassungen vorgenommen, die insbesondere aufgrund der anderen Perspektive auf Datenmanagement, die dieser Arbeit zugrunde gelegt wird, nötig werden.

[26] Die Darstellung des Kategoriensystems erfolgt auf Englisch, um potenzielle Begriffsverzerrungen durch die Übersetzung der Fachbegriffe des DAMA-DMBoK zu vermeiden.

Basierend auf diesen Überlegungen wurde daher ein Kategoriensystem entwickelt, das neben den vorgestellten Funktionen als Hauptkategorien auch diverse zusätzliche aus dem DAMA-DMBoK abgeleitete Unterkategorien beinhaltet. Damit entstand das folgende Ausgangs-Kategoriensystem[26]:

1. data development
    a) data modeling
    b) implementation
        i. database management system
        ii. query language
2. document & content management
    a) acquisition / retrieval
    b) storage
    c) backup & recovery
    d) content management
    e) retention
    f) purging
3. data security management
    a) data security
    b) data privacy
    c) access control
    d) encryption
4. meta data management
5. data quality management
    a) integrity
        i. consistency
        ii. redundancy
6. data usage / data analysis

Festlegung von Kodiereinheit und Auswahlkriterien

Um eine detaillierte Kodierung der Dokumente zu ermöglichen, wurde entschieden, die Länge der Kodiereinheiten möglichst minimal zu halten. Daher wurde versucht, nur einzelne Begriffe zu kodieren. Da dies kaum an allen Stellen sinnvoll möglich ist, wurde, soweit nötig, auch die Kodierung längerer Textstücke zugelassen. Es wurde jedoch darauf geachtet, dass einem Textstück alle in diesem vorkommenden bzw. mit diesem in Beziehung stehenden Themen/Begriffe zugeordnet werden, weswegen Mehrfachzuordnungen eines Textstücks zu mehreren Kategorien explizit zugelassen sind. Außerdem wird der Kontext berücksichtigt, in dem ein Begriff im jeweiligen Dokument genannt wurde: Dadurch können unter anderem Begriffe, die beispielsweise nur zur Abgrenzung in einem Dokument genannt werden, erkannt und von der Analyse ausgeschlossen werden, um zu vermeiden, dass die Ergebnisse durch solche Begriffsnennungen verzerrt werden. Entsprechend wird als Analysekriterium festgelegt, dass alle kodierten Textpassagen einen klaren Bezug zum Fachgebiet Datenmanagement aufweisen und – aufgrund ihrer Bedeutung im Dokument – auf eine Berücksichtigung des jeweiligen Themas als Unterrichtsinhalt schließen lassen.

Durchführung

Nach Festlegung des initialen Kategoriensystems folgte die eigentliche qualitative Inhaltsanalyse, die mit Unterstützung durch die Analysesoftware MaxQDA [27] durchgeführt wurde. Dazu wurden die ausgewählten Dokumente in zufälliger Reihenfolge durchsucht und Textpassagen, die den festgelegten Auswahlkriterien genügen, kodiert und, sofern möglich, in das zuvor entwickelte Kategoriensystem eingeordnet. Dabei wurde versucht, Begriffe auf einer möglichst tiefen bzw. detaillierten Ebene des Systems zu kodieren. Eine Zuordnung zur obersten Ebene wurde soweit möglich vermieden, da der Detailgrad dieser Ebene im Allgemeinen eher gering ist. Falls ein Begriff in einem der Dokumente genannt wurde, der zwar den Auswahlkriterien entsprach, aber der nicht (bzw. nur auf oberster Ebene) in das Kategoriensystem einordenbar war, wurde eine entsprechende induktive Ergänzung desselben vorgenommen. In allen vorgekommenen Fällen handelte es sich dabei um eine weitere Detaillierung des Systems. Auf oberster Ebene mussten indes keine Ergänzungen vorgenommen werden, sodass das System nicht in der Breite erweitert werden musste. Es ergaben sich daher folgende induktive Ergänzungen:

- Ausdifferenzierung von *1.1 data modeling*: Es wurden die drei Unterkategorien *1.1.1 non-relational model*, *1.1.2 object-oriented model* und *1.1.3 relational model* ergänzt.
- Ausdifferenzierung von *1.2.1 database management system*: Es wurden die beiden Unterkategorien *1.2.1.1 non-relational* und *1.2.1.2 relational* ergänzt.
- Ergänzung von *5. data quality management* um die Unterkategorie *5.2 data accuracy, reliability & completeness*.
- Ausdifferenzierung von *6. data usage/data analysis*: Der Aspekt *data analysis* wurde in eine eigene Unterkategorie *6.1* ausgelagert und die weiteren Unterkategorien *6.2 data interpretation*, *6.3 data sharing*, *6.4 large amounts of data*, *6.5 legal, social and ethical aspects* ergänzt.

Auswertung

Nach der Analyse der ausgewählten Dokumente konnte festgestellt werden, dass das entstandene induktiv ergänzte Kategoriensystem – wie auch das Ausgangssystem – Kategorien unterschiedlichsten Detailgrads enthält. Deswegen und aufgrund des unterschiedlichen Umfangs und Detailgrads der zugrundeliegenden Dokumente, sind die gewonnenen Ergebnisse ohne weitere Verarbeitung und zusätzliche Informationen kaum interpretierbar: Eine quantitative Betrachtung der Vorkommnisse der Themen in den Dokumenten würde womöglich fehlerhafte Schlüsse auf den Umfang bzw. die Relevanz der Themen im Unterricht nahelegen. Daher wurden die Anzahlen der Kodierungen zu einem Wahrheitswert pro Kategorie und Dokument zusammengefasst. Dies ist den Analysezielen dienlich, da es nicht darum geht, zu ermitteln, welche Bedeutung ein Thema in den Dokumenten hat

[27] `http://www.maxqda.de`

bzw. wie oft es darin vorkommt, sondern darum, aus einer distanzierten Perspektive auf den Informatikunterricht zu ermitteln, welche Themen üblicherweise betrachtet werden.

In Tabelle 4.2 werden die Analyseergebnisse dargestellt, indem für jedes Thema und jedes analysierte Dokument angegeben wird, ob ein Thema darin vorkommt. Zusätzlich wird für jedes Thema die Prozentzahl der Dokumente, die dieses abdecken, und für jedes Dokument die Prozentzahl an Themen, die dieses abdeckt, angegeben. Die prozentuale Häufigkeit der Themen in den verschiedenen Dokumenten wird außerdem, da diese eine wesentliche Grundlage für die Beantwortung der beiden Fragestellungen darstellt, in Abbildung 4.2 visualisiert.

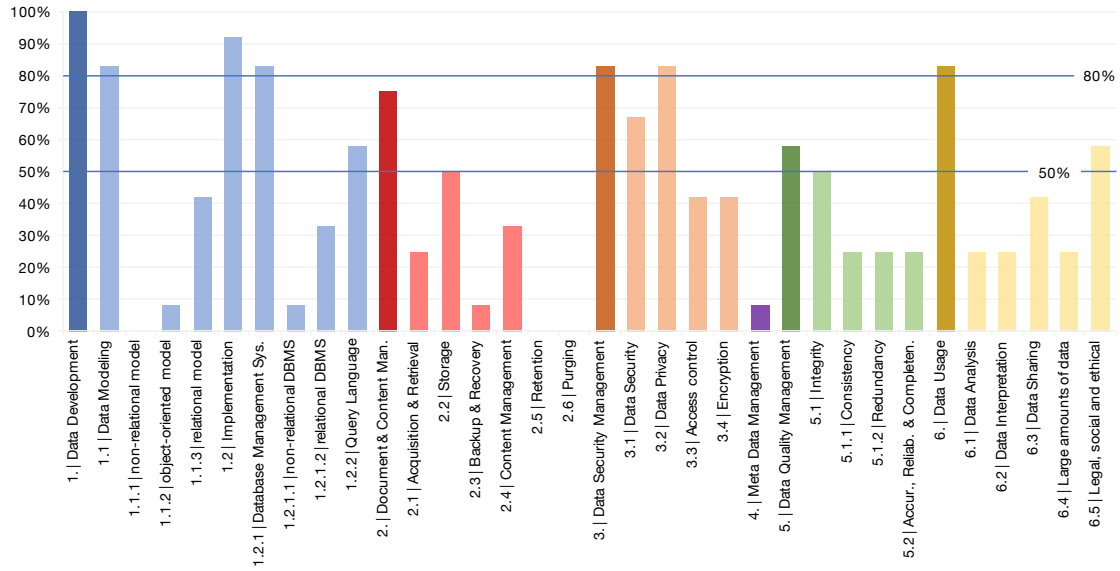


Abbildung 4.2: Verbreitung von Datenmanagementthemen in Bildungsstandards und Curricula.

Kategorie	EPA 2004	GI 2008	K12 2011	BY 2009	HE 2010	HH 2011	NRW 2013	RLP 2011	AT 2012	CA 2007	CAS 2012	IS 1999	Prozentuale Abdeckung
1. Data Development	×	×	×	×	×	×	×	×	×	×	×	×	100 %
1.1 Data Modeling	×	×	×	×	×	×	×	×			×	×	83 %
1.1.1 non-relational model													0 %
1.1.2 object-oriented model				×									8 %
1.1.3 relational model	×	×						×			×	×	42 %
1.2 Implementation	×	×		×	×	×	×	×	×	×	×	×	92 %
1.2.1 Database Management System	×	×		×	×	×	×	×	×	×		×	83 %
1.2.1.1 non-relational DBMS					×								8 %
1.2.1.2 relational DBMS				×				×		×		×	33 %
1.2.2 Query Language				×	×	×	×	×		×		×	58 %
2. Document & Content Management		×	×	×	×	×	×		×	×	×		75 %
2.1 Acquisition & Retrieval			×						×	×			25 %
2.2 Storage		×		×	×		×			×	×		50 %
2.3 Backup & Recovery										×			8,3 %
2.4 Content Management		×		×		×	×						33 %
2.5 Retention													0 %
2.6 Purging													0 %
3. Data Security Management	×	×	×	×	×	×	×	×		×	×		83 %
3.1 Data Security	×		×	×	×		×	×		×	×		68 %
3.2 Data Privacy	×	×	×	×	×	×	×	×		×	×		83 %
3.3 Access control			×	×		×	×	×					42 %
3.4 Encryption		×	×			×	×	×					42 %
4. Meta Data Management					×								8 %
5. Data Quality Management		×		×	×		×	×			×	×	58 %
5.1 Integrity				×	×		×	×			×	×	50 %
5.1.1 Consistency				×				×			×		25 %
5.1.2 Redundancy				×				×				×	25 %
5.2 Data Accuracy, Reliability & Completeness		×		×							×		25 %
6. Data Usage	×	×	×	×		×	×	×	×	×	×		83 %
6.1 Data Analysis		×	×							×			25 %
6.2 Data Interpretation	×	×					×						25 %
6.3 Data Sharing		×		×		×	×			×			42 %
6.4 Large amounts of data			×						×		×		25 %
6.5 Legal, social and ethical aspects		×	×			×	×	×	×	×			58 %
Prozentuale Abdeckung	30 %	53 %	38 %	62 %	41 %	47 %	56 %	53 %	24 %	47 %	53 %	30 %	

Tabelle 4.2: Überblick über die Analyseergebnisse: In den Spalten wird das Dokument, in den Zeilen die Kategorien angegeben. Kategorien die in mindestens 80 % der analysierten Dokumente enthalten sind, sind durch graue Hinterlegung hervorgehoben.

4.3.4 Interpretation

Die Ergebnisse der Untersuchung können hinsichtlich der beiden zugrundeliegenden Fragestellungen bzw. der zweiten Forschungsfrage dieser Arbeit wie folgt zusammengefasst und interpretiert werden:

Datenmanagement im heutigen Informatikunterricht

Während des Analyseprozesses hat sich deutlich gezeigt, dass das deduktiv ermittelte Kategoriensystem für den Analysezweck gut geeignet war, da nur wenige der in den Dokumenten enthaltenen Themen des Datenmanagements induktiv ergänzt werden mussten. Dass es sich dabei insbesondere um solche Kategorien handelt, die weitere Details ergänzen, bestätigt diesen Eindruck. Der *Data Management Body of Knowledge* deckt daher die für die Schule relevanten Themenbereiche im Wesentlichen ausreichend ab. In dieser professionellen Beschreibung von Datenmanagement wird allenfalls an einigen Stellen auf Details verzichtet, die bei der speziellen Adaption des Fachgebiets für den Unterricht herausgearbeitet werden müssen.

Die Ergebnisse zeigen, dass der Informatikunterricht sich im Umfeld des Datenmanagements heute üblicherweise auf einen eher kleinen Ausschnitt des Fachgebiets konzentriert: Während alle Kategorien der obersten Ebene in mindestens 58 % und vier dieser sechs Kategorien sogar in mehr als 80 % der analysierten Dokumente betrachtet werden, ist ab den Kategorien ab der zweiten Ebene eine wesentlich geringere Abdeckung erkennbar – nur drei der 19 Kategorien dieser Ebene sind in mindestens 80 % der Dokumente repräsentiert (vgl. Abbildung 4.2). Dabei handelt es sich erwartungsgemäß um die traditionellen Unterrichtsthemen im Datenbankunterricht, insbesondere *Datenmodellierung* und *Datenbankmanagementsysteme*, aber auch um den Themenbereich *Data Privacy*, der heute immer größeren Stellenwert im gesellschaftlichen Diskurs erlangt. Neben diesen Themen werden jedoch auch in mindestens 50 % der analysierten Dokumente *Anfragesprachen*, *Datenspeicherung* im Allgemeinen, *Datensicherheit*, *Integrität* von Daten sowie *rechtliche, soziale und ethische Aspekte* im Zusammenhang mit Daten und Datenmanagement erwähnt. Neben diesen Themen werden auch weitere Bereiche des Datenmanagements zwar im Unterricht am Rande angeschnitten, eine detailliertere Betrachtung findet jedoch nur an wenigen Stellen statt und bleibt relativ oberflächlich.

Die Analyse zeigt außerdem einen hohen Konsens zwischen den analysierten Dokumenten: Obwohl die Empfehlungen für Bildungsstandards in Informatik für die Sekundarstufe I in Deutschland nicht verbindlich eingeführt sind, weisen die betrachteten deutschen Curricula (BY, HH, HE, RLP, NRW) eine hohe Übereinstimmung mit diesen auf. Der hohe Konsens ist jedoch nicht auf eine nationale Betrachtung beschränkt, sondern kann auch international erkannt werden. Insbesondere mit dem Ontario State Curriculum (CA), dem Computing at School Curriculum (CAS) sowie dem österreichischen AHS-Curriculum (AT) weisen die deutschsprachigen Dokumente im Allgemeinen eine hohe Übereinstimmung auf.

Charakterisierung der Lücke zwischen Informatikunterricht und fachlicher Sichtweise

Aus einer anderen Perspektive betrachtet können die Ergebnisse die Lücke zwischen dem Fachgebiet Datenmanagement und dem in den betrachteten Dokumenten festgehaltenen Stand des Informatikunterrichts charakterisieren. Diese zeigt sich insbesondere durch die Themen, die kaum im Unterricht betrachtet werden: Nur 41 % der betrachteten Themen sind in mindestens 50 % der Dokumente repräsentiert. Selbst bei Absenkung des Schwellwerts für diese Betrachtung auf 30 %, werden nur knapp 60 % der Themen in den Dokumenten abgedeckt.

Diese Lücke zeigt sich beispielsweise am Vergleich der Bedeutung von strukturierten und unstrukturierten Daten in den analysierten Dokumenten: Während den strukturierten Daten (Kategorien 1.1 und 1.2) im Unterricht eine hohe Relevanz zukommt, werden weniger strukturierte Daten (Kategorien 2.1–2.6) kaum betrachtet, obwohl diese im Alltag der Schülerinnen und Schüler sicherlich eine wesentlich größere Rolle spielen, da dazu unter anderem unstrukturierte Textdokumente, Bilder, Videos usw. zählen. Dieser Unterschied zeigt sich beispielsweise im Curriculum der kanadischen Provinz Ontario (2007): Dieses erwähnt zwar die Kompetenz, dass Schülerinnen und Schüler Daten aus externen Quellen (wie beispielsweise aus sequentiellen Dateien, Datenbanken, XML-Dateien oder relationalen Datenbanken unter Nutzung SQL) lesen und in solche schreiben können sollen. Gleichzeitig ist jedoch kein Hinweis darauf zu finden, dass das dabei erworbene Wissen auch auf unstrukturierte oder weniger strukturierte Daten übertragen werden soll, wie sie z. B. in einem Cloud-Speicher oder einem Content-Management-System gespeichert werden. Ähnliches zeigt sich auch bei weiteren Themen: Beispielsweise ist zwar *data development* (das u. a. „Datenbanken“, enthält) in allen Materialien repräsentiert, *data security management* (u. a. mit Aspekten der Privatsphäre und des Datenschutzes) konnte hingegen nur in zehn von zwölf Dokumenten, die Datensicherheit selbst sogar nur noch in acht von zwölf Dokumenten erkannt werden. Die Verwaltung weniger strukturierter Daten wird nur in sieben Dokumenten genannt, *data quality management* sogar nur in drei. Auch das heute im Umgang mit Daten sehr zentrale Thema *Metadaten* wird nur in einem Curriculum – und dort nur am Rande – thematisiert: Im hessischen Lehrplan (HE) geht es dabei insbesondere um die Verwendung von Metatags auf Webseiten und die damit einhergehenden Vorteile für die Indizierung durch Suchmaschinen, dass dabei erworbene Wissen wird jedoch nicht verallgemeinert und auf andere Anwendungsfälle von Metadaten übertragen. In nahezu allen Curricula werden jedoch verschiedenste Arten der Datennutzung explizit erwähnt, es kann aber davon ausgegangen werden, dass solche Beispiele auch in den anderen Fällen zumindest oberflächlich einfließen, da die Nutzung der zentrale Zweck einer Verwaltung von Daten ist.

Es zeigt sich also, wie bereits vermutet, ein klarer Unterschied zwischen dem Forschungsstand im Datenmanagement und dem Stand des Informatikunterrichts. Insbesondere ist dabei erkennbar, dass modernere Themen bisher, zumindest laut den entsprechenden Curricula und Bildungsstandards, kaum unterrichtlich aufgegriffen werden, während eher tradierte Aspekte klar vertreten sind.

Zusammenhang zwischen Erscheinungszeitpunkt und abgedeckten Themen

Als dritter Teilaspekt dieser Untersuchung wurde überprüft, ob ein Zusammenhang zwischen dem Entstehungszeitpunkt eines Dokuments und dem Umfang der darin abgedeckten Datenmanagementthemen erkennbar ist. Ein erkennbarer Zusammenhang würde es zulassen, auf eine derzeit stattfindende Entwicklung zu schließen, insbesondere wenn neuere Dokumente die Themen und Entwicklungen der letzten Jahre eher berücksichtigen. Auch diese Hypothese kann anhand der Analyseergebnisse widerlegt werden: Eine Darstellung der prozentualen Abdeckung der Datenmanagementthemen in einem Dokument über dessen Erscheinungsjahr (Abbildung 4.3) zeigt, dass ein solcher Zusammenhang nur eingeschränkt bestehen kann. Es ist zwar eindeutig erkennbar, dass die beiden ältesten Dokumente nur einen sehr geringen Anteil der Datenmanagementthemen berücksichtigen, während die meisten modernen Dokumente zumindest ein höheres Gewicht auf diese legen. Gleichzeitig ist jedoch innerhalb dieser neueren Dokumente kein klarer Zusammenhang zwischen Erscheinungsjahr und prozentualer Abdeckung erkennbar: Beispielsweise hat das Computing-At-School-Curriculum (CAS) von 2012 dieselbe Abdeckung wie der Lehrplan von Rheinland-Pfalz (RLP) von 2011 und die Empfehlungen für Bildungsstandards für die Sekundarstufe I der Gesellschaft für Informatik (GI) von 2008. Auch der Lehrplan des Bundeslands Nordrhein-Westfalen von 2013 deckt nur marginal mehr ab als diese beiden Dokumente. Es kann daher auf Basis dieser Untersuchung keine derzeitig stattfindende Entwicklung in diesem Bereich identifiziert werden.

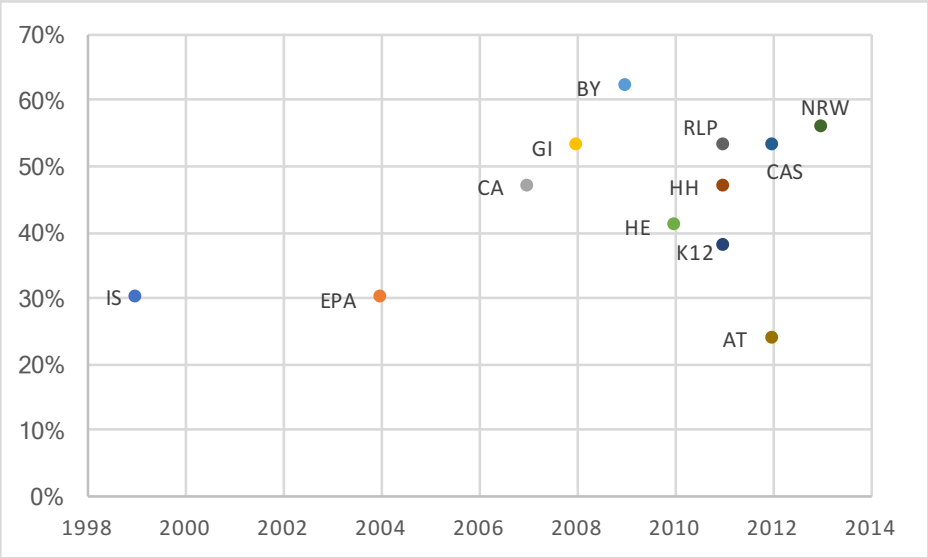


Abbildung 4.3: Auftragung der prozentualen Abdeckung von Datenmanagementthemen innerhalb eines Dokuments über dessen Erscheinungsjahr.

4.3.5 Zusammenfassung und Fazit

Hinsichtlich der Abdeckung von Datenmanagementthemen im aktuellen Unterricht kann, zumindest nach den analysierten Bildungsstandards und Curricula, klar gesagt werden, dass der Informatikunterricht sich auch heute weiterhin insbesondere auf die bereits anfangs der 1990er Jahre etablierten Themen *Datenmodellierung* und *(relationale) Datenbanken* fokussiert[28]. Dies deckt sich mit der Darstellung aus Abschnitt 4.3, die auf einer explorativen Charakterisierung der Inhalte von Unterrichtskonzepten und Schulbüchern beruhte. In den analysierten Dokumenten werden beispielsweise Lernziele betont wie die *„Realisierung von Objekten, Klassen und Beziehungen in einem relationalen Datenbanksystem" Staatsinstitut für Schulqualität und Bildungsforschung (2009)*. Modernere Themen, wie *Data Mining*, *nicht-relationale Datenbanken* oder *Big Data* im Allgemeinen werden bisher weder ausführlich im Unterricht thematisiert, noch wurden sie im Rahmen informatikdidaktischer Forschung detaillierter untersucht. Trotz der interessanten Ergebnisse, die auch die Hypothese stützen, dass Datenmanagement bisher allenfalls als Randthema im Unterricht vertreten ist, müssen jedoch auch die Grenzen der Analyse betrachtet werden: Insbesondere können keine direkten Rückschlüsse auf die Realität des Informatikunterrichts gezogen werden, da es sich um eine Analyse reiner Richtliniendokumente handelt. Diese lassen nicht zwingend auf die reale Umsetzung im Unterricht schließen, da bei dieser vielfältige Schwerpunkte gesetzt werden können, die es auch ermöglichen im Rahmen des aktuellen Informatikunterrichts bereits weitere Themen des Datenmanagements zu thematisieren – was sicherlich zum Teil auch geschieht. Es kann jedoch erfahrungsgemäß angenommen werden, dass ein Großteil des Unterrichts sich eher relativ eng an den vorliegenden Vorgaben orientiert, sodass die Analyse einen guten Einblick in große Teile des Informatikunterrichts gibt. Durch die Betrachtung dieser Richtliniendokumente kann außerdem nicht auf den Umfang eines Themas im Unterricht geschlossen werden: Beispielsweise könnte ein Unterricht zwar die beiden Aspekte Datenbanken und Datensicherheit thematisieren, aber einen deutlichen Schwerpunkt auf den Sicherheitsaspekt legen und dabei weitere Themen des Datenmanagements passenderweise mindestens mit anreißen, ohne dass es aus den Bildungsstandards und Curricula hervorgehen würde.

[28] Auch gegen Ende dieses Promotionsprojekts sind hier allenfalls geringe Veränderungen erkennbar, die jedoch diese Ausgangslage für die Arbeit nicht relevant beeinflussen.

5 Daten und Datenmanagement in Gesellschaft, Alltag und Beruf

Nachdem in den vorherigen Kapiteln die Bedeutung von Datenmanagement aus fachlicher und professioneller Sicht herausgearbeitet und der aktuelle Forschungsstand in der Informatikdidaktik charakterisiert wurde, wird im Folgenden der Bezug des Themenfeldes zur Alltagswelt hergestellt und dessen Bedeutung in der heutigen *digitalen Gesellschaft*[29] herausgestellt. Dazu werden einerseits die Anforderungen der Gesellschaft an den Umgang mit Informatiksystemen zusammengefasst, andererseits auch relevante Alltagskontexte aufgezeigt, in denen Datenmanagement heute eine immer größere Bedeutung einnimmt. Aus diesen beiden Bereichen werden daraufhin exemplarische Phänomene herausgegriffen, denen heute jeder im tagtäglichen Umgang mit Informatiksystemen begegnen kann. Als Abschluss dieses Kapitels wird die in den letzten Jahren neu entstandene und viel beachtete Profession des *Data Scientist* charakterisiert, die eine weitere Perspektive auf die vielfältige Verarbeitung und Nutzung von Daten eröffnet und die berufliche Bedeutung des Themengebiets verdeutlicht.

5.1 Anforderungen der digitalen Gesellschaft

> *Die Digitalisierung birgt riesige Chancen für eine gerechtere und fortschrittliche Gesellschaft. Um diese zu nutzen, braucht es mehr als den Ausbau technischer Infrastruktur. Alle müssen aktiv teilhaben können – und dafür die richtigen Rahmenbedingungen und notwendigen digitalen Kompetenzen haben. (Bundesministerium für Familie, Senioren, Frauen und Jugend, 2017)*

Im Rahmen der digitalen Gesellschaft ergeben sich durch die Speicherung, Analyse und allgemein Nutzung von Daten vielfältige neue Anforderungen, Herausforderungen und Chancen, die verschiedene Bereiche des Zusammenlebens in der Gesellschaft betreffen und die daher – auch zur Schaffung eines kritischen Weltbildes – im Rahmen von allgemeinbildendem Unterricht thematisiert werden sollten. Der Begriff *Digitalisierung* wird dabei oft nicht mehr im Sinne der informatischen Definition als die Abbildung analoger Informationen in durch Ziffern beschriebene maschinell verarbeitbare Daten verstanden, sondern weiter gefasst, sodass Digitalisierung nach diesem Verständnis eine *„Veränderung von Prozessen und Objekten durch den zunehmenden Einsatz von digitaler Technik" (Wissenschaftliche Dienste des Deutschen Bundestags, 2017)* darstellt. Die damit einhergehenden Veränderungen werden durch eine Vielzahl von Akteuren charakterisiert: Der Arbeitskreis der Technologietransferstellen niedersächsischer Hochschulen beschreibt beispielsweise in der Publikation

[29] Der Begriff *digitale Gesellschaft* wird in dieser Arbeit im Sinne des aktuell vorherrschenden Diskurses zu diesem Thema so verstanden, dass er die zunehmende Digitalisierung aller Lebensbereiche sowie die gesellschaftlichen Auswirkungen dieser Entwicklung beinhaltet.

„Technologie-Informationen“ die Veränderungen, die eine digitale Dorfgemeinschaft, Industrie 4.0, die zunehmende Verbreitung von Robotern und Ähnliches mit sich bringen *(Arbeitskreis der Technologietransferstellen niedersächsischer Hochschulen, 2017)*: All diese Veränderungen beruhen unter anderem auf der großen Menge an Daten, die heute über Alles und Jeden zur Verfügung steht. Im gesellschaftlichen Kontext steht dabei oft insbesondere eine bisher nicht möglich gewesene Individualisierung vielfältiger Aspekte des täglichen Lebens im Zentrum. Die FDP beschreibt in einem Beschlussdokument sechs Chancen der digitalen Gesellschaft: *individuelle Bildung, bessere Arbeit, digitale Autonomie, Chancen für Wirtschaft und Mobilität, besseres Gesundheitssystem* und ein *unkomplizierter Staat (FDP, 2016)*. Ähnliche und weitere Bereiche, *Zukunft, Lernen, Arbeit, Transparenz* und *Ethik*, werden beispielsweise durch *Arnold und Köhler (2018)* aufgegriffen. Auch in diesen beiden Beispielen ist eine klare Individualisierung bzw. Fokussierung auf den Einzelnen erkennbar, die durch die vielfältigen Daten, die heute zur Verfügung stehen, erst ermöglicht wird, aber auch, mit der *digitalen Autonomie* und der *Ethik*, Ansätze für die Diskussion von Gefahren, die mit den vielfältigen neuen Möglichkeiten einhergehen.

Einer der Bereiche, die regelmäßig im Kontext der digitalen Gesellschaft aufgegriffen werden, ist die Bildung: Beispielsweise setzt sich die Kultusministerkonferenz in ihrem Strategiedokument „Bildung in der digitalen Welt“ unter anderem das Ziel, dass die Länder Kompetenzen in ihre Lehr-/Bildungs-/Rahmenlehrpläne einbeziehen, die *„für eine aktive, selbstbestimmte Teilhabe in einer digitalen Welt erforderlich sind“ (Kultusministerkonferenz, 2016)*. Dieses Strategiedokument wurde durch den *Fachbereich Informatik und Ausbildung/Didaktik der Informatik* der *Gesellschaft für Informatik* in einer Stellungnahme aufgegriffen und weiter ausgearbeitet. Dabei ist ein Modell entstanden, das *„verschiedene Sichtweisen, die ein Gesamtkonzept der digitalen Bildung ermöglichen“ (Brinda, 2016)* zusammenfasst. Dieses in Abbildung 5.1 dargestellte Modell und dessen Begründung in der Stellungnahme betonen, neben anderen, die technologische Perspektive mit der Fragestellung *„Wie und warum funktioniert das?“ (Brinda, 2016)* und dem Ziel der *„Gestaltung aktiver digitaler Medien und Technologie zur Problemlösung“ (Brinda, 2016)*. Somit sind für ein Weltverständnis in der heutigen digitalen Gesellschaft und für eine aktive Mitgestaltung dieser Welt, fachliche Kompetenzen aus der Informatik unerlässlich und stellen eines der Fundamente der digitalen Gesellschaft dar. Zu diesen Kompetenzen gehören auch solche aus dem näheren Umfeld des Datenmanagements, da viele der Entwicklungen und Technologien, die die heutige digitale Gesellschaft prägen, unter anderem auf Aspekten dieses Fachgebiets basieren.

Auch im Rahmen des gesellschaftlichen Diskurses werden immer mehr Themen der Informatik aufgegriffen und umfassend thematisiert. Dies trifft aus dem Datenmanagement insbesondere auf das Schlagwort *Big Data* zu, das zwar aus informatischer Perspektive durch die drei zuvor bereits beschriebenen Eigenschaften *volume*, *velocity* und *variety* charakterisiert wird, im umgangssprachlichen Verständnis aber eine zunehmend weitere Bedeutung hat: Dabei wird unter *Big Data* heute im Zweifel alles verstanden, das mit der maschinellen Speicherung und Verarbeitung (mehr oder weniger großer) Datenmengen in Zusammenhang steht. Dies führt zu einer hohen Verbreitung dieses Begriffs und insbesondere zur Diskussion dieses eigentlich eher fachlich gearteten Themas selbst in Tageszeitungen, deren

Abbildung 1: Haus der digitalen Bildung[4]

Abbildung 5.1: Haus der digitalen Bildung *(Brinda, 2016)*.

eigentliche Zielgruppe nicht unbedingt tiefgreifendes informatisches Vorwissen aufweist. Auch im wissenschaftlichen Diskurs zu Big Data ist durchaus erkennbar, dass dieses Thema auch als gesellschaftlich relevantes Phänomen wahrgenommen wird: So untersuchen beispielsweise *Pentzold und Fischer (2017)*, welche Aspekte im Diskurs zur Sammlung und Nutzung von Daten für Journalisten und die breite Öffentlichkeit bedeutsam sind. Im Kontext der *Handygate-Affaire*[30] ermittelten die Autoren vier verschiedene Ausprägungen des Diskurses zu diesem Thema: Die *Rechtfertigung* der Aktivitäten, die *Kritik* an diesen, die *Resignation* vor ihnen und die *Selbstverantwortlichkeit* für diese. Diese vier Ausprägungen zeigen, dass gerade Negativbeispiele vielfältig und vielschichtig von der Öffentlichkeit hinterfragt werden.

Damit wird deutlich, dass für ein kritisches Weltverständnis heute vielfältiges Wissen notwendig ist, das aufgrund der Bedeutung der Informatik in der heutigen digitalen Gesellschaft auch aus dieser und insbesondere dem Datenmanagement stammt: Da die Informatik *„die Wissenschaft von der systematischen Verarbeitung von Informationen" (Rechenberg und Pomberger, 2002)* darstellt, sind sowohl *Information* als auch *Daten*, die die maschinenverarbeitbare Repräsentation von Informationen darstellen, zentrale Begriffe nicht nur der Informatik, sondern auch der heutigen Gesellschaft im Allgemeinen. Durch die zunehmende Digitalisierung und den erkennbaren Versuch, alles Digitalisierbare auch wirklich zu digitalisieren (und die dabei entstehenden Daten im Folgenden vielfältig und nutzbringend einzusetzen), stellt Datenmanagement eine der zentralen Herausforderungen in allen

[30] Die *Handygate-Affaire* bezeichnet eine, vom Landeskriminalamt Sachsen im Rahmen eines Ermittlungsverfahrens durchgeführte, massenweise Auswertung von Handy-Verbindungsdaten. Dabei wurden persönliche Daten von mehr als 40.000 Personen erfasst, die sich während einer Demonstration am 18. und 19. Februar 2011 innerhalb bestimmter Mobilfunkzellen in Dresden aufhielten *(vgl. Lischka, 2011)*.

Bereichen der heutigen Gesellschaft dar. Dies trifft nicht nur auf die kommerzielle Verarbeitung von Daten durch Unternehmen zu, sondern führt auch zu einer immer stärker werdenden Bedeutung im täglichen Leben, in dem heute immer häufiger Phänomene der Informatik im Allgemeinen und des Datenmanagements im Speziellen auftreten und deren Grundlagen – für ein umfassendes Weltverständnis – auch verstanden werden müssen.

5.2 Alltagskontexte und Phänomene des Datenmanagements

Neben ihrer gesellschaftlichen Rolle nehmen Daten heute auch eine Schlüsselrolle im persönlichen Leben ein: Jeder erzeugt und verwaltet täglich riesige Datenmengen, unter anderem bei der Speicherung von Dokumenten oder Musik, der Nutzung sozialer Medien oder auch in weniger offensichtlicher Form, z. B. bei der Nutzung von elektronischen Fahrkarten im öffentlichen Personennahverkehr. Der Umgang mit Daten hat dabei viele Facetten: Daten werden beabsichtigt oder unbeabsichtigt erzeugt, können lokal oder in der Cloud gespeichert werden, sind von verschiedenster Art und haben unterschiedlichste Strukturen. Da Daten nahezu überall erfasst werden, ermöglichen Datenanalysen die Rekonstruktion großer Teile des Privatlebens, oft mit hoher statistischer Signifikanz. Zum Beispiel reicht es aus, die Daten, die durch elektronische Fahrkarten generiert werden, mit den Daten, die intelligente Stromzähler („Smart Meter“) erfassen, zusammenzuführen, um Arbeitszeiten und Einkaufsgewohnheiten einer Person herauszufinden (Beispiele für diese Möglichkeiten finden sich bei *Beckel et al. (2014)*). Während sich dieses Beispiel mit Daten beschäftigt, die von Dritten über eine Person erfasst werden, werden heute gleichzeitig immer umfangreichere Datensätze durch uns selbst erfasst: Mit dem Ziel die eigene Lebensqualität zu erhöhen oder eine Verbesserung der Gesundheit herbeizuführen, wird das „Quantified Self“ bzw. „Life Logging“ mit der zunehmenden Nutzung von beispielsweise Smart Watches oder Fitness-Armbändern immer alltäglicher, sodass heute große Teile des Alltags protokolliert werden. Durch den Einzug immer vielfältigerer, ständig mithörender und uns kontinuierlich „beobachtender“ intelligenter persönlicher Assistenten (im Sinne von Amazons Alexa, Apples Siri oder Googles Assistant) wird die Bedeutung solcher Daten und damit einhergehend auch die Anzahl interessanter Kontexte für den Informatikunterricht auch in den nächsten Jahren sicherlich weiterhin deutlich zunehmen.

Aufgrund des deutlichen und weiterhin zunehmenden Einflusses von Daten wird der verantwortungsbewusste Umgang mit eigenen und fremden Daten ein immer integralerer Bestandteil des Lebens. Insbesondere Speichern, Bearbeiten, Löschen und Nutzen von Daten sind bereits heute alltägliche Aufgaben, die trotzdem verschiedene Herausforderungen bergen: Das Speichern beinhaltet nicht nur die Wahl eines geeigneten Speichermediums, sondern beispielsweise auch Entscheidungen, wie die Daten strukturiert und organisiert werden, ob Backups nötig sind (und wie diese angefertigt werden sollen) oder ob die Synchronisierung von Daten zwischen einer Vielzahl von Geräten (und möglicherweise Benutzern) angestrebt und wie diese realisiert wird. Dabei gewinnt auch der Schutz eigener und fremder Daten vor Manipulation, Verlust oder missbräuchlicher Nutzung wesentlich

an Bedeutung, genauso wie Methoden zur Sicherstellung der Authentizität von Daten. Gleichzeitig findet eine umfassende Erfassung von Daten und deren Auswertung in vielen Kontexten bereits heute statt, beispielsweise werden durch Smartphones bzw. Smartphone-Apps kontinuierlich Standort- und Bewegungsdaten erhoben, das Smart Home „weiß“, ob jemand zu Hause ist und vieles mehr. Zum Teil lassen diese Dienste Spielräume für die Einschränkung einer Datennutzung durch den Anbieter. Um hier fundierte Entscheidungen treffen und Dienste bewusst nutzen zu können, ist es notwendig, den Wert der eigenen Daten zu erkennen sowie ein Verständnis für die Funktionsweise und somit auch die Möglichkeiten und Grenzen moderner Datenerfassung und Datennutzung zu erwerben. Die obigen Kontexte dienen somit als Rahmen für den Unterricht, die es den Lernenden erlauben, das im Schulunterricht erworbene Wissen mit ihren Alltagserfahrungen zu verknüpfen.

Für eine weitergehende Orientierung an den Alltagserfahrungen der Schülerinnen und Schüler, ist es hilfreich, aus diesen Kontexten solche Phänomene im Unterricht herauszugreifen, denen die Lernenden bereits im Rahmen des allgegenwärtigen Umgangs mit Daten begegnet sind bzw. zumindest sein können (vgl. z. B. *Diethelm und Dörge (2011)*).

Unter einem informatischen Phänomen wird in dieser Arbeit, entsprechend der Definition nach *Diethelm und Dörge (2011)*, ein Ereignis verstanden, *„das durch automatisierte Informationsverarbeitung verursacht wird und im realen oder mentalen Handlungsumfeld der Schülerinnen und Schüler stattfindet“*. Phänomene verknüpfen daher den (Informatik-)Unterricht mit der Lebenswelt der Lernenden, sodass sie für einen motivierenden und schülerorientierten Unterricht, der zu einem Weltverständnis beiträgt, essentiell sind. Aus diesem Grund kommt den Phänomenen auch eine zentrale Bedeutung im Modell der Didaktischen Rekonstruktion für den Informatikunterricht zu *(Diethelm, Dörge et al., 2011)*. Eine vertiefte Betrachtung der Phänomene kann aufgrund deren Vielfalt im Rahmen dieser Arbeit nicht stattfinden. Stattdessen wird im Folgenden eine exemplarische Auswahl von Phänomenen aus verschiedenen Bereichen des alltäglichen Umgangs mit Daten charakterisiert, die so ausgewählt wurden, dass sie die Breite des betrachteten Fachgebiets widerspiegeln. Gleichzeitig wurde darauf geachtet, dass sie auch solche Fragestellungen eröffnen, zu deren Klärung grundlegende Kenntnisse aus dem Datenmanagement notwendig oder hilfreich sind, sodass sie Ausgangspunkte für einen motivierenden Datenmanagementunterricht darstellen. Ein Teil der vorgestellten Phänomene wird im vierten Teil der Arbeit im Rahmen der praktischen Unterrichtserprobung aufgegriffen.

Bereich *Datenspeicherung und -verwaltung*. Die Speicherung und Verwaltung von Daten sind heute im Alltag allgegenwärtig. Jeder speichert große Datenmengen und nutzt dabei verschiedene Möglichkeiten zur Datenspeicherung und -strukturierung, beispielsweise indem Textdokumente häufig direkt im Dateisystem gespeichert, Musik, Videos oder E-Mails aber üblicherweise in speziell zur Verwaltung solcher Dateitypen orientierten Datenspeichern verwaltet werden. Die dabei auftretenden Phänomene betreffen verschiedene Aspekte der Datenspeicherung und -verwaltung:

- Je nachdem wo eine Datei gespeichert ist, stehen verschiedene Möglichkeiten zur Verfügung, beispielsweise Versionierung.
- Trotz gleicher Dateigröße können manche Dateien auf einen Cloud-Datenspeicher schneller hochgeladen werden als andere.
- Die Speicherung derselben Fotos in einer Fotoverwaltungssoftware benötigt zum Teil weniger Speicherplatz als direkt im Dateisystem.
- Nicht alle Dateitypen können von einer Suchfunktion gleichermaßen gut und effizient durchsucht werden.

Bereich *Metadaten*. Ergänzend zu den eigentlichen Daten werden meist umfangreiche Zusatzinformationen in Form von Metadaten erzeugt und gespeichert. Es handelt sich dabei z. B. um Dateiattribute, wie den Zeitstempel der letzten Änderung oder den Dateinamen, aber auch um detailliertere Informationen, wie Änderungsprotokolle und Autoreninformationen, den Entstehungsort von Fotos oder einer Datei zugeordnete Stichworte. Die Masse und Ausprägung dieser Daten hängt stark von der genutzten Anwendung, dem Gerät und dem Dateiformat ab. Entsprechend ergeben sich auch in diesem Bereich verschiedene Phänomene, wie beispielsweise:

- Bei Fotos ist es möglich, diese nach Entstehungsort zu filtern, bei anderen Dateitypen, wie Dokumenten, nicht.
- Obwohl Informationen aus einem Dokument gelöscht wurden, können diese oft – auch durch andere Personen – wiederhergestellt werden.
- Je nach verwendetem Programm und der Art der Datei können Änderungen rückgängig gemacht werden.
- Auch wenn der Ersteller einer Datei nicht genannt wird bzw. anonymisiert wurde, ist dieser zum Teil einfach identifizierbar.

Bereich *Datensynchronisation*. Bei der Speicherung und Verwaltung von Daten besteht häufig die Schwierigkeit, Inkonsistenzen zu vermeiden, wenn eine redundante Datenspeicherung vonnöten ist. Dies ist insbesondere der Fall, wenn Daten nicht nur durch eine Person und auf einem Gerät genutzt werden, sondern über mehrere Geräte hinweg bzw. zwischen Personen synchronisiert werden. Je nach verwendetem Dateiformat und Werkzeug treten dabei unterschiedliche Phänomene auf:

- Nur bei manchen Dateien erkennt der Computer, wenn diese bereits durch andere bearbeitet werden und verweigert es, diese schreibend zu öffnen.
- Nach dem Bearbeiten derselben Datei auf mehreren Geräten bzw. durch mehrere Nutzer entstehen häufig Konflikte.

- Synchronisationskonflikte können nur in manchen Fällen automatisch aufgelöst werden.
- Trotz eines Backups von Daten in die Cloud können Daten versehentlich unwiderruflich gelöscht werden.

Bereich *Datenanalyse*. Daten werden heute überall erfasst und ausgewertet, um anhand dieser Entscheidungen zu treffen. Dabei sind zum Teil erstaunlich genaue Vorhersagen möglich, die kausal kaum erklärbar scheinen und oft völlig unerwartet sind. Da die für ein Verständnis der Funktionsweise dieser Analysen notwendigen Grundlagen oft nicht bekannt sind, sind diese für die meisten Personen kaum verständlich und einschätzbar, sodass eine Vielfalt an Phänomenen in diesem Bereich entsteht.

- Anhand von Datenanalysen können unerwartete Dinge in unerwarteter Genauigkeit vorhergesagt werden.
- Die anhand von Datenanalysen getroffenen Entscheidungen erscheinen oft willkürlich bzw. kaum nachvollziehbar.
- Durch schnelle Datenanalysen können z. B. Trends teils in Echtzeit erkannt werden.
- Selbst beim Erstkontakt mit einem Unternehmen/Dienst scheint dieses/dieser schon Informationen über eine Person zu haben.

Bereich *Datensicherheit und Datenschutz*. Daten werden meist nicht kontinuierlich erfasst, sondern üblicherweise auch über (oft ungeschützte) Kommunikationskanäle wie das Internet übertragen oder auf mobilen Datenträgern, meist ohne geeignete Schutzmechanismen, gespeichert. Dadurch entstehen deutliche Herausforderungen für die Sicherstellung der Vertraulichkeit, den Schutz vor Manipulation und den Schutz der Privatsphäre.

- Trotz bestmöglicher Anonymisierung können personenbezogene Daten unerwartet oft deanonymisiert werden.
- Obwohl ein PC mit einem Passwort geschützt ist, können andere Personen oft relativ einfach Daten davon „stehlen“.
- Eine Webseite kann ihre Nutzer anscheinend identifizieren, ohne dass sich diese einloggen und ohne dass Cookies aktiviert sind.
- Datenanalysen scheinen Dinge über Personen zu wissen, die diese nirgends angegeben haben.

Diese exemplarisch herausgegriffenen Phänomene weisen wiederum vielfältige Bezüge zu verschiedenen Kontexten auf, in denen Daten im Alltag zutage treten. Beispielsweise sind im Kontext von *Smart Home* Datenanalysen und damit einhergehende Phänomene von Bedeutung: Wenn beispielsweise erkannt werden soll, ob eine Person zuhause ist, oder wenn das Smart Home Gewohnheiten erlernen und dadurch proaktiv reagieren soll, ist re-

lativ einfach erkennbar, dass aus den erfassten Daten zum Teil mehr herausgelesen werden kann als erwartet. Gleichzeitig stellt sich aber auch die Frage, warum und welche Metadaten dabei aufgezeichnet werden, wie die Sicherheit der Daten bestmöglich sichergestellt wird und vieles mehr. Bei Betrachtung verschiedener Kontexte sind somit noch viele weitere Phänomene zu entdecken, die vorher nicht genannt wurden, da diese exemplarische Auswahl nur einen Eindruck der Vielfalt geben kann und soll. In Kombination mit den Kontexten bilden Phänomene einen motivierenden Rahmen für den Informatikunterricht: Beispielsweise kann der Kontext in Abstimmung mit den Schülerinnen und Schülern bzw. in Orientierung an deren Interessen ausgewählt werden, gleichzeitig durch die gezielte Thematisierung konkreter Phänomene aber trotzdem der Unterricht so gelenkt werden, dass die Förderung der angestrebten Kompetenzen ermöglicht wird. Eine entsprechende Nutzung der Phänomene als Einstieg in den Unterricht wird beispielsweise in der in Kapitel 12 vorgestellten Unterrichtseinheit gewählt.

5.3 Datenmanagement im beruflichen Umfeld

Neben der Bedeutung im Alltag hat Datenmanagement mittlerweile auch eine weiterhin stetig zunehmende Bedeutung im beruflichen Umfeld erlangt: Der Umgang mit Daten ist heute nicht mehr nur ein Thema das informatiknahe Berufe beschäftigt, im Gegenteil können große Teile der auch im Privatleben auftretenden Herausforderungen direkt auf das Berufsleben übertragen werden und sind dort mindestens ebenso relevant. Insbesondere die *Datenanalyse* ist im beruflichen Umfeld bereits heute wesentlich relevanter und in den verschiedensten Berufszweigen noch deutlicher vertreten als im Privatleben. Mit der zunehmenden ökonomischen Bedeutung von Daten – das World Economic Forum *(Thirani und Gupta, 2017)* setzt die Bedeutung von Daten heute sogar mit der von Öl gleich (*„data is definitely the new age ‚oil'"*) – wurde die Beschäftigung mit Big Data und die Positionierung eines Unternehmens dazu zu einem zentralen wirtschaftlichen Faktor. Aufgrund dieser immer stärker zunehmenden Relevanz von Datenanalysekompetenzen in verschiedenen Bereichen entwickelt sich seit einigen Jahren das neue Berufsbild des *Data Scientist*. Dieser Datenwissenschaftler ist ein Experte sowohl im Umgang mit Daten als auch in deren Analyse und bringt umfangreiche Kompetenzen aus dem Datenmanagement, aber auch beispielsweise aus der Mathematik, mit. In den letzten Jahren wurden immer mehr Studiengänge eingerichtet, die zum Erwerb der für diese neue Berufsrichtung nötigen Kompetenzen beitragen. Ein bekannter Vorreiter ist dabei der Studiengang *Master of Information and Data Science* an der Universität Berkeley, der verschiedene grundsätzliche Fähigkeiten eines Data Scientists ausbilden soll (vgl. Abbildung 5.2). Über die genaue Ausrichtung der Data-Science-Studiengänge scheint aber noch keine klare Einigkeit zu bestehen, sodass verschiedene Studiengänge durchaus alle Facetten zwischen theoretischer, praktischer und wissenschaftlicher Prägung annehmen können. Einigkeit besteht jedoch darüber, dass das Berufsbild des Data Scientists sich nicht nur aus einer informatischen Perspektive mit Datenanalysen befassen kann, sondern auch klare Verbindungen zu Mathematik und Statistik aufweist.

Abbildung 5.2: Angestrebte Fähigkeitsbereiche im Studiengang *Master of Information and Data Science* der Universität Berkeley. *(Quelle: `https://datascience.berkeley.edu/academics/`).*

Gerade in großen Unternehmen nehmen Daten und Datenanalysen, meist unter dem Stichwort Big Data, heute bereits eine immer zentralere Rolle ein: *„Betrachtet man die kommerziellen Aspekte bei der Nutzung von Big Data erweist sich die schnelle Nutzung von Markt-, Kunden- und Nutzerdaten zunehmend als wichtiger Wettbewerbsfaktor“ (Dorschel, 2015)*. Diese Nutzung von Daten ist jedoch nicht nur auf große Konzerne beschränkt, auch in kleinen und mittelständischen Unternehmen nimmt deren Bedeutung zu. Entsprechend stellt Datenmanagement heute nicht nur ein interessantes Thema für allgemeinbildende Schulen, sondern auch einen wichtigen Bereich der beruflichen Bildung dar, der nicht vernachlässigt werden sollte. Obwohl in dieser Arbeit der allgemeinbildende Charakter von Datenmanagement im Vordergrund steht, können die Ergebnisse dieser Arbeit auch als Basis für eine Weiterentwicklung der Konzepte beruflicher Bildung in diesem Bereich dienen.

6 Ausgangslage für den Informatikunterricht im Bereich Datenmanagement

Neben der fachlichen und fachdidaktischen Forschung sowie der Bedeutung von Datenmanagement im täglichen Leben, stellt die Perspektive der am Schulunterricht beteiligten Akteure eine wichtige Grundlage für diese Arbeit dar. Daher wird in einer Fragebogenstudie die Lehrerperspektive auf das Themenfeld betrachtet und untersucht, welches Wissen die Lehrerinnen und Lehrer bereits zu den Themen mitbringen, welche sie als besonders interessant für den Unterricht erachten und welche Herausforderungen sie erwarten. Als dritte Perspektive wurden Schülerinnen und Schüler in einer weiteren Fragestellung befragt, um ihr Vorwissen und ihre Erfahrungen mit ausgewählten Bereichen des Datenmanagements zu charakterisieren und so die dritte Forschungsfrage zu beantworten.

6.1 Lehrerperspektive auf das Fachgebiet Datenmanagement

Gerade aufgrund der deutlichen Veränderungen, die sich im Fachgebiet Datenmanagement heute gegenüber dem aktuellen Stand in der Schule zeigen, ist es bei einer fachdidaktischen Aufbereitung dieses Themengebiets essenziell, auch die Sichtweise der Lehrerinnen und Lehrer einzubeziehen: Diese müssen die Veränderungen der letzten Jahre nachvollziehen und ihre Fachkenntnisse auf einem aktuellen Stand halten, aber auch den Umgang mit solchen Themen im Unterricht neu überdenken und diesen ggf. neu gestalten. Dies wird insbesondere deswegen relevant, weil erwartet werden kann, dass das Wissen der Lehrkräfte zu modernen Themen deutlich geringer ausgeprägt ist, als zu den in der Schule bereits seit Jahren etablierten und typischerweise auch im Lehramtsstudium fest verankerten Themen. Selbst in Fortbildungen und Workshops werden aktuellere Themen des Datenmanagements bis dato kaum thematisiert, sodass eine kontinuierliche Lehrerfortbildung kaum sichergestellt ist. Um im Rahmen dieser Arbeit einen Eindruck über die Ausprägung des Wissens der Lehrerinnen und Lehrer in Zusammenhang mit Datenmanagement zu erhalten, wurde eine quantitative Untersuchung der Selbsteinschätzung von Lehrkräften durchgeführt.

6.1.1 Ziele der Untersuchung

Zur Erforschung der Perspektiven von Lehrerinnen und Lehrern auf das Fachgebiet zu bekommen, wurden zuerst relevante Bereiche ermittelt, denen dieses Wissen zugeordnet werden kann und hinsichtlich derer die Lehrkräfte daraufhin befragt werden konnten. Solche Wissensbereiche werden durch verschiedene Modelle systematisiert, beispielsweise das in der Informatikdidaktik bewährte, *TPACK-Modell* (vgl. Abbildung 6.1; *Mishra und Koehler (2006)*). Dieses Modell gliedert das Wissen von Lehrerinnen und Lehrern in die Bereiche

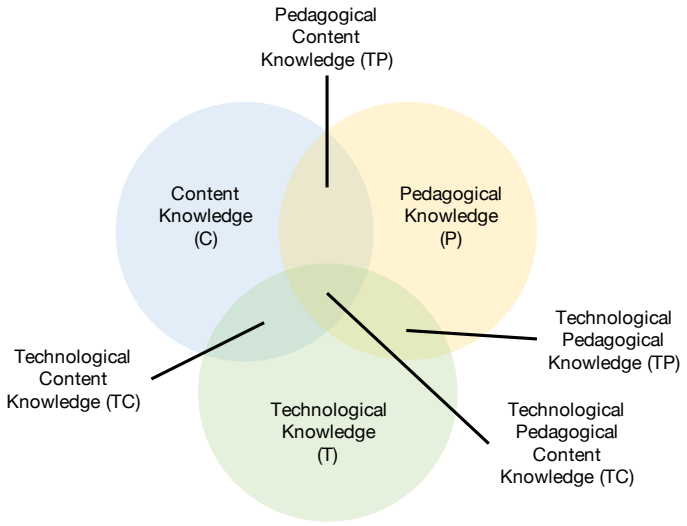


Abbildung 6.1: TPACK-Modell nach *Mishra und Koehler (2006)*.

Technological Knowledge (TK), *Pedagogical Knowledge* (PK) und *Content Knowledge* (CK) sowie deren Überschneidungsbereiche (PCK, TCK, PCK, TPK, TPCK). Diese Dimensionen können auch bei einer Untersuchung des aktuellen Wissenstandes der Lehrkräfte hilfreich eingesetzt werden: Während davon auszugehen ist, dass das pädagogische Wissen der Lehrer themenübergreifend und daher auch im Bereich Datenmanagement anwendbar ist, sind themenbezogene Unterschiede insbesondere im *content knowledge*, aber auch in den damit zusammenhängenden Überschneidungsdimensionen *technological content knowledge*, *pedagogical content knowledge* und *technological pedagogical content knowledge* zu erwarten, da diese Wissensdimensionen durch die inhaltlichen Veränderungen im Datenmanagement beeinflusst werden.

Um die diese betroffenen Dimensionen des TPACK-Modells abzudecken, wurden aus der Vielzahl möglicher Schwerpunkte für eine derartige Untersuchung drei Fragestellungen ausgewählt. Diese können einen wertvollen Einblick in die Lehrerperspektive geben:

- Wie tiefgreifendes Wissen haben Lehrerinnen und Lehrer zu zentralen Themen des Datenmanagements? *(content knowledge)*
- Welche Themen des Datenmanagements erachten Lehrerinnen und Lehrer als besonders interessant für den Informatikunterricht? *(pedagogical content knowledge)*
- Welche Herausforderungen erwarten Lehrerinnen und Lehrer, wenn sie Datenmanagementthemen im Informatikunterricht thematisieren? *(content knowledge, pedagogical content knowledge, technological content knowledge, technological pedagogical content knowledge)*

6.1.2 Untersuchungsmethode: Fragebogenstudie

Zur Untersuchung der Fragestellungen wurde eine Fragebogenstudie mit Lehrerinnen und Lehrern verschiedener Schulen, Schularten und Bundesländer geplant und durchgeführt. Dazu wurde zu jeder der drei Fragestellungen je eine Frage erstellt. Zusätzlich wurde eine Liste von Datenmanagementthemen entwickelt, die aus der bereits vorhandenen Vorarbeit im Rahmen der Lehrplananalyse in Abschnitt 4.3 und damit auch indirekt aus dem DAMA-DMBoK, abgeleitet wurde. Die ausgewählten Themen waren: *(klassische) Datenbanken, NoSQL/non-relationale Datenbanken, verteilte Datenbanken, Cloud-Speicher, Cloud-Computing, Datenanalyse (klassisch), Data Mining, Big Data, Open Data, Verschlüsselung von Daten, Datenmodellierung, Funktionsweise von Suchmaschinen, CAP-Theorem, ACID-Prinzip, BASE-Prinzip, Metadaten, Datensicherheit (z. B. Backup), Datenschutz* sowie *Gefahren bei der bzw. durch die maschinelle Verarbeitung von Daten*.

In einem Fragebogen (vgl. Anhang A), der den Lehrerinnen und Lehrern zur Beantwortung vorgelegt wurde, sollte jedes der drei folgenden Fragebogenitems für jedes der Datenmanagementthemen ausgewertet werden:

- Wie schätzen Sie ihr eigenes Wissen zu den Themen ein?
- Wie interessant finden Sie diese Themen für den Informatikunterricht?
- Wo sehen Sie die Schwierigkeiten bei der Umsetzung für den Informatikunterricht?

Die Einschätzung zu den ersten beiden Fragen sollte dabei auf einer Likert-Skala erfolgen, die folgende Antwortmöglichkeiten erlaubte:

- Zu Frage 1: *unbekannt*, *kaum Wissen*, *grundlegendes Wissen*, *detailliertes Wissen*. Zur klareren Differenzierung der Antwortmöglichkeiten wurde allen Teilnehmern die Erklärung mitgegeben, dass unter *kaum Wissen* solches verstanden wird, das sie als nicht ausreichend zur Thematisierung des Themas im Unterricht einschätzen. *Grundlegendes Wissen* erlaubt die oberflächliche Betrachtung eines Themas im Unterricht und *detailliertes Wissen* ist auch zur Thematisierung innerhalb eines Oberstufenkurses ausreichend.
- Zu Frage 2: *nicht interessant*, *kaum interessant*, *eher interessant*, *sehr interessant*. An dieser Stelle wurde keine weitere Erklärung der Dimensionen mitgeliefert, da es sich um eine subjektive Einschätzung der Lehrerinnen und Lehrer handeln sollte, die insbesondere dazu dienen sollte, das unterschiedliche Interesse an den verschiedenen Themen zu ermitteln. Eine Normierung der Antworten, wie sie durch die Erklärung zu Frage 1 bezweckt wurde, war daher an dieser Stelle weder nötig noch sinnvoll, da sich Abweichungen auf alle Antworten eines Fragebogens und somit alle Themen zugleich beziehen würden.

Im Gegensatz zu den ersten beiden Fragen wurde der dritten keine Likert-Skala zugrunde gelegt, sondern es werden die drei zentralen Schwierigkeiten, die in vorherigen Diskussionen in Rahmen von Lehrerfortbildungen und anderen Tagungen häufig genannt worden

sind, als Antwortmöglichkeiten angeboten. Die Lehrerinnen und Lehrer sollten dabei entscheiden, ob sie erwarten, dass diese Schwierigkeit im Rahmen eines möglichen Datenmanagementunterrichts auftritt oder nicht. Mehrfachantworten waren daher in diesem Fall explizit erwünscht.

6.1.3 Durchführung und Auswertung

Der Fragebogen wurde im Rahmen von drei Lehrerfortbildungsworkshops vor der eigentlichen Workshopdurchführung an insgesamt 80 Lehrerinnen und Lehrer unterschiedlichen Hintergrunds verteilt:

- 63 Lehrkräfte kamen aus Bayern, 17 aus der Region Berlin-Brandenburg.
- 72 Lehrkräfte unterrichteten an Gymnasien (60 Bayern, 12 Berlin-Brandenburg), eine/r an einer Fachoberschule (Bayern), zwei an Integrierten Sekundarschulen (Berlin), eine/r an einem Oberstufenzentrum (Berlin), zwei an Realschulen (Bayern), eine/r an einer Gesamtschule mit gymnasialer Oberstufe (Berlin-Brandenburg) und eine/r an einer außerschulischen Bildungseinrichtung (Berlin-Brandenburg).
- Mindestens 58 der bayerischen Lehrkräfte waren als Fachbetreuer bzw. Fachbetreuerinnen an ihren jeweiligen Schulen eingesetzt.

Da die Befragung im Rahmen von Fortbildungsveranstaltungen zum Thema Datenmanagement stattfand, kann eine Beeinflussung der Ergebnisse nicht völlig ausgeschlossen werden: Obwohl die Verteilung und Beantwortung der Fragebögen noch vor der eigentlichen Fortbildung erfolgte, sind unter den Teilnehmenden sicherlich eher grundsätzlich am Thema interessierte Lehrkräfte vertreten. Da jedoch die Befragung für die 58 Teilnehmenden bayerischen Fachbetreuer bzw. Fachbetreuerinnen im Rahmen von Pflichtfortbildungen stattfand und nur die verbleibenden 22 Personen freiwillig an der Fortbildung teilgenommen haben, kann durch einen Vergleich der jeweiligen Ergebnisse eine Einschätzung dieser Beeinflussung erfolgen. Tendenziell muss jedoch mit einem etwas erhöhten Interesse gerechnet werden. Obwohl sowohl aufgrund der kleinen Stichprobengröße als auch der möglicherweise vorhandenen Vorprägung der Gruppe keine validen Rückschlüsse auf die Gesamtpopulation der Lehrkräfte möglich sind, geben die Ergebnisse einen klaren Einblick in die Lehrerperspektive auf das Thema Datenmanagement und sind damit als erste Basis hilfreich für die weitere Forschung zu diesem Thema.

Die Auswertung der Daten erfolgte mithilfe von statistischen Methoden. Bevor diese angewendet werden konnten, mussten jedoch die digitalisierten Ergebnisse der Fragebögen zum Teil noch bereinigt werden: Zum Teil wurden von manchen Teilnehmern mehr als eine Antwort bei Likert-skalierten Fragen angekreuzt, vermutlich mit dem Ziel ein Schwanken zwischen den beiden Antworten anzuzeigen. Da dies nicht valide auswertbar war, wurden solche Antworten als ungültig und die jeweilige Teilfrage als nicht beantwortet gewertet. Genauso wurde zum Teil im Rahmen der ersten Frage zum Wissen der Lehrkräfte angegeben, dass ein Thema *unbekannt* sei, dann jedoch Antworten zu Interesse und

erwarteten Schwierigkeiten gegeben. Auch hier kann davon ausgegangen werden, dass diese Antworten kaum valide sind, sodass bei Einstufung eines Themas als unbekannt die möglicherweise vorhandenen Antworten zu Fragen 2 und 3 nicht berücksichtigt wurden.

Nach der Bereinigung der Daten wurden diese aggregiert: Für Fragen 1 und 2 wurden der Median und Modus der Antworten für jedes der betrachteten Themen berechnet. Beide Werte eignen sich gut zur Beschreibung der Ergebnisse, da es sich dabei um kardinalskalierte Daten handelt. Der Median beschreibt dabei den Wert, der bei Sortierung aller Werte in der Mitte steht und somit die Stichprobe in zwei gleiche Teile trennt. Der Modus bzw. Modalwert ist der am häufigsten vorkommende Wert in einer Stichprobe. Während diese beiden Werte zwar eine Beschreibung der Stichprobe über ihre zentralen Werte erlauben, sagen sie jedoch nichts über die Streuung der Werte aus. Daher wurde zusätzlich die mittlere Standardabweichung bezüglich des Medians berechnet (in den Ergebnistabellen bezeichnet als $\bar{d}_{0.5}$). Für die Auswertung der dritten Frage wurde (pro Thema) aufsummiert, wie viele Lehrkräfte mit den entsprechenden Schwierigkeiten rechnen sowie der prozentuale Anteil an den abgegebenen Antworten berechnet. Die Ergebnisse sind in Tabelle 6.1 dargestellt.

Um die Validität der Ergebnisse trotz der geringen Stichprobengröße sicherzustellen, wurde überprüft, ob in den Daten Muster erkennbar sind, die auf eine Beeinflussung der Ergebnisse schließen lassen. Dabei konnte festgestellt werden, dass die Ergebnisse sich bei Vergrößerung der Stichprobe nur relativ wenig geändert haben: Die Auswertung wurde dazu zuerst mit einem kleineren Teil der Antworten durchgeführt und dann um zusätzliche Daten erweitert. Dabei konnte festgestellt werden, dass schon bei ca. 30 ausgewerteten Fragebögen (die jeweils zu ca. 50% aus Bayern und Berlin-Brandenburg stammten) das Ergebnis relativ gut mit dem am Ende ermittelten übereinstimmte. Die weiteren Fragebögen dienten daher insbesondere zur Absicherung der Ergebnisse und damit zur Verringerung der Streuung. Die Untersuchung, ob der Verpflichtungsgrad der Fortbildung einen Einfluss auf die Ergebnisse hatte, erwies sich als negativ: Ein Vergleich der beiden Subgruppen hinsichtlich der ersten beiden Fragen (vgl. Abbildung 6.2 und 6.3) zeigt zwar Unterschiede, diese sind jedoch nur sehr gering ausgeprägt und betreffen nur vereinzelte Themen, sodass eine generelle Beeinflussung nicht erkennbar ist. Entsprechend verhält es sich bei den erwarteten Schwierigkeiten, bei denen auch keine relevanten Unterschiede erkannt werden können. Gleichzeitig kann diese Betrachtung auch einen potenziellen regionalen Einfluss ausschließen: Da der Besuch der Fortbildung für 58 der 63 bayerischen Lehrkräfte im Rahmen ihrer dienstlichen Aufgaben verpflichtend war, während der Rest der bayerischen Lehrkräfte und die 17 Lehrkräfte aus Berlin/Brandenburg keinerlei Teilnahmepflicht hatten, müsste sich ein regionaler Einfluss bei dieser Betrachtung zeigen und kann daher ausgeschlossen werden.

	Einschätzung des Wissens 0 = unbekannt 1 = kaum Wissen 2 = grundlegendes Wissen 3 = detailliertes Wissen				geschätztes Interesse 0 = nicht interessant 1 = kaum interessant 2 = eher interessant 3 = sehr interessant				erwartete Herausforderungen % der befragten Lehrkräfte		
	# Antworten	Modus	Median	$\bar{d}_{0.5}$	# Antworten	Modus	Median	$\bar{d}_{0.5}$	fehlendes Fachwissen	fehlende Werkzeuge	Komplexität
(klassische) Datenbanken	79	3	3	0,42	79	3	2	0,54	1,3 %	16,5 %	10,1 %
NoSQL / non-relationale Datenbanken	78	1	1	0,46	62	1	1	0,52	49,4 %	10,1 %	15,2 %
Verteilte Datenbanken	78	1	1	0,53	63	1	1	0,63	44,3 %	17,7 %	13,9 %
Cloud-Speicher	79	2	2	0,39	77	2	2	0,55	26,6 %	22,8 %	5,1 %
Cloud-Computing	76	2	2	0,59	70	2	2	0,64	31,6 %	31,6 %	10,1 %
Datenanalyse (klassisch)	77	2	2	0,47	75	2	2	0,61	20,3 %	21,5 %	7,6 %
Data Mining	78	1	1	0,44	62	2	2	0,81	46,8 %	20,3 %	6,3 %
Big Data	75	1	1	0,56	65	2	2	0,63	36,7 %	22,8 %	12,7 %
Open Data	74	1	1	0,59	45	1	1	0,82	38,0 %	12,7 %	2,5 %
Verschlüsselung von Daten	78	2	2	0,35	78	3	2	0,62	12,7 %	16,5 %	20,3 %
Datenmodellierung	78	2	2	0,50	77	3	2	0,69	8,9 %	6,3 %	3,8 %
Funktionsweise von Suchmaschinen	77	2	2	0,36	76	2	2	0,59	16,5 %	15,2 %	5,1 %
CAP-Theorem	73	0	0	0,15	10	1	1	0,50	45,6 %	0,0 %	2,5 %
ACID-Prinzip	75	0	0	0,55	26	0	1	0,73	34,2 %	0,0 %	3,8 %
BASE-Prinzip	75	0	0	0,16	11	1	1	0,45	41,8 %	0,0 %	3,8 %
Metadaten	77	2	2	0,61	71	2	2	0,72	25,3 %	17,7 %	3,8 %
Datensicherheit (z. B. Backup)	77	2	2	0,35	77	3	2	0,75	11,4 %	15,2 %	0,0 %
Datenschutz	79	2	2	0,29	79	3	3	0,54	12,7 %	13,9 %	6,3 %
Gefahren bei der bzw. durch die maschinelle Verarbeitung von Daten	78	2	2	0,28	78	3	2	0,69	10,1 %	20,3 %	6,3 %

Tabelle 6.1: Ergebnisse des Lehrerfragebogens.

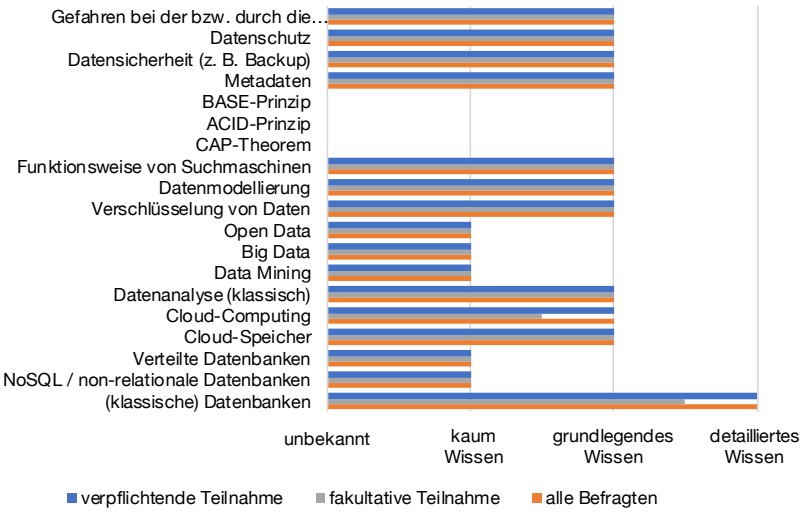


Abbildung 6.2: Vergleich der Ergebnisse bezüglich des Wissens der Lehrerinnen und Lehrer nach Subgruppen (aufgetragen ist jeweils der Median der Subgruppen bezüglich der Themen).

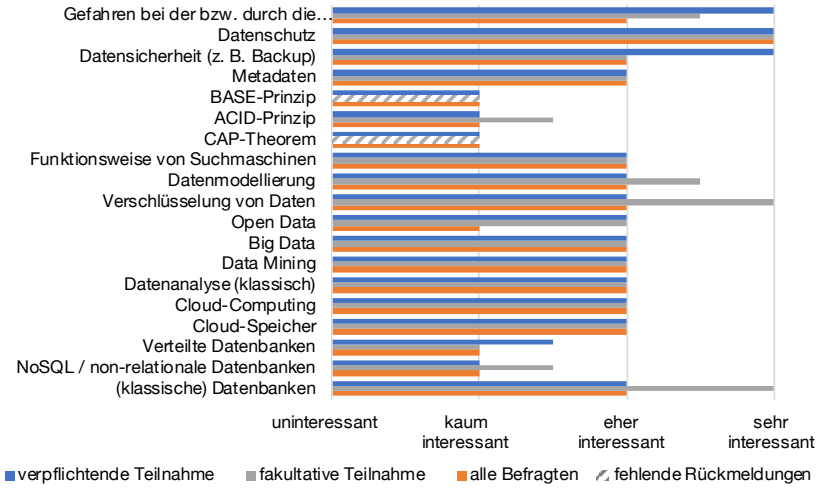


Abbildung 6.3: Vergleich der Ergebnisse bezüglich der Interessantheit der Themen nach Subgruppen (aufgetragen ist jeweils der Median der Subgruppen bezüglich der Themen). *Hinweis: Aufgrund fehlender Rückmeldungen zum CAP-Theorem und BASE-Prinzip in der Subgruppe mit fakultativer Teilnahme können hier keine Ergebnisse angegeben werden.*

6.1.4 Interpretation

Frage 1: Fachwissen der Lehrerinnen und Lehrer

Die Ergebnisse zur Selbsteinschätzung der Lehrerinnen und Lehrer bezüglich ihres Wissens zu aktuellen Datenmanagementthemen zeigen, dass relativ unterschiedliches Vorwissen auch innerhalb der einzelnen Themen herrscht: Für die meisten liegt die Streuung im Bereich $\bar{d}_{0.5} \leq 0.50$, für wenige wurde dieser Grenzwert geringfügig überschritten. So-

mit gibt es bei den meisten Themen jeweils Lehrerinnen bzw. Lehrer, die nach eigener Einschätzung stark ausgeprägtes Fachwissen besitzen als auch solche, die dieses als eher gering einschätzen. Nur bei vier Themen (CAP-Theorem, BASE-Paradigma, Datenschutz und Gefahren einer automatisierten Datenverarbeitung) war eine besonders hohe Übereinstimmung in der Einschätzung der verschiedenen Lehrerinnen und Lehrer zu erkennen, die Streuung lag in diesen Fällen bei $\bar{d}_{0.5} \leq 0.3$. Im Allgemeinen zeigen die Ergebnisse, dass die Lehrerinnen und Lehrer *kaum bis grundlegendes Wissen* zu den meisten Datenmanagementthemen haben. Gleichzeitig sind ihnen jedoch im Allgemeinen bereits alle Themen außer drei (CAP-Theorem, ACID-Paradigma, BASE-Paradigma) überwiegend zumindest bekannt (vgl. Abbildung 6.4). Das ist auch der Fall für Lehrerinnen bzw. Lehrer, die die jeweiligen Themen in der zweiten Frage als eher uninteressant für den Unterricht eingeschätzt haben. Es kann daher angenommen werden, dass selbst ohne spezielles Interesse am Thema, Informatiklehrkräfte mit den relevanten Begriffen aus diesem Bereich in Berührung kommen.

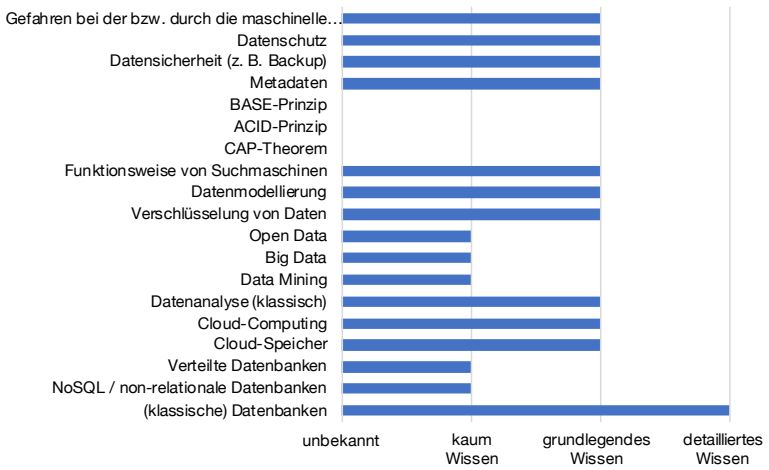


Abbildung 6.4: Median der Selbsteinschätzung des Wissens von Lehrerinnen und Lehrern zu verschiedenen Datenmanagementthemen.

In Bezug auf *relationale Datenbanken* schätzen nahezu alle Befragten ihr Wissen als detailliert ein, während zu anderen bereits häufig im Unterricht berücksichtigten oder mit solchen eng verwandten Themen das Wissen eher als *grundlegend* eingeschätzt wird. Dies ist insbesondere für *Datenanalysen*, *Verschlüsselung* und *Metadaten* der Fall. Themen, die typischerweise kaum im Informatikunterricht thematisiert werden, wie beispielsweise *verteilte Datenbanken*, *Big Data* oder *Data Mining*, sind weniger bekannt – die Befragten geben an, *kaum Wissen* zu diesen Themen zu haben. Die einzigen, zu denen die Lehrerinnen und Lehrer angeben (mindestens) *grundlegendes Wissen* zu haben, die aber gleichzeitig nicht bereits übliches Unterrichtsthema sind, stellen *Cloud-Speicher* und *Cloud-Computing* dar.

Nur drei der betrachteten Themen waren den Teilnehmerinnen und Teilnehmern größtenteils unbekannt, zwei davon auch mit einer geringen Streuung von $\bar{d}_{0.5} = 0.15$ bzw. $\bar{d}_{0.5} = 0.16$. Bei diesen handelt es sich um die drei für das Datenmanagement zentralen

Prinzipien *CAP-Theorem*, *ACID-Paradigma* und *BASE-Paradigma*. Diese stellen zentrale theoretische Grundlagen für die Funktionsweise von Datenmanagementsystemen dar. Somit handelt es sich dabei um eine relevante Wissenslücke der Lehrerinnen und Lehrer, die es für einen vertieften Unterricht zu diesen Themen zu schließen gilt.

Frage 2: Attraktivität für den Informatikunterricht

Im Allgemeinen stuften die befragten Lehrerinnen und Lehrer die genannten Datenmanagementthemen als eher interessant ein (vgl. Abbildung 6.5). Dabei ist jedoch hervorzuheben, dass kein Thema im Median als *uninteressant* eingeordnet wurde, einzig das ACID-Prinzip wurde im Modus als *uninteressant* eingeschätzt. Nur ein geringer Anteil der Lehrkräfte hat Themen überhaupt als *uninteressant* eingestuft. Insbesondere *Datensicherheit* und *Gefahren bei der bzw. durch die maschinelle Verarbeitung von Daten* wurden im Median als für den Unterricht *interessant* eingestuft, *Datenschutz* sogar als sehr interessant. Im Gegensatz zu diesen gesellschaftlich auch häufig klar diskutierten Themen, wurden eher technologisch ausgerichtete Aspekte wie *nicht-relationale* und *verteilte Datenbanken*, *Open Data* sowie die *ACID-* und *BASE-Prinzipien* und das *CAP-Theorem* als weniger interessant eingestuft. Trotzdem kann im Gesamtbild erkannt werden, dass diese Themen ein grundlegendes Interesse wecken. Aufgrund der großen Streuung der Ergebnisse, die in fast allen Fällen über $\bar{d}_{0.5} = 0.50$ lag, kann jedoch vermutet werden, dass hier noch keine gefestigte Einschätzung existiert: Nahezu jedes Thema wurde von einem Teil der Befragten als interessant oder sehr interessant gewertet, während ein anderer Teil es für eher uninteressant einschätzte. Im mündlichen Gespräch nach den Fortbildungsveranstaltungen zeigte sich jedoch eine höhere Einschätzung des Interessantheitsgrades durch die Lehrerinnen und Lehrer als diese es vorher bekundet hatten. Die Streuung der Ergebnisse kann daher möglicherweise durch das geringe Vorwissen und eine unterschiedlich ausgeprägte Neugierde bzw. Offenheit neuen Themen gegenüber erklärt werden.

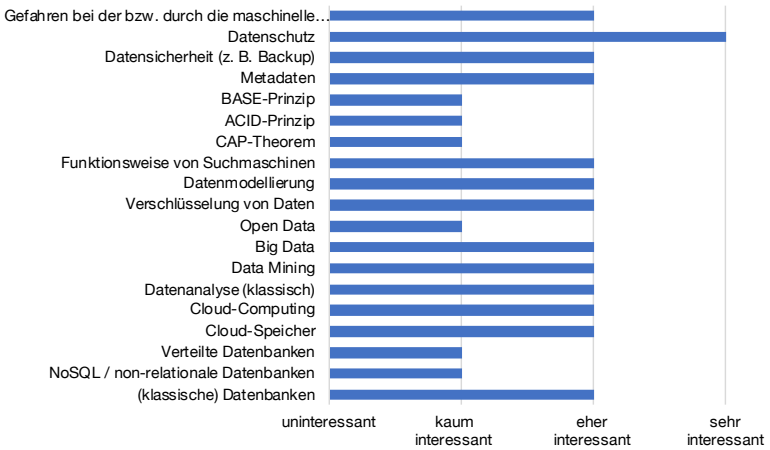


Abbildung 6.5: Median der Einschätzung der Interessantheit von Datenmanagementthemen durch die Lehrerinnen und Lehrer.

Frage 3: Erwartete Schwierigkeiten im Informatikunterricht

Bei der Thematisierung von Datenmanagementthemen im Unterricht sehen die meisten Lehrerinnen und Lehrer die größte Schwierigkeit darin, dass ihnen selbst das notwendige *Fachwissen fehlt* (vgl. Abbildung 6.6). Dies zeigt, zusammen mit den Ergebnissen zur ersten Frage, dass für eine erfolgreiche Berücksichtigung von Datenmanagementthemen im Unterricht geeignete Fortbildungsmaßnahmen, die den Aufbau von ausreichend tiefgreifendem Wissen und Kompetenzen aus dem Bereich Datenmanagement fördern, genauso unabdingbar sind wie Materialien, die die Lehrkräfte bei der unterrichtlichen Umsetzung geeignet unterstützen können. Außerdem sehen die Lehrerinnen und Lehrer auch die *fehlenden Werkzeuge* als problematisch an: Gerade zu moderneren Themen scheinen diese nicht zu existieren, den Lehrenden bisher unbekannt zu sein oder schlichtweg als ungeeignet wahrgenommen zu werden. Jedoch werden selbst beim bewährten Unterrichtsthema *Datenbanken* die zur Verfügung stehenden Werkzeuge von einem kleinen Teil der Befragten als eher ungenügend angesehen. Die dritte Antwortmöglichkeit – eine zu hohe Komplexität – wurde nur für wenige Themen als relativ relevant erachtet: Immerhin 20 % der Teilnehmenden gaben an, das Thema *Verschlüsselung* als für den Unterricht zu komplex einzuschätzen, obwohl es hier schon vielfältige durchaus positive Unterrichtserfahrungen gibt. Nur wenige weitere Themen (*NoSQL-Datenbanken*, *Verteilte Datenbanken*, *Big Data*, *Cloud Computing*) wurden von mehr als 10 % der Teilnehmenden als zu komplex erachtet. Nach der Befragung wurde dies auf Nachfrage oft so erläutert, dass die Lehrerinnen und Lehrer davon ausgehen, dass diese Themen entsprechend didaktisch reduziert im Unterricht thematisiert werden können, wie es auch für viele andere komplexe Themen der Fall ist. Die Schwierigkeit liegt nach Meinung der Befragten daher eher in der geeigneten didaktischen Reduktion als in der fehlenden Eignung der Themen für den Unterricht.

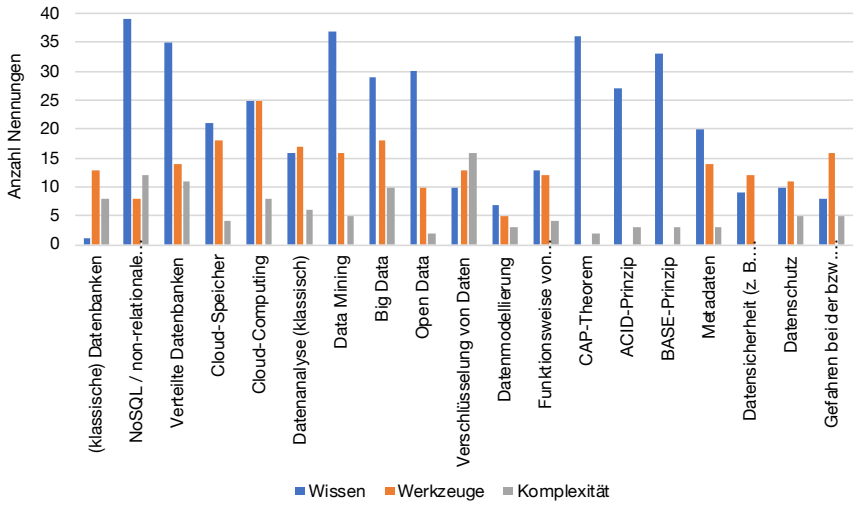


Abbildung 6.6: Von den Lehrerinnen und Lehrern erwartete Schwierigkeiten im Unterricht zu verschiedenen Datenmanagementthemen.

6.1.5 Zusammenfassung und Fazit

Die derzeit kaum vorherrschende Thematisierung des Themas Datenmanagement im unterrichtlichen Kontext zeigt sich auch an dem Vorwissen der Lehrerinnen und Lehrer: Die Hypothese, dass diese nur sehr eingeschränktes Wissen zu solchen Datenmanagementthemen haben, die über die klassischen Themen hinausgehen, konnte durch die Untersuchung bestätigt werden. Trotzdem zeigten viele Lehrerinnen und Lehrer Interesse an diesen Themen, wobei hier aufgrund der hohen Streuung der Ergebnisse valide Aussagen schwierig sind. Es kann vermutet werden, dass diese hohe Streuung insbesondere auch durch die mangelnden Kenntnisse der Lehrkräfte bedingt ist, weswegen ihnen eine tiefergehende Einschätzung der Themen nur eingeschränkt möglich war. Dies bestätigt sich durch die Herausforderungen, die die Befragten insbesondere bei dem Mangel an eigenem Fachwissen und zusätzlich bei den fehlenden Werkzeugen für den Unterricht sahen, nicht aber in der Komplexität der Themen. Entsprechend können durch diese Untersuchung drei zentrale Herausforderungen für die Steigerung der Relevanz von Datenmanagementthemen im Unterricht gesehen werden: Die Fortbildung der Lehrkräfte hinsichtlich der Prinzipien des Datenmanagements, die Gestaltung geeigneter Unterrichtskonzepte die auch für Lehrkräfte wichtige Umsetzungshinweise enthalten und die Entwicklung von für den Einsatz im Unterricht geeigneten Werkzeugen.

6.2 Schülerperspektive auf das Fachgebiet Datenmanagement

Neben der Lehrerperspektive stellt auch die Schülerperspektive auf ein Thema ein wichtiges Fundament für erfolgreichen Unterricht dar. Obwohl Datenmanagement für Schülerinnen und Schüler möglicherweise derzeit weniger greifbar ist als andere Themen der Informatik und es auch kaum im Informatikunterricht thematisiert wird, ist aufgrund der Allgegenwärtigkeit im Alltag zu erwarten, dass jeder Jugendliche schon mit verschiedenen Aspekten des Datenmanagements in Berührung gekommen ist. Beispielsweise in Zusammenhang mit der Nutzung von Smartphones, sozialen Netzwerken oder der Synchronisierung eigener Daten zwischen verschiedenen Geräten sind daher bereits vor einer potenziellen Thematisierung im Unterricht vielfältige Vorstellungen bei den Schülerinnen und Schülern zu erwarten. Um auch in diesem Bereich eine Basis für die weitere Forschung zu schaffen, wurden die Erfahrungen von Schülerinnen und Schülern mit und deren Wissen zu verschiedenen Themen des Datenmanagements explorativ untersucht.

6.2.1 Ziele der Untersuchung

Um einen Überblick über die Erfahrungen der Lernenden mit diesen Themen zu bekommen und um das Wissen, das sie dabei bereits vorunterrichtlich aufgebaut haben, zu explorieren, wurde die zweite Forschungsfrage untersucht: *Wie sind das Vorwissen und die*

Kenntnisse der Schülerinnen und Schüler zu dem Datenmanagement zugehörigen Themen ausgeprägt?. Diese Untersuchung wurde auf drei zentrale Themenbereiche fokussiert:

- *Datenbanken und Datenanalysen:* Diese Themen sind bereits heute Teil des Informatikunterrichts, aus fachlicher Perspektive gehen sie jedoch deutlich über die dort thematisierten Aspekte hinaus. Die beiden Themen stellen daher eine Möglichkeit dar, den Umfang des ggf. schon erworbene Fachwissen der Schülerinnen und Schüler zu ermitteln.
- *Metadaten im Alltag:* Metadaten stellen heute ein breit diskutiertes Thema der Informatik dar, das in vielen Alltagssituationen Bedeutung hat und mit dem heute jeder in Berührung kommt. Dieses Thema gibt damit Aufschluss über die Wahrnehmung von Daten im täglichen Leben.
- *Wert der eigenen Daten:* Jeder verwaltet heute eine große Menge eigener und fremder Daten. Es ist daher wichtig, den Wert, der diesen beigemessen wird, zu erkennen. Zusätzlich zeigt dieser Aspekt auch, wie relevant das Thema für die Lernenden ist und kann dazu dienen, ihnen die Möglichkeiten und Gefahren von Datenanalysen zu verdeutlichen.

6.2.2 Untersuchungsmethode: Fragebogenstudie

Da man davon ausgehen kann, dass Schülerinnen und Schüler bisher kaum in Kontakt mit den meisten der Fachbegriffe gekommen sind, die in der vorherigen Untersuchung den Lehrkräften präsentiert worden sind, kann der Schülerfragebogen nicht analog zum Lehrerfragebogen aufgebaut werden. Stattdessen werden die Schülerinnen und Schüler in einem ersten Teil indirekt zu ihren Erfahrungen mit und ihrem Wissen über Metadaten befragt, indem sie mit entsprechenden Alltagssituationen, in denen Metadaten eine Rolle spielen, konfrontiert wurden:

- Du erstellst mit deinem Smartphone ein Foto. Welche Informationen werden dem Bild automatisch mitgeliefert (d. h. sind in der Bilddatei zusätzlich zum eigentlichen Bild als sog. Metadaten gespeichert)?
  Antwortmöglichkeiten: Datum/Uhrzeit; genauer Ort der Aufnahme (GPS-Daten); Name aller Personen auf dem Bild; Beschreibung was auf dem Bild zu sehen ist; Name des Fotografen; Informationen zur Kamera, mit der das Bild erstellt wurde
- Du besuchst eine Webseite im Internet. Welche Informationen kann diese über dich herausfinden? (Ohne dass du spezielle Einstellungen machen oder spezielle Programme installieren musst)
  Antwortmöglichkeiten: von welcher Webseite ich komme; welchen Browser ich benutze; welches Betriebssystem ich benutze; meinen genauen Standort (GPS-Daten); meinen Namen; Name einiger Programme, die ich installiert habe; meine E-Mail-Adresse; meine Interessen; mich eindeutig identifizieren; ob ich ein mobiles Gerät oder einen PC nutze; meine Monitorauflösung; meine Sprache; in welchem Land ich mich befinde; mein Alter

Beiden Fragen sind stark auf den Alltag der Schülerinnen und Schüler ausgerichtet: Es wurden zwei Situationen des alltäglichen Umgangs mit Daten ausgewählt und im Fragebogen zusammen mit verschiedenen potenziell in diesen Situationen erhobenen Daten präsentiert. Diese wurden so ausgewählt, dass eine der beiden eher direkten und bewussten Umgang der Schüler mit den Daten (hier in Form von Fotos und deren Metadaten) beinhaltet, während in der anderen Situation der Kontakt eher indirekt und unterbewusst stattfindet. Die Antworten auf diese beiden Fragen geben daher einen Eindruck davon, ob Schülerinnen und Schüler sich der kontinuierlichen Erfassung von Metadaten durch Informatiksysteme und über den Umfang dieser Daten bewusst sind.

Im zweiten Teil des Fragebogens wurden sie konkreter zu zwei Datenmanagementthemen befragt, die zur Vermeidung begrifflicher Schwierigkeiten so gewählt wurden, dass deren Grundsätze den befragten Schülerinnen und Schülern bereits aus dem Schulunterricht bekannt sein sollten[31]. Es wurden daher die Themen *Datenbanken* und *Datenanalysen* ausgewählt. In den dazu gestellten Fragen wird jedoch gezielt auch über das aus dem Schulunterricht zu erwartende Wissen hinausgegangen, um erkennen zu können, ob die Schülerinnen und Schüler das erworbene Wissen auf ähnliche Situationen übertragen können. Es werden den Teilnehmenden daher die folgenden Fragen und Antwortmöglichkeiten vorgelegt:

- Welche der folgenden Aussagen über Datenbanken sind deiner Meinung nach korrekt?
  Antwortmöglichkeiten: alle Daten müssen konsistent gespeichert sein; bei kleineren Datenmengen lohnt sich eine Datenbank nicht; große Datenmengen schaffen die meisten Datenbanken nicht; nur bis zu 5 Personen können gleichzeitig mit einer Datenbank arbeiten; jede Datenbank liegt auf einem eigenen Server; um Fotos, Videos und so weiter zu speichern, sind Datenbanken kaum geeignet; Cloud-Dienste basieren typischerweise auf Datenbanken

- Welche der folgenden Aussagen über Datenanalysen sind deiner Meinung nach korrekt?
  Antwortmöglichkeiten: Datenanalysen dauern sehr lange; kleine Datenmengen sind besser, da die Analyse schneller geht; aus großen Datenmengen kann man wenig herauslesen; oft ist es möglich Informationen über Personen herauszufinden, die gar nicht in den Daten stehen; es ist kaum möglich, solch große Datenmengen wie sie beispielsweise die NSA vorhält zu analysieren; die Metadaten sind oft wesentlich interessanter als die eigentlichen Daten; meine Daten dürfen ruhig analysiert werden, die finden sowieso nichts Neues heraus; ich habe nichts zu verbergen

Diese Fragen können einen gewissen Einblick in das Wissen Schülerinnen und Schüler über Datenbanken (und Datenspeicher im Allgemeinen), aber auch über Datenanalysen, geben. Die Fragen zielen teilweise auch auf aktuelle Themen im gesellschaftlichen Diskurs ab, beispielsweise indem eine Antwortmöglichkeit vorgesehen wurde, die die massive

[31] Es wurde sich dabei passend zur befragten Gruppe an den entsprechenden Lehrplanausschnitten der bayerischen Realschule und des bayerischen Gymnasiums orientiert.

Datenspeicherung durch Geheimdienste wie die NSA und den Aspekt der Analysierbarkeit solcher Datenmengen thematisiert.

Als letzter Aspekt der Untersuchung wurden die Schülerinnen und Schüler nach ihren Gewohnheiten bezüglich der Datensicherung befragt. Dabei wurden die Fragen explizit auf den Wert, den die Befragten ihren eigenen Daten beimessen, hin ausgerichtet. Die Fragen lauten daher:

- Wie schützt du deine Daten vor Verlust?
  Antwortmöglichkeiten: ich erstelle regelmäßig ein Backup auf USB-Stick oder externer Festplatte; ich synchronisiere sie in die Cloud (z. B. Dropbox); meine Daten sind nicht so wertvoll, dass ich sie schützen muss; habe ich mir noch keine Gedanken gemacht
- Typischerweise sichere ich folgende Daten:
  (Freitextantwort)

Durch diese beiden Fragen kann ein Einblick in den Wert, den die Schülerinnen und Schüler ihren verschiedenen persönlichen Daten beimessen, gewonnen werden.

Der gesamte Fragebogen ist dieser Arbeit in Anhang B beigefügt.

6.2.3 Durchführung und Auswertung

Die Fragebögen wurden an 42 bayerische Schülerinnen und Schüler in zwei Gruppen verteilt. Aufgrund der Einbettung in eine Praktikumsveranstaltung[32] bzw. einen Vortrag zum Thema Datenmanagement (vor Thematisierung der eigentlichen Inhalte), konnte eine Rücklaufquote von 100 % erreicht werden. Die beiden Gruppen waren wie folgt zusammengesetzt:

- Gruppe 1 bestand aus 20 Schülerinnen und Schülern unterschiedlicher Klassen, Schulen und Jahrgangsstufen: Vier Schüler/-innen kamen aus Realschulen (alle neunte Jahrgangsstufe). 16 Schüler-/innen kamen aus Gymnasien (1× achte, 6× neunte, 7× zehnte, 2× elfte Jahrgangsstufe).
- Gruppe 2 bestand aus 22 Schülerinnen und Schülern einer neunten Klasse eines Gymnasiums.

Während für die zweite Gruppe eine klare Aussage möglich ist, dass *Datenbanken und Datenmodellierung* bereits, wie im Lehrplan *(Staatsinstitut für Schulqualität und Bildungsforschung, 2009)* vorgesehen, thematisiert wurde, kann dies für die erste Gruppe nur abgeschätzt werden: Das Thema ist laut Lehrplan für das Bayerische Gymnasium in der neunten Klasse vorgesehen, in der Bayerischen Realschule wird es typischerweise auch in der neunten Klas-

[32] Die Befragung fand im Rahmen eines Praktikumsversuchs des Mädchen-und-Technik- bzw. Jugend-und–Technik-Praktikums an der Technischen Fakultät der Friedrich-Alexander-Universität statt, der nicht in Richtung Datenmanagement o. Ä. ausgerichtet war.

se oder vorher thematisiert[33]. Da die Befragung nach Schuljahresende in den Sommerferien stattfand, kann davon ausgegangen werden, dass ein Großteil der befragten Schülerinnen und Schüler bereits Vorkenntnisse aus dem Unterricht mitbringt.

Bei der Auswertung wurden alle Antwortmöglichkeiten auf die Fragen als Unterfragen betrachtet. Für diese wurde jeweils die Anzahl der Befragten, die diese angekreuzt haben, aufsummiert. Die Freitextfrage zur Sicherung der Daten wurde getrennt betrachtet: Hier wurden ähnliche bzw. sinngemäß identische Antworten zusammengeführt, z. B. Bilder und Fotos, aber auch Excel-, Worddateien und Dokumente. Die aggregierten Ergebnisse sind in Tabelle 6.2 dargestellt.

6.2.4 Interpretation

Metadaten im Alltag

Obwohl Metadaten kein etabliertes oder üblicherweise thematisiertes Unterrichtsthema darstellen und sich im Rahmen des zweiten Fragebogenteils auch andeutet, dass die Schülerinnen und Schüler diesen Begriff nicht klar fassen können, scheinen sie insbesondere im ersten präsentierten Kontext trotzdem einen relativ guten Eindruck von diesen zu haben (vgl. Abbildung 6.7): Nahezu alle Teilnehmer geben an, dass *Datum und Uhrzeit* bei der Erstellung eines Fotos mit dem Smartphone als Metadatum gespeichert wird. Zumindest 60 % der Befragten sind sich auch bewusst, dass der *Aufnahmeort als GPS-Koordinaten* und *Informationen über die Kamera* mit dem das Foto erzeugt wurde, üblicherweise gespeichert werden. Außerdem haben die meisten Teilnehmer korrekt angenommen, dass die *Namen der Personen* auf dem Foto oder eine *Beschreibung* desselben typischerweise nicht automatisch gespeichert werden. Die Ergebnisse dieses ersten Teils zeigen daher, dass Schülerinnen und Schüler anscheinend mit dem Themengebiet Metadaten im Alltag prinzipiell in Berührung kommen und auch ein Bewusstsein dafür aufgebaut haben, dass Metadaten mehr oder weniger unbemerkt und automatisch erzeugt und gespeichert werden können.

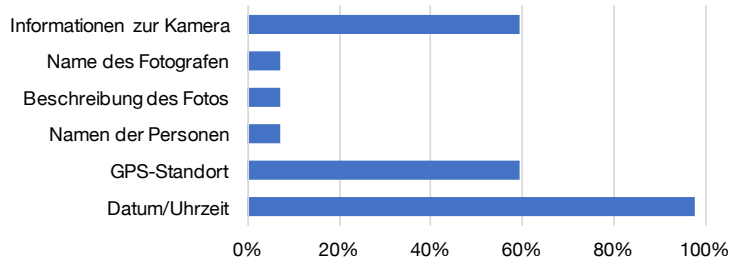


Abbildung 6.7: Übersicht über die von Schülerinnen und Schülern erwarteten Metadaten im Kontext der Erstellung eines Fotos mit dem Smartphone (prozentualer Anteil der Befragten).

[33] Aufgrund der Modulstruktur des Lehrplans für die Bayerische Realschule *(Staatsinstitut für Schulqualität und Bildungsforschung, 2008)* kann keine 100 %ige Aussage getroffen werden, wann welches Modul im Unterricht aufgegriffen wird, da diese Entscheidung den Schulen überlassen wird.

		Anzahl Antworten	Anteil von $N = 42$
Q1:	**Metadaten von Smartphone-Fotos**		
1.1	Datum/Uhrzeit	41	97,6 %
1.2	GPS-Standort	25	59,5 %
1.3	Namen der Personen am Foto	3	7,1 %
1.4	Beschreibung des Fotos	3	7,1 %
1.5	Name des Fotografen	3	7,1 %
1.6	Informationen zur Kamera	25	59,5 %
Q2:	**Metadaten beim Web-Surfen**		
2.1	Ursprungswebseite	21	50,0 %
2.2	Browsername	32	76,2 %
2.3	Betriebssystem	22	52,4 %
2.4	GPS-Standort	15	35,7 %
2.5	Name des Nutzers	7	16,7 %
2.6	Namen von Programmen	5	11,9 %
2.7	E-Mail-Adresse des Nutzers	10	23,8 %
2.8	Interessen des Nutzers	13	31,0 %
2.9	Eindeutige Nutzer-ID	8	19,0 %
2.10	Stationäres oder mobiles Gerät	26	61,9 %
2.11	Bildschirmauflösung	3	7,1 %
2.12	Sprache	22	52,4 %
2.13	Land	33	78,6 %
2.14	Alter des Nutzers	2	4,8 %
Q3:	**Datenbanken**		
3.1	Konsistente Speicherung	12	28,6 %
3.2	Nicht lohnenswert bei kleineren Datenmengen	12	28,6 %
3.3	Große Datenmengen problematisch	0	0 %
3.4	Max. fünf gleichzeitige Nutzer	1	2,4 %
3.5	Eine Datenbank nutzt genau einen Server	9	21,4 %
3.6	Ungeeignet für Fotos, Videos, ...	11	26,2 %
3.7	Cloud-Dienste basieren auf Datenbanken	25	59,5 %
Q4:	**Backup von Daten**		
4.1	Regelmäßige Backups auf z. B. USB-Sticks	31	73,8 %
4.2	Synchronisation mit der Cloud	16	38,1 %
4.3	Daten sind nicht wertvoll genug	7	16,7 %
4.4	Keine Gedanken gemacht	3	7,1 %
4.5	Gesicherte Daten		
4.5.1	Fotos	22	52,4 %
4.5.2	Dokumente	6	14,3 %
4.5.3	Videos	14	33,3 %
4.5.4	Daten mit Schulbezug	2	4,8 %
4.5.5	Speicherstände von Spielen	1	2,4 %
4.5.6	Programme	1	2,4 %
4.5.7	Programmdaten	3	7,1 %
4.5.8	Musik	4	9,5 %
4.5.9	Kontakte	4	9,5 %
Q5:	**Datenanalysen**		
5.1	Datenanalysen dauern sehr lange	6	14,3 %
5.2	Kleine Datenmengen sind zu bevorzugen, weil Analyse schneller	19	45,2 %
5.3	Aus großen Datenmengen kann man wenig herauslesen	6	14,3 %
5.4	Datenanalysen können helfen neue Informationen zu entdecken	16	38,1 %
5.5	Große Datenmengen können kaum analysiert werden	6	14,3 %
5.6	Metadaten sind oft wertvoller als die eigentlichen Daten	23	54,8 %
5.7	Über mich finden die sowieso nichts Neues heraus	2	4,8 %
5.8	Ich habe nichts zu verbergen	5	11,9 %

Tabelle 6.2: Ergebnisse des Schülerfragebogens.

Es fällt ihnen jedoch schwer, dieses Wissen auf andere bekannte Situationen zu übertragen, in denen Metadaten schwerer einseh- und erkennbar sind: Im Rahmen der zweiten Situation (vgl. Abbildung 6.8), dem Aufruf einer Webseite, nimmt zwar die klare Mehrheit der Teilnehmenden an, dass der Webserver Informationen über den Standort des Clients in Form des Herkunftslandes der Anfrage bekommt (78,6 %), und auch die Übermittlung des Browsernamens scheint ca. 77 % der Schülerinnen und Schüler bewusst zu sein. Bei anderen Daten, die real auch erhoben werden, sind sie sich aber eher unsicher: Nur 52 % nehmen an, dass die Sprache und das Betriebssystem erfasst werden. In anderen Bereichen scheint wieder mehr Sicherheit zu herrschen, dabei unterschätzen viele der Befragten jedoch den Umfang der erfassten Daten: Nur 7 % gehen davon aus, dass die Bildschirmauflösung übermittelt wird, nur 12 % vermuten dasselbe über den Namen verschiedener installierter Programme. Gleichzeitig überschätzen einige Schülerinnen und Schüler den Umfang der Daten jedoch auch: So scheint es für die Befragten nicht abwegig, dass die Interessen der Nutzer Webseiten direkt übermittelt werden (31 %), genauso eine eindeutige Benutzer-ID (19 %) oder die E-Mail-Adresse des Nutzers (24 %).

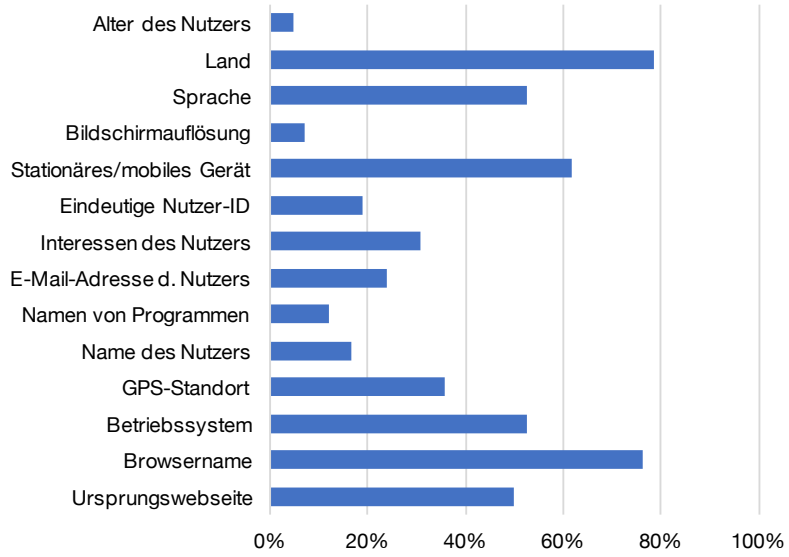


Abbildung 6.8: Übersicht über die von Schülerinnen und Schülern erwarteten Metadaten im Kontext des Besuchs einer Internetseite (prozentualer Anteil der Befragten).

Die Ergebnisse zeigen daher, dass die Schülerinnen und Schüler im Allgemeinen zwar ein Bewusstsein für Metadaten entwickelt haben, aber insbesondere die Möglichkeiten zur Erfassung solcher Daten und damit den Umfang der erfassten Metadaten kaum einschätzen können. Dies fällt insbesondere in Situationen, in denen die sie direkt mit den Metadaten in Kontakt kommen, weniger ins Gewicht, als in Situationen, in denen die Nutzung der Daten außerhalb ihres Einflussbereichs stattfindet.

Datenbanken und Datenanalysen

Hinsichtlich des Themengebiets *Datenbanken und Datenanalysen* zeigen die Ergebnisse, dass die befragten Schülerinnen und Schüler eher vage Kenntnisse haben. Obwohl die meisten von ihnen bereits Unterricht zum Thema Datenbanken besucht haben, sind kaum Unterschiede zwischen Fragen, die relativ nahe am üblicherweise thematisierten Schulstoff sind, und solchen die darüber klar hinausgehen, erkennbar. Es fällt dabei auf, dass nur 38 % der Schülerinnen und Schüler vermuten, dass Datenanalysen genutzt werden können, um *nicht-offensichtliche Informationen* aus Daten zu extrahieren, obwohl dieses Thema heute zentrale gesellschaftliche Bedeutung hat. Gleichzeitig gaben 55 % der Schüler an, dass *Metadaten oft interessanter für Datenanalysen* sind, als die Originaldaten, was bei einer Frage mit zwei Antwortmöglichkeiten darauf hindeutet, dass hier kein gefestigtes Wissen vorhanden ist und die Jugendlichen bisher auch kaum (beispielsweise in den Nachrichten) mit dem Begriff *Metadaten* in Berührung gekommen zu sein scheinen. Dies trifft genauso auf die 45 % der Teilnehmer zu, die vermuten, dass *kleine Datenmengen für Datenanalysen bevorzugt* werden sollten. Die Ergebnisse zeigen, dass an dieser Stelle, trotz vorherigen Unterrichts zum Thema Daten, das Fachwissen aus diesem Themenbereich kaum auf allgemeinere Problemstellungen übertragen werden kann und insbesondere auch gesellschaftlich relevante Aspekte von den Schülerinnen und Schülern mit deren aktuellem Wissen kaum kritisch hinterfragt bzw. bewertet werden können.

Wert der eigenen Daten

Durch die Frage zur Sicherung von Daten kann ein Einblick in den Wert, den Schülerinnen und Schüler ihren Daten beimessen, gewonnen werden. Mit 74 % gab ein Großteil der Befragten an, eigene Daten regelmäßig auf externe Medien wie USB-Sticks zu sichern, während ca. 38 % ihre Daten mit Clouddatenspeichern synchronisieren. Ein genauerer Blick auf die Ergebnisse zeigt, dass ca. 29 % der Schülerinnen und Schüler beide Methoden nutzen, während jedoch auch 17 % an, kein Backup der eigenen Daten zu erstellen. Von den Schülerinnen und Schülern, die keine Backups erstellen, gibt der Großteil an, dass die eigenen Daten nicht wertvoll genug für ein Backup sind, während drei der Befragten sich darum noch keine Gedanken gemacht haben. Die gesicherten Daten und damit vermutlich auch die wertvollsten Daten der Schüler, sind allen voran Fotos (52 % der Befragten sichern diese), Videos (33 %) und Dokumente (14 %). Weitere Daten wie Musik, Kontakte oder Anwendungen wurden nur von weniger als 10 % der Befragten genannt (vgl. Abbildung 6.9).

6.2.5 Zusammenfassung und Fazit

Auf Schülerseite zeigen sich die Auswirkungen der bisher kaum stattfindenden Thematisierung von Datenmanagementthemen im Unterricht insbesondere daran, dass ein Großteil der Befragten zwar bereits sowohl von Datenbanken und Datenanalysen als auch Metada-

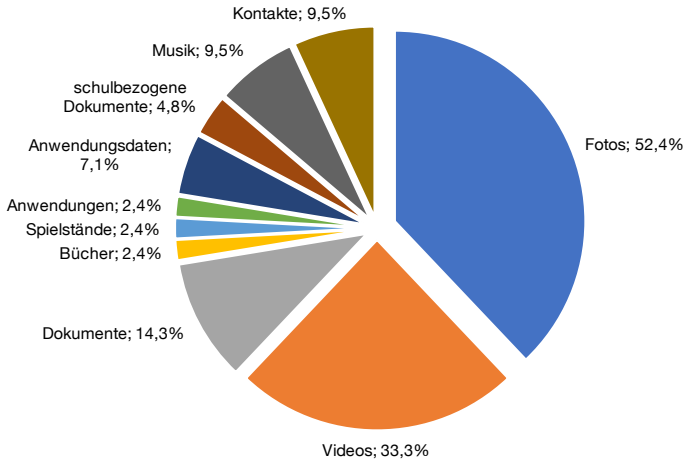


Abbildung 6.9: Übersicht über die von Schülerinnen und Schüler am meisten gesicherten Daten.

ten gehört und zu diesen ein gewisses Vorwissen und Vorstellungen erworben hat, dieses Wissen aber kaum auszureichen scheint, um Möglichkeiten und Gefahren aktueller Entwicklungen auf dem Gebiet des Datenmanagements einschätzen und bewerten zu können. Diese Befähigung zum kritischen Umgang mit gesellschaftlich relevanten Entwicklungen stellt jedoch ein wichtiges Ziel allgemeinbildenden Informatikunterrichts dar. Da die Schülerinnen und Schüler im Alltag mit verschiedenen Arten von Daten umgehen, die einen (individuell unterschiedlichen) Wert für sie haben, müssen jene relevanten Themen, zu denen die Befragten zwar grundlegendes aber kaum ausreichendes Wissen haben, im Unterricht aufgegriffen und vertieft werden, beispielsweise Metadaten. Auf diese Weise kann ein besseres Verständnis für diese erworben werden, sodass die Schülerinnen und Schüler grundlegende Kompetenzen für ein Leben in der digitalen Gesellschaft erwerben.

6.3 Zusammenfassung der Ausgangslage

Zusammenfassend zeigen die Ergebnisse der drei Untersuchungen daher, dass obwohl ein klarer Bedarf für die Vermittlung von Kompetenzen aus dem Bereich des Datenmanagements besteht, diese derzeit aus verschiedenen Gründen bisher kaum stattfinden kann. Insbesondere fehlt dazu die fachliche Fundierung des Wissens der Lehrkräfte und die Einbettung solcher Themen in den Bildungsstandards und Curricula, aber es muss auch eine Entwicklung von bisher kaum vorhandenen Werkzeugen für den Unterricht stattfinden. Gerade solche Themen, denen heute im Alltag eine immer größere Bedeutung zukommt, scheinen bisher – zumindest nach den dem Unterricht zugrunde liegenden Richtlinien – kaum in der Schule thematisiert zu werden. Es kann daher eine deutliche Lücke, nicht nur zwischen der fachwissenschaftlichen Sichtweise auf Datenmanagement und dem stattfindenden Unterricht, sondern gleichzeitig auch zwischen Unterricht und den Anforderungen des täglichen Lebens, festgestellt werden. Dadurch zeigt sich der Bedarf an einer fachdi-

daktischen Aufarbeitung des Themengebiets, da trotz dieser Lücken nicht direkt und ohne weitere Forschung davon ausgegangen werden kann, dass es sinnvoll ist, fehlende Themen direkt in den Unterricht aufzunehmen. Im Gegenteil kann und soll es nicht das Ziel des allgemeinbildenden Schulunterrichts sein, das komplette Fachgebiet detailliert auf den Unterricht abzubilden. Stattdessen muss auf den allgemeinbildenden Charakter potenzieller Unterrichtsinhalte geachtet und dieser herausgearbeitet werden. Entsprechend ist eine weitere Aufarbeitung des Themengebiets, wie sie im weiteren Verlauf dieser Arbeit geschieht, und eine weitergehende Diskussion von Datenmanagement aus informatikdidaktischer Sicht essenziell, da auf diese Weise das Datenmanagement für die Lehrerinnen und Lehrer besser greifbar wird. Durch geeignete Fortbildung von Lehrkräften und Erstellung geeigneter Materialien für diese, muss ihr Fachwissen auf- und ausgebaut und ihnen somit die Möglichkeit gegeben werden, diese immer wichtiger werdenden Themen im Informatikunterricht zu thematisieren.

Teil III:

Datenmanagement und Data Literacy aus informatikdidaktischer Sicht

7 Charakterisierung der Informatik durch Ideen, Konzepte und Prinzipien

Aufbauend auf der vorher ermittelten Ausgangslage für den Unterricht und den sich damit ergebenden Herausforderungen, kann nun die vierte Forschungsfrage in Angriff genommen werden: Die Ermittlung der zentralen Konzepte und Praktiken des Fachgebiets Datenmanagement unter Berücksichtigung der Perspektive der Sekundarschulinformatik.

In den vergangenen Jahren wurde bereits eine Vielzahl von Ansätzen entwickelt, um die Informatik (oder eines ihrer Themen bzw. Fachgebiete) durch Ideen, Konzepte oder Prinzipien zu beschreiben. Nicht alle solchen Ansätze stammen dabei aus der Informatikdidaktik, häufig sind sie auch stark fachwissenschaftlich geprägt und entstammen der jeweiligen Forschung im Fachgebiet. Für die Planung von Unterricht und Curricula spielen sie eine zentrale Rolle. *Shaw (1992)* betont die Wichtigkeit ideenbasierter Ansätze für den Unterricht: *„Let's organize our courses around ideas rather than around artifacts. This helps make the objectives of the course clear to both students and faculty. Engineering schools don't teach boiler design – they teach thermodynamics."* Auch innerhalb der Fachdidaktik Informatik besteht Einigkeit, dass Informatikunterricht sich primär auf die für das Fach zentralen und langfristig relevanten Aspekte konzentrieren sollte, wie sie durch solche Ideen, Konzepte oder Prinzipien beschrieben werden. So charakterisieren beispielsweise *Hartmann, Näf und Reichert (2006)* die fundamentalen Ideen als Konzept, *„mit dem die Bedeutsamkeit eines Themas oder Sachverhaltes überprüft werden kann. Diese Überprüfung liefert wichtige Anhaltspunkte für die Aufbereitung des Stoffes und die Unterrichtsgestaltung."*

Die Ermittlung der Ideen, Konzepte oder Prinzipien der Informatik oder eines ihrer Teilbereiche ist dabei oft durch die jeweils durch den/die Forschenden eingenommene Perspektive auf die Informatik geprägt, sodass eine gewisse Subjektivität nicht ausgeschlossen werden kann. Gleichzeitig unterliegen die jeweils entstandenen Kataloge, trotz der angestrebten langfristigen Gültigkeit, aufgrund der hohen Dynamik der Informatik einem gewissen Wandel. Es ist daher nicht verwunderlich, dass in den bisherigen Arbeiten aus diesem Bereich, die oft auf eine langjährige Tradition zurückblicken, das Fachgebiet Datenmanagement bzw. insbesondere dessen Veränderungen der letzten Jahre bisher allenfalls am Rande berücksichtigt werden. Bevor nun zur weiteren fachlichen Klärung die dem Fachgebiet zugrundeliegenden Ideen, Konzepte oder Prinzipien ermittelt werden können, werden im Folgenden zuerst diese drei Begriffe definiert und voneinander abgegrenzt. Daraufhin werden zur Fundierung der weiteren Arbeit bewährte Ansätze zur Beschreibung der Informatik oder eines ihrer Teilbereiche vorgestellt und voneinander abgegrenzt. Darauf aufbauend erfolgt die Entwicklung eines Ansatzes zur Ermittlung der Schlüsselkonzepte auf Basis der verschiedenen vorgestellten Ansätze und die Anwendung auf das Fachgebiet Datenmanagement.

7.1 Ideen, Konzepten und Prinzipien

Die drei Begriffe *Idee*, *Konzept* und *Prinzip* werden häufig im Zusammenhang mit der Charakterisierung eines Faches bzw. der Ermittlung dessen wesentlicher Inhalte verwendet. In der Informatikdidaktik wird dabei beispielsweise von den *fundamentalen Ideen der Informatik (Schwill, 1993)*, den *Great Principles of Computing (Denning, 2003b)* oder den *Zentralen Konzepten im Informatikunterricht (Zendler und Spannagel, 2006)* gesprochen. *Zendler und Spannagel (2006)* erläutern, dass der Begriff Idee schwierig zu fassen ist, sodass die Verwendung dieses Begriffs, aber auch der anderen beiden, teils unklar und in verschiedenen Arbeiten uneinheitlich oder sogar widersprüchlich ist.

Die Interpretation des Begriffs *Idee*, wie ihn *Schwill (1993)* bei den fundamentalen Ideen der Informatik verwendet, entstammt den Arbeiten von Jerome S. Bruner. Dieser baut auf der Idee im Kantschen Sinne auf, umschreibt den Begriff und dessen Verständnis selbst jedoch nur vage (für eine weitere Klärung des Begriffs im Brunerschen Sinne vgl. *Schubert und Schwill (2011)* sowie *Schwill (2004)*). Basierend auf den Arbeiten von Bruner entwickelte Schwill *(Schwill, 1993)* eine Definition des Begriffs der *fundamentalen Idee*, die auf verschiedenen Kriterien basiert, die eine Idee, d. h. ein Denk-, Handlungs-, Beschreibungs- oder Erklärungsschema, erfüllen muss, um als fundamental zu gelten. Neben diesen Kriterien für die Fundamentalität einer Idee beschreibt Schwill selbst jedoch den Begriff der Idee auch nur vage, insbesondere betont er aber ihren idealisierten Charakter: *„Ideas are certain abstract ideal imaginations of objects that are not available in reality but that act as models for human behavior or real objects and thus define objectives which humans try to achieve approximately.“ (Schwill, 2004)*

Der Begriff *Konzept* entstammt hingegen der Psychologie. Das englische Wort *concept* wird in dieser eigentlich gleichbedeutend mit dem deutschen *Begriff* verwendet, trotzdem hat sich auch im Deutschen die Bezeichnung „Konzept“ heute durchgesetzt. Ein Begriff stellt dabei eine Denkeinheit dar, die *„aus einer Menge von Gegenständen unter Ermittlung der diesen Gegenständen gemeinsamen Eigenschaften mittels Abstraktion gebildet wird“ (DIN 2342:2011-08, 2011)*. Das semiotische Dreieck (vgl. Abbildung 7.1 sowie *Ogden und Richards (1923)*) betont die Verknüpfung von *Begriff* und *Benennung* (alternativ oft *Symbol*) mit dem *Gegenstand*, dem verschiedene *Merkmale* zugeordnet werden. Konzepte entstehen im psychologischen Sinn, indem der Mensch einem Objekt bzw. Gegenstand Eigenschaften zuordnet, weswegen auch diesen ein gewisser Grad an Idealisierung zugrunde liegt. Diese ist jedoch weniger deutlich ausgeprägt als bei Ideen: Bei Konzepten wird von einem konkreten Objekt ausgegangen und dessen zentrale Aspekte analysiert und hervorgehoben. Ein Konzept beschreibt daher, *wie ein Objekt ist*, eine Idee hingegen beschreibt idealisiert, *wie ein Objekt sein soll*. Beide Begriffe konzentrieren sich dabei jedoch im Sinne einer Abstraktion auf zentrale und hervorhebenswerte Aspekte. Konzepte und Ideen können nach diesem Verständnis daher miteinander in Verbindung gesetzt werden, da sie ein und dasselbe Objekt mit unterschiedlich starken Idealisierungsgrad beschreiben können. Während die Idealisierung beim Konzept sich daher insbesondere auf eine Fokussierung auf das Wesentliche beschränkt,

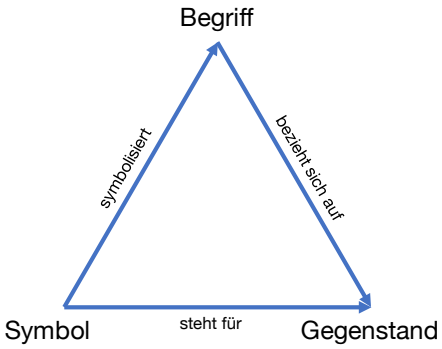


Abbildung 7.1: Das semiotische Dreieck nach *Ogden und Richards (1923)*.

liegt dem Ideenbegriff eine (meist nicht erreichbare) Zielvorstellung zugrunde, wie sie auch das Zielkriterium der fundamentalen Ideen aufgreift.

Als *Prinzipien* werden häufig solche Konzepte oder Ideen herausgestellt, denen eine besonders herausragende Bedeutung im betrachteten Bereich zukommt. Abgeleitet vom lateinischen Wortursprung wird Prinzipien oft zugeschrieben, dass sie zentrale Grundlagen im betrachteten Umfeld sind. Dies ist beispielsweise in den Naturwissenschaften der Fall, in denen Prinzipien die im jeweiligen Fachgebiet als allgemeingültig erachteten bzw. bewiesenen Regeln darstellen, die oft Basis für weitere Annahmen, Axiome oder Gesetze sind, und die als unwiderlegbar gelten. Eine andere Verwendung des Begriffs wird beispielsweise bei Denning deutlich: *„What we call principles are almost always distilled from recurrent patterns observed in practice.“ (Denning, 2003b)* Im Gegensatz zu anderen Quellen verwendet Denning den Begriff, ohne einen speziellen Fokus auf die Zentralität und Unwiderlegbarkeit zu richten. Stattdessen legt er als wichtigstes Kriterium das regelmäßige Auftreten in der Wissenschaft zugrunde.

Zusammenfassend können alle drei Begriffe, *Idee*, *Konzept* und *Prinzip*, als eng verwandt charakterisiert werden: Sie beschreiben jeweils ein Objekt oder einen Teil davon, indem zentrale Eigenschaften und gegebenenfalls dessen Funktionsweise herausgestellt werden. Sie unterscheiden sich aber in den Anforderungen hinsichtlich der Bedeutung im Fachgebiet sowie dem Grad an Idealisierung (vgl. Abbildung 7.2). Da im Folgenden das Ziel die Charakterisierung des Fachgebiets *Datenmanagement* ist, wurde der Fokus auf die Suche nach den *Konzepten* des Datenmanagements gelegt: Diese beschreiben, was Datenmanagement bzw. dessen Themen sind, ohne eine starke Idealisierung zugrunde zu legen. Üblicherweise kann innerhalb eines Fachgebiets eine große Anzahl an Konzepten gefunden werden, deren Bedeutung unterschiedlich hoch ist. Somit muss hier eine Auswahl stattfinden, um diese auf eine erfassbare und greifbare Menge zu reduzieren. In dieser Arbeit wird der Fokus daher auf die *Schlüsselkonzepte* des Datenmanagements gelegt: Unter einem Schlüsselkonzept wird im Folgenden ein Konzept verstanden, das einerseits einen zentralen Aspekt aus dem Fachgebiet beschreibt und damit zu dessen Strukturierung und Charakterisierung beiträgt, andererseits aber auch als Schlüssel zu diesem dienen kann, indem es zu einem grundlegenden Verständnis beiträgt und somit einen Zugang zum Fachgebiet eröff-

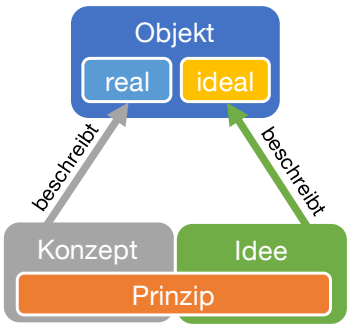


Abbildung 7.2: Beziehung zwischen den Begriffen *Konzept*, *Idee*, *Prinzip* und *Objekt*.

net. Gleichzeitig wird jedoch nicht zwingend vorausgesetzt, dass den Schlüsselkonzepten eine gleichermaßen zentrale Bedeutung im Fachgebiet zukommt, wie es bei Prinzipien der Fall ist. Es ist aber davon auszugehen, dass verschiedene der im Folgenden ermittelten Schlüsselkonzepte prototypisch für Prinzipien des Datenmanagements stehen.

7.2 Bisherige Arbeiten in Informatik und Informatikdidaktik

Die Strukturierung eines Fachgebiets durch Ermittlung der zugrundeliegenden Ideen, Konzepte oder Prinzipien ist insbesondere in den naturwissenschaftlichen Fächern und in der Informatik ein anerkannter und schon seit langem genutzter Ansatz. Dabei werden verschiedene Ziele verfolgt: Oft wird angestrebt, eine Wissenschaftsdisziplin oder Teile davon zu charakterisieren, um diese für nicht-Experten verständlicher und nachvollziehbarer zu machen oder um diese für den Schulunterricht aufzubereiten. Als Basis für die weitere Arbeit in diesem Bereich werden daher im Folgenden durch ihre Bedeutung oder ihren methodischen Ansatz hervorstechende Arbeiten auf diesem Gebiet beschrieben, zueinander kontrastiert und auf dieser Basis Leitlinien für die im nächsten Kapitel durchgeführte Ermittlung der Schlüsselkonzepte des Datenmanagements gezogen.

7.2.1 Fundamentale Ideen der Informatik, der Theoretischen Informatik und der Schulinformatik

Der insbesondere in der deutschen Informatikdidaktik bekannteste Ansatz zur Charakterisierung der Informatik wurde von *Schwill (1993)* vorgestellt. Er sieht die Softwareentwicklung als eine der zentralen Aufgaben der Informatik und hat daher den Softwareentwicklungsprozess als Ausgangspunkt für die Ermittlung der *fundamentalen Ideen der Informatik* gewählt. Dazu hat er, basierend insbesondere auf den Arbeiten von Jerome S. Bruner, vier bzw. später fünf Kriterien für fundamentale Ideen der Informatik definiert.

Eine fundamentale Idee der Informatik muss. . . *(vgl. Schwill, 1993; Schwill, 1998)*

- *„in verschiedenen Bereichen der Informatik vielfältig anwendbar oder erkennbar sein“*: Das *Horizontalkriterium* sortiert Begriffe bzw. Themen aus, die nur in speziellen Bereichen der Informatik auftreten und daher nicht als fundamental für die Informatik im Ganzen gelten können.
- *„auf jedem intellektuellen Niveau aufgezeigt und vermittelt werden können“*: Das *Vertikalkriterium* stellt sicher, dass fundamentale Ideen schon auf niedrigem intellektuellem Niveau verstanden werden können, gleichzeitig aber auch ausbaufähig sind und damit beispielsweise auch im Sinne eines Spiralcurriculums thematisiert werden können.
- *„in der historischen Entwicklung des Bereichs deutlich wahrnehmbar sein und längerfristig relevant bleiben“*: Das *Zeitkriterium* verhindert, dass Ideen als fundamental angenommen werden, die nur kurzfristige Relevanz haben und in Kürze wieder obsolet sind. Es stellt somit eine gewisse Zeitbeständigkeit des Ideenkatalogs sicher.
- *„einen Bezug zu Sprache und Denken des Alltags und der Lebenswelt besitzen“*: Das *Sinnkriterium* stellt einen Bezug zur Alltagswirklichkeit der Lernenden her. Es stellt sicher, dass nur Aspekte als fundamental angenommen werden, die auch praktische Relevanz haben und trägt damit zu den Zielen einer Allgemeinbildung bei.
- *„zur Annäherung an eine gewisse idealisierte Zielvorstellung dienen, die jedoch faktisch möglicherweise unerreichbar ist“*: Das im ursprünglichen Kriterienkatalog nicht enthaltene *Zielkriterium* stellt sicher, dass eine Idee nicht ohne konkretes Ziel, dass mit ihrer Hilfe erreicht werden soll, als fundamental angenommen wird. Es verdeutlicht auch den idealisierten Charakter einer Idee, indem das Ziel zwar vorstellbar, aber nicht unbedingt real erreichbar sein muss.

In einer weiteren Explizierung des Begriffs der fundamentalen Idee verdeutlicht *Schwill (1998)*, dass die Erfüllung des Sinn- und des Zielkriteriums insbesondere aus dem Ideencharakter folgt, während die Fundamentalität der Ideen das Horizontal-, Vertikal- und Zeitkriterium absichern (vgl. Abbildung 7.3). Es können jedoch auch Einflüsse des Ideencharakters auf das Vertikalkriterium und der Fundamentalität auf das Sinnkriterium festgestellt werden, sodass Fundamentalität und Ideencharakter nicht klar getrennt betrachtet werden können.

Basierend auf diesen Kriterien und der Untersuchung des Softwareentwicklungsprozesses hat Schwill einen Katalog von 63 fundamentalen Ideen der Informatik entwickelt, die er den drei Masterideen *Algorithmisierung*, *Sprache* und *Strukturierte Zerlegung* hierarchisch unterordnet (vgl. Abbildungen 7.4 bis 7.6).

Ausgehend vom offensichtlichen Herausfallen der Masteridee *Sprache* aus dem Schema und von dem relativ geringen Anteil an Aspekten der theoretischen Informatik im Ideenkatalog, wurde Schwills Ansatz von *Modrow (2003)* aufgegriffen und durch stärkere Einbeziehung der theoretischen Informatik der Ideenkatalog weiterentwickelt, nicht aber die diesem zugrundeliegenden Kriterien. Durch diese Überarbeitung wurde die Masteridee *Formalisierung* (vgl. Abbildung 7.7) eingeführt, welche die ursprüngliche Masteridee *Sprache* ersetzt.

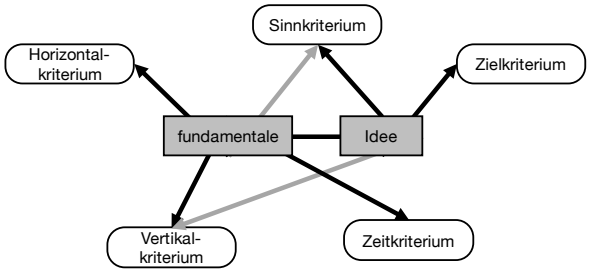


Abbildung 7.3: Bezug der Kriterien für fundamentale Ideen zum Ideencharakter und deren Fundamentalität nach *Schwill (1998)*.

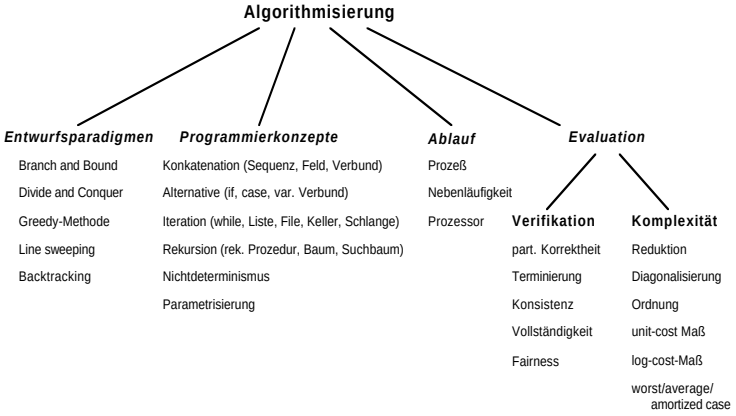


Abbildung 7.4: Masteridee *Algorithmisierung* nach *Schwill (1993)*.

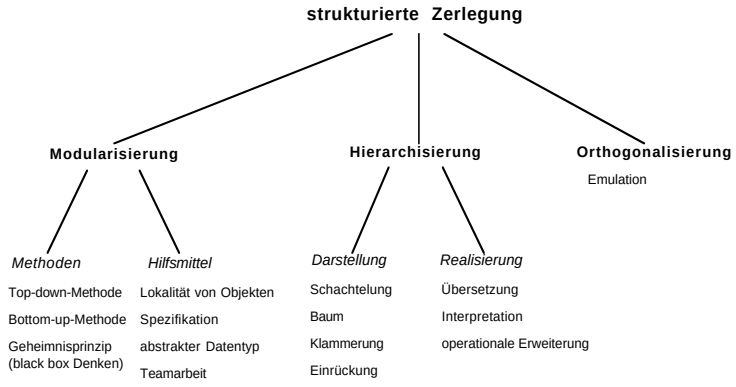


Abbildung 7.5: Masteridee *Strukturierte Zerlegung* nach *Schwill (1993)*.

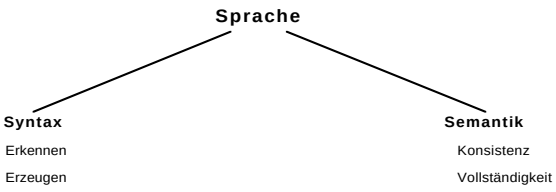


Abbildung 7.6: Masteridee *Sprache* nach *Schwill (1993)*.

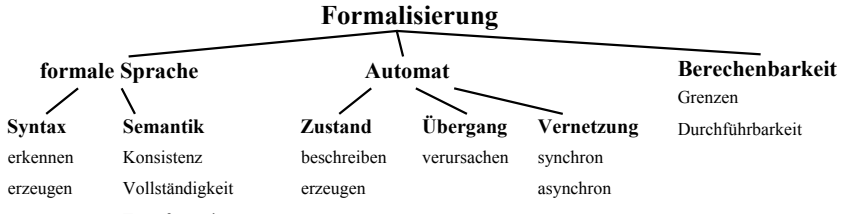


Abbildung 7.7: Masteridee *Formalisierung* nach *Modrow (2003)*.

Einige Jahre später haben *Modrow und Strecker (2016)* den Begriff der fundamentalen Ideen erneut herangezogen und hinsichtlich verschiedener Kritikpunkte (beispielsweise der hohen Anzahl der Einzelideen und der Verkürzung der Informatik auf den Softwareentwicklungsprozess) neu gedacht. Ausgehend von der Sichtweise auf die Informatik, dass diese das Ziel hat, *„reale oder virtuelle Arbeitsabläufe an einzelne Automaten oder Automatennetze zu übertragen, also Menschen Arbeit abzunehmen oder Arbeiten durchführen zu lassen, die diese nicht ausführen können oder wollen“ (Modrow und Strecker, 2016)*, wurden sechs *fundamentale Ideen der Schulinformatik* ermittelt: *Modellierbarkeit, Vernetzbarkeit, Kontextualisierbarkeit, Algorithmisierbarkeit, Digitalisierbarkeit* und *Realisierbarkeit*. Dabei wurde, im Gegensatz zu den ursprünglichen Arbeiten von Schwill und von Modrow, insbesondere die schulische Sichtweise auf die Informatik betont, ohne das Fach an sich vollumfänglich beschreiben zu wollen.

7.2.2 Great Principles of Computing

Einen anderen Ansatz zur Charakterisierung der Informatik hat Denning gewählt. Er ging der Fragestellung nach, ob die Informatik eine Naturwissenschaft darstellt und charakterisierte diese als solche *(Denning, 2005)*. Entsprechend handelt es sich hierbei um eine stark fachwissenschaftlich orientierte Charakterisierung. Denning hebt in seinen Arbeiten insbesondere den andauernd stattfindenden starken Wandel der Informatik hervor: Während in den 1950er Jahren sechs Kerntechnologien, d. h. in der Informatik besonders zentrale und in verschiedenen Kontexten relevante Technologien, ausgereicht haben, um die Informatik zu beschreiben, ist diese Zahl 1989 auf neun und 2003 bereits auf 30 angewachsen *(vgl. Denning, 2003b)*. Basierend auf diesen Kerntechnologien hat Denning einen Ansatz zur Beschreibung ihrer zentralen Prinzipien entwickelt, mit dem Ziel die Informatik im Gesamten durch diese Prinzipien beschreiben zu können (im Gegensatz zu den von ihm herangezogenen bisherigen Werken, die sich auf verschiedene Teilbereiche der Informatik konzentrieren). Im Modell der *Great Principles of Computing (Denning, 2003b)* betrachtet er diese *Kerntechnologien* zusammen mit den *Praktiken*, *Entwurfsprinzipien* und *Mechanismen* der Informatik (vgl. Abbildung 7.8). Denning selbst definiert diese drei Kategorien nicht genauer, sondern beschreibt sie eher exemplarisch. Aus seinen Beschreibungen abgeleitet, können sie jedoch wie folgt verstanden werden: Die Praktiken spiegeln den Umgang mit Informatiksystemen wider und bezeichnen häufige Tätigkeiten von Informatikerinnen und Informatikern, z. B. Programmierung und Modellierung. Die Entwurfsprinzipien

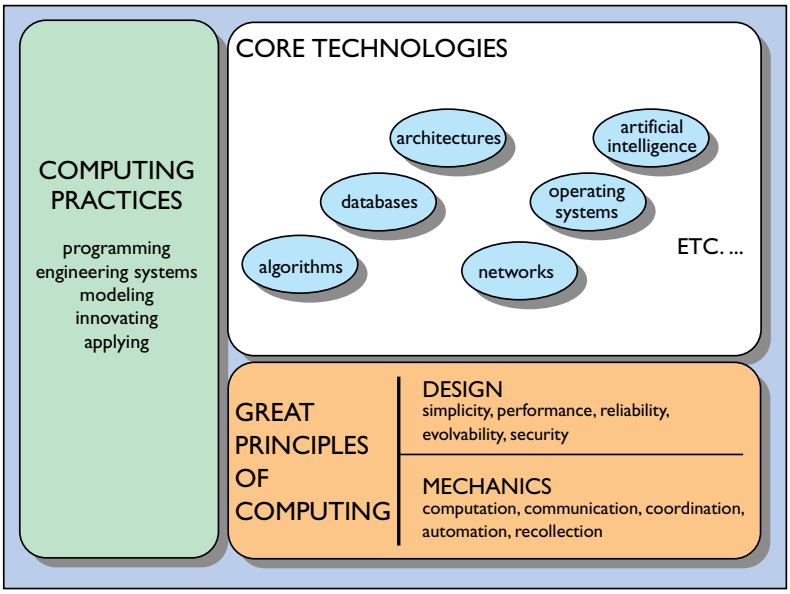


Abbildung 7.8: Modell der Great Principles of Computing nach *Denning (2003b)*.

repräsentieren Ziele, die bei Entwurf und Entwicklung von Informatiksystemen üblicherweise verfolgt oder zumindest beachtet werden, wie beispielsweise die Einfachheit oder Zuverlässigkeit. Die Mechanismen betonen, orientiert am Mechanikbegriff der Physik, das Zusammenwirken verschiedener Komponenten eines Systems zur Erreichung eines gemeinsamen Ziels, beispielsweise durch Kommunikation oder Koordination.

7.2.3 Konzepte und Prozesse der Informatik

Zendler und Spannagel (2006 bzw. 2007) wählten zur Ermittlung der *Konzepte* und *Prozesse* der Informatik einen Ansatz, der sich methodisch deutlich von den meisten anderen in diesem Bereich unterscheidet: Durch eine empirische Herangehensweise basierend auf der Befragung von Experten mithilfe von Fragebogenstudien wurde insbesondere ein höherer Grad an Objektivität der ermittelten Konzepte bzw. Prozesse angestrebt. Dazu wurden zwei Untersuchungen durchgeführt: In einer ersten Fragebogenstudie stellten *Zendler und Spannagel (2006)* fachlichen Experten aus der Informatik eine Liste von 49 potenziellen Konzepten vor, die aus dem *ACM Computing Classification System* extrahiert wurde. Diese wurden von den ausgewählten Experten hinsichtlich der Erfüllung der vier ursprünglichen Kriterien für fundamentale Ideen der Informatik nach Schwill, die vorher knapp eingeführt wurden, eingeschätzt. In einem zweiten Schritt wurde eine analog angelegte Untersuchung mit den 44 allgemeinbildenden Prozessen nach *Costa und Liebmann (1996)* durchgeführt. Durch eine statistische Auswertung und Bildung von Clustern konnten Zendler und Spannagel 15 für die Informatik zentrale Konzepte sowie sechs zentrale Prozesse ermitteln (vgl. Abbildung 7.9).

central concepts	problem, data, computer, test, algorithm, process, system, information, language, communication, software, program, computation, structure, model
central processes	problem solving and problem posing, classifying, finding relationships, investigating, analyzing, generalizing

Abbildung 7.9: Zentrale Konzepte und Prozesse der Informatik nach Zendler und Spannagel (2006/2007).

7.2.4 Big Ideas of K–12 Computer Science Education

Während die bisher präsentierten Ansätze mit Ausnahme der fundamentalen Ideen der Schulinformatik nach *Modrow und Strecker (2016)* aus einer eher fachlich geprägten Perspektive entstanden sind, befassen sich *Bell, Tymann und Yehudai (2011)* aus dem Blickwinkel der informatischen Bildung mit den *Big Ideas of K-12 Computer Science Education*. Orientiert am Konzept der *Big Ideas in Science* versuchten sie die für die Schulinformatik zentralen Ideen zu ermitteln. Basierend auf Rückmeldungen verschiedener Experten weltweit, wurden so zehn für die Schulinformatik zentrale *Big Ideas* identifiziert:

- *Information is represented in digital form.*
- *Algorithms interact with data to solve computational problems.*
- *The performance of algorithms can be modelled and evaluated.*
- *Some computational problems cannot be solved by algorithms.*
- *Programs express algorithms and data in a form that can be implemented on a computer.*
- *Digital systems are designed by humans to serve human needs.*
- *Digital systems create virtual representations of natural and artificial phenomena.*
- *Protecting data and system resources is critical in digital systems.*
- *Time dependent operations in digital systems must be coordinated.*
- *Digital systems communicate with each other using protocols.*

In der zugehörigen Beschreibung der Ideen *(vgl. Bell, Tymann und Yehudai, 2011)* wird auf verschiedene Beispiele eingegangen, die Bedeutung der Ideen erläutert und eine Reihe von Beispielen gegeben, wie diese Ideen sich in der Informatik manifestieren. Durch ihren Fokus auf die allgemeinbildende Schulinformatik beschreiben die Big Ideas die Informatik aus einem deutlich anderen Blickwinkel als andere Ansätze und können somit direkter für den Informatikunterricht verwendet und gegebenenfalls den Schülerinnen und Schülern auch direkt als Fazit einer Unterrichtssequenz an die Hand gegeben werden. Gleichzeitig zeigt sich jedoch, insbesondere durch ihre in Vergleich mit den anderen Ansätzen wesent-

lich oberflächlichere und offenere Betrachtung, auch ein höherer Interpretationsspielraum und entsprechend eine geringere Präzision als bei anderen Ansätzen.

7.2.5 Quarks of Object-Oriented Development

Während die bisher präsentierten Ansätze grundsätzlich die Informatik im Ganzen betrachten, findet bei den *Quarks of Object-Oriented Development* nach *Armstrong (2006)* eine Fokussierung auf einen ihrer Teilbereiche statt – die objektorientierte Entwicklung. Das Ziel dieses Ansatzes ist die Ermittlung zentraler Konzepte dieses Teilgebiets und dessen Beschreibung durch ein geeignetes Schema. Dazu wurden in einem empirischen Ansatz verschiedene Publikationen zur objektorientierten Entwicklung ausgewählt und untersucht, aus denen die darin thematisierten Konzepte extrahiert wurden. Auf quantitative Weise wurden die acht zentralsten dieser Konzepte, die daraufhin als *Quarks* bezeichnet wurden, bestimmt und in ein Schema eingeordnet, das die objektorientierte Programmierung beschreibt (vgl. Abbildung 7.10).

Construct	Concept	Definition
Structure	Abstraction	Creating classes to simplify aspects of reality using distinctions inherent to the problem.
	Class	A description of the organization and actions shared by one or more similar objects.
	Encapsulation	Designing classes and objects to restrict access to the data and behavior by defining a limited set of messages that an object can receive.
	Inheritance	The data and behavior of one class is included in or used as the basis for another class.
	Object	An individual, identifiable item, either real or abstract, which contains data about itself and the descriptions of its manipulations of the data.
Behavior	Message Passing	An object sends data to another object or asks another object to invoke a method.
	Method	A way to access, set, or manipulate an object's information.
	Polymorphism	Different classes may respond to the same message and each implement it appropriately.

Abbildung 7.10: Quarks of Object-Oriented Development nach *Armstrong (2006)*.

7.3 Kontrastierung und Diskussion der Ansätze

Wie schon anhand der kurzen Vorstellung dieser ausgewählten Ansätze erkennbar ist, weisen sie verschiedene Gemeinsamkeiten auf. Insbesondere ist klar erkennbar, dass alle Ansätze ein ähnliches Ziel verfolgen: Die Charakterisierung der Informatik oder eines ihrer

Teilgebiete durch eine Liste bzw. ein Modell von Konzepten, Ideen oder Prinzipien (zum Teil optimiert für eine bestimmte Nutzung). Trotzdem unterscheiden sich die verschiedenen Ansätze teils deutlich, insbesondere hinsichtlich ihrer eingenommenen Perspektive, aber auch in der genutzten Methodik. Während ein Großteil der Ansätze auf einer theoretisch-argumentativen Herleitung basiert, ist nur ein kleiner Teil empirisch geprägt. Den meisten Arbeiten ist es daher gemein, dass sie stark subjektiv geprägt und nur eingeschränkt reliabel und valide erscheinen. Trotzdem haben sich insbesondere die *fundamentalen Ideen der Informatik* und die *Great Principles of Computing* (trotz zum Teil vorhandener kritischer Stimmen) als sinnvolle und praxistaugliche Charakterisierungen erwiesen und eine hohe Bekanntheit und Relevanz nicht nur im Fach, sondern auch in der Fachdidaktik und Schulinformatik erlangt. Die zentralen Unterschiede und Gemeinsamkeiten der vorgestellten Ansätze, die diese bzw. deren Unterschiede charakterisieren, sind in Tabelle 7.1 zusammengefasst.

Obwohl es neben den hier präsentierten Ansätzen noch viele weitere gibt, konnte keiner gefunden werden, der das Fachgebiet Datenmanagement oder Teile davon detaillierter betrachtet. Im Gegenteil wird dieses Fachgebiet typischerweise höchstens als Randaspekt miterfasst: Beispielsweise ist mit dem Mechanismus *recollection* in den Great Principles of Computing nur ein Begriff genannt, der Teile dieses Fachgebiets mehr als nur anschneidet. Dieser deckt laut dessen Beschreibung Aspekte wie *Hierarchisierung*, *Persistenz* oder *Sharing* (im Sinne der Weitergabe von Daten) ab, die klar dem Datenmanagement zuzuordnen sind. Andere Bereiche des Fachgebiets scheinen aber weiterhin unterrepräsentiert, auch in anderen Ansätzen, die meist einen Schwerpunkt in der Softwareentwicklung haben.

Basierend auf den beschriebenen methodischen Ansätzen, wird daher im Folgenden eine Methodik zur Ermittlung der Schlüsselkonzepte des Datenmanagements entwickelt, die versucht, die Vorteile der verschiedenen Methoden miteinzubeziehen und gleichzeitig mögliche Kritikpunkte an diesen zu vermeiden. Dazu soll insbesondere eine möglichst hohe Objektivität erreicht werden, indem die Schlüsselkonzepte nicht rein auf subjektiven Überlegungen basierend identifiziert werden. Durch Beschreibung eines nachvollziehbaren Prozesses mit kriteriengeleiteten Entscheidungen wird außerdem versucht, eine möglichst hohe Nachvollziehbarkeit zu erreichen. Basierend auf diesen Überlegungen und den beschriebenen Arbeiten wurden daher folgende Leitlinie für die Analyse festgelegt:

Wahl eines empirischen Ansatzes. Während in anderen Teilen der Informatik, beispielsweise der Softwareentwicklung, seit Jahren ein relativ hoher Konsens über die zentralen Aspekte herrscht, scheint dies bisher im Bereich des Datenmanagements nur in wesentlich geringerem Maße der Fall zu sein. Eine theoretische Herleitung der Schlüsselkonzepte aus existierenden Arbeiten ist daher aufgrund einer fehlenden oder schwachen Basis nur unzureichend möglich und würde zu einer hohen Varianz der Ergebnisse in Abhängigkeit von der als Ausgangspunkt betrachteten Charakterisierung des Fachgebiets führen. Hingegen wurde, beispielsweise durch *Zendler und Spannagel (2006)* und durch *Armstrong (2006)*, bereits gezeigt, dass ein empirischer Ansatz in diesem Bereich durchaus erfolgversprechend sein kann und eine hohe Nachvollziehbarkeit sowie eine gewisse Objektivität der Ergebnisse sicherstellt. Daher wurde für die angestrebte Analyse ein empirischer Ansatz

gewählt, der eine systematische Identifizierung der Schlüsselkonzepte ermöglichen und somit durch Einbeziehung einer Vielzahl an Perspektiven auf das Fachgebiet auch eine höhere Objektivität sicherstellen soll.

Trennung der Analyse in zwei Phasen. Um das Analyseziel zu erreichen, reicht jedoch ein Ansatz wie von Zendler und Spannagel nicht aus: Diese konnten zwar eine Auswahl an Konzepten ermitteln, in dieser Arbeit ist aber zusätzlich angedacht, über diese hinauszugehen und ein verständliches und übersichtliches Modell der Schlüsselkonzepte des Datenmanagements zu entwickeln. Entsprechend wird die Analyse in zwei Phasen aufgeteilt, von denen sich eine mit der explorativen Erforschung des Fachgebiets *(wie bei Armstrong, 2006)*, die andere mit der Ermittlung einer geeigneten Modellstruktur und der Einordnung der gefundenen Schlüsselkonzepte in diese (ähnlich zu den *Great Principles of Computing Denning (2004)*) beschäftigt.

Entscheidung für einen Aufbau des Modells erst nach der explorativen Phase. Statt den Aufbau des angestrebten Modells direkt zu Beginn der Untersuchung festzulegen, wird diese Entscheidung erst nach der explorativen Phase getroffen. Auf diese Weise kann sichergestellt werden, dass das entwickelte Modell den Charakter des Fachgebiets möglichst gut widerspiegeln und alle zentralen Elemente beinhalten kann.

Nutzung sowohl einer fachlichen als auch einer didaktischen Perspektive. Während im Rahmen der Exploration des Fachgebiets eindeutig eine fachliche Perspektive im Vordergrund steht, wird im Rahmen der zweiten Analysephase, bei der Entwicklung des Modells, auch eine didaktische Perspektive eingenommen. Diese wird insbesondere bei der Entscheidung für den Aufbau des Modells, aber auch bei der durch das Modell berücksichtigten Breite und Tiefe des Fachgebiets eine große Rolle spielen.

Ansatz	Betrachteter Bereich	Perspektive	Zentrale Aspekte	Methodischer Ansatz
Fundamentale Ideen der Informatik *(Schwill, 1993)*	Informatik im Gesamten; Fokus auf Softwareentwicklungsprozess	fachlich	Schema der fundamentalen Ideen; Kriterien; Ideenkatalog	theoretisch-argumentativ
Fundamentale Ideen der theoretischen Informatik *(Modrow, 2003)*	Informatik im Gesamten; Fokus auf Theoretische Informatik	fachlich	Erweiterung des Ideenkatalogs nach Schwill	theoretisch-argumentativ
Fundamentale Ideen der Schulinformatik *(Modrow und Strecker, 2016)*	Schulinformatik	didaktisch	Identifizierung einer überschaubaren Menge an fundamentalen Ideen der Schulinformatik	theoretisch-argumentativ
Great Principles of Computing *(Denning, 2003b)*	Informatik im Gesamten	fachlich	Modell der Great Principles; Einbeziehung von Praktiken und Kerntechnologien	theoretisch-argumentativ
Konzepte und Prozesse der Informatik *(Zendler und Spannagel, 2006; Zendler, Spannagel und Klaudt, 2007)*	Informatik im Gesamten	fachlich	Differenzierung in Konzepte und Prozesse	fragenbogenbasierte empirische Ermittlung
Big Ideas of K-12 Computer Science Education *(Bell, Tymann und Yehudai, 2011)*	Informatik im Gesamten	didaktisch	Beschreibung von im Kontext informatischer Bildung relevanten Ideen; relativ offener Ansatz	theoretisch-argumentativ
Quarks of Object-Oriented Programming *(Armstrong, 2006)*	objektorientierte Entwicklung	fachlich	Beschreibung nur eines Fachgebiets	literaturbasierte empirische Ermittlung

Tabelle 7.1: Überblick über verschiedene Ansätze zur Charakterisierung der Informatik oder ihrer Teilbereiche.

8 Schlüsselkonzepte des Datenmanagements

Um die *Schlüsselkonzepte des Datenmanagements* zu ermitteln und ein Modell zu entwickeln, das einen übersichtlichen und verständlichen Überblick über diese Schlüsselkonzepte ermöglicht und somit das Fachgebiet charakterisiert, muss dieses unter Beachtung der Entwicklungen der letzten Jahre aufgearbeitet werden. Dazu werden zentrale Inhalte und Themen auf die jeweils zugrundeliegenden Konzepte zurückgeführt und diese insbesondere hinsichtlich ihrer zeitlichen Stabilität und ihrer Bedeutung im Fachgebiet bewertet. Im Folgenden wird zuerst der gewählte und in dieser Arbeit entwickelte methodische Ansatz zur Ermittlung dieser Konzepte und zur Strukturierung in Form des angestrebten Modells beschrieben, bevor er dann direkt auf das Fachgebiet Datenmanagement angewendet wird.

8.1 Beschreibung der Methodik und Analyse der Schlüsselkonzepte

Wie in den Vorüberlegungen bereits festgelegt und erläutert wurde, wurde die Ermittlung der Schlüsselkonzepte in zwei Phasen aufgeteilt:

- **Phase 1 – Explorative Analyse des Fachgebiets:** Die erste Phase umfasst die explorative Erfassung potenzieller Schlüsselkonzepte des Datenmanagements, indem basierend auf einer Analyse zentraler fachwissenschaftlicher Literatur ein umfassender Überblick über die Themen des Datenmanagements gewonnen wird und potenzielle Schlüsselkonzepte extrahiert werden.
- **Phase 2 – Ermittlung und Strukturierung der Schlüsselkonzepte:** In der zweiten Phase wird die als eher umfangreich erwartete Menge an ermittelten Begriffen zu einem übersichtlichen Modell des Datenmanagements hin kondensiert. Dazu werden diese kriterienbasiert in das Modell eingeordnet, dessen Aufbau zuvor im Rahmen dieser Phase definiert wird. Durch diese Einordnung werden die ermittelten Schlüsselkonzepte und damit auch das Fachgebiet Datenmanagement strukturiert.

8.1.1 Phase 1: Explorative Analyse des Fachgebiets

Um einen umfassenden Überblick über die Themen des Fachgebiets zu gewinnen, konnte nicht nur eine relativ enge Perspektive auf das Feld berücksichtigt werden, sondern es müssen verschiedenste Strömungen einbezogen werden. Entsprechend reicht es an dieser Stelle nicht aus, die einzige bestehende Arbeit zur umfangreichen Charakterisierung von Datenmanagement als Basis zu nutzen: Aufgrund seiner Zielsetzung beschreibt der Data Management Body of Knowledge (DAMA-DMBoK, *DAMA International (2010)*) das Fachgebiet zwar ausführlich, aber insbesondere aus professioneller Sicht und ohne me-

thodische Absicherung der Vollständigkeit (beispielsweise durch Wahl und Überprüfung einer repräsentativen Stichprobe an Experten). Trotzdem wurde diese Charakterisierung, neben anderen Arbeiten aus dem Fachgebiet, mit in die Analyse einbezogen, da sie zentrale Themen und Ideen des Fachgebiets klar hervorhebt.

Anstatt wie *Zendler und Spannagel (2006) und Zendler, Spannagel und Klaudt (2007)* Experten zu befragen, wurde die Exploration des Fachgebiets, wie beispielsweise auch bei *Armstrong (2006)*, in Form einer Literaturstudie bzw. Dokumentenanalyse durchgeführt. Neben den üblichen Methoden zur Datengewinnung, wie der direkten Beobachtung oder der gezielten Befragung, stellt die Dokumentenanalyse in den empirischen Sozialwissenschaften „*eine weitere eigenständige Verfahrensgruppe dar, um empirische Daten zu gewinnen und auszuwerten. Dabei wird bei einer genuinen Dokumentenanalyse auf bereits vorhandene bzw. vorgefundene Dokumente (‚extant documents') zurückgegriffen, die völlig unabhängig vom Forschungsprozess produziert wurden [...]*" *(Döring und Bortz, 2016)*. Eine Adaption dieser Methodik ist an dieser Stelle sinnvoll und möglich, da hier, wie bei der Beobachtung *menschlichen Erlebens und Verhaltens (Döring und Bortz, 2016)*, eine externe Beobachterposition eingenommen und bereits vorliegende Dokumente analysiert werden. Während Armstrong hauptsächlich auf eine Auswahl von Konferenzpapieren und Zeitschriftenartikeln zur objektorientierten Entwicklung zurückgreift und sich dabei auf solche beschränkt, die (implizit oder explizit) Konzepte der objektorientierten Entwicklung thematisieren und identifizieren, wurde im Rahmen dieser Arbeit ein anderer Ansatz gewählt: Da das Fachgebiet Datenmanagement im Verhältnis zur objektorientierten Entwicklung auf eine wesentlich kürzere Historie zurückblickt, wurden auch dessen Konzepte bisher weniger thematisiert und es besteht eingeschränkterer Konsens über diese, als bei der objektorientierten Entwicklung. Daher konnte als Basis für diese Analyse nicht auf derlei Material zurückgegriffen werden. Stattdessen wurden diverse Lehrbücher aus dem Umfeld analysiert, die die Breite des Fachgebiets darstellen und verschiedene Bereiche mehr oder weniger detailliert und mit verschiedenen Fokussen aufgreifen. Die Literaturanalyse wurde methodisch an der qualitativen Inhaltsanalyse nach *Mayring (2010)* orientiert, wobei aufgrund der speziellen Zielsetzung und der Trennung der beiden Analysephasen verschiedene im Sinne der Methodik zulässige Anpassungen vorgenommen wurden. Aufgrund des Analyseziels, ein zur Charakterisierung des Fachgebiets dienliches Kategoriensystem zu erstellen, das einen umfangreichen Überblick über Datenmanagement gibt, konnte nur ein induktiver Ansatz gewählt werden: Für einen (teilweise) deduktiven Ansatz müssten geeignete Vorarbeiten existieren, die als Basis für ein Kategoriensystem dienen könnten, was im Datenmanagement jedoch nicht der Fall war. Die durchgeführte Analyse entsprach daher methodisch einer induktiven zusammenfassenden qualitativen Inhaltsanalyse. Diese wurde in folgende Phasen unterteilt (vgl. auch Abbildung 8.1):

1. Auswahl der Literatur
2. Festlegung der
    - Kodiereinheit
    - Auswahlkriterien

3. Durchführung der Analyse: Entwicklung des Kategoriensystems
4. Clustering der Ergebnisse[34]
5. Überprüfung der Vollständigkeit[35]

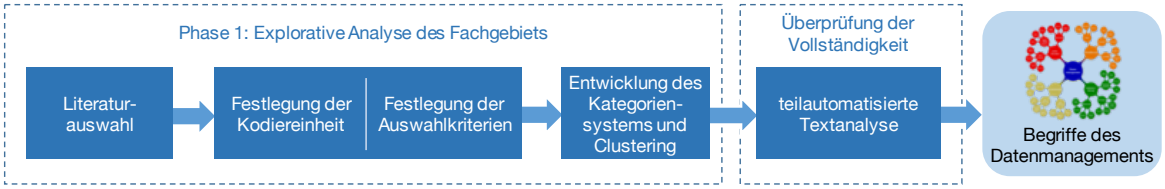


Abbildung 8.1: Ablauf der explorativen Analyse des Fachgebiets.

Literaturauswahl

Als Basis für die Analyse wurden sechs anerkannte Fachbücher aus dem Fachgebiet Datenmanagement ausgewählt, zusätzlich wurde der DAMA-DMBoK berücksichtigt, da er die einzige extensive Charakterisierung des Fachgebiets aus professioneller Sicht darstellt. Bei der Auswahl der Literatur wurde insbesondere darauf geachtet, dass zentrale Grundlagenwerke aus dem Fachgebiet berücksichtigt werden. Da diese zwar typischerweise aus traditionellen Gründen einen starken Fokus auf das Themenfeld Datenbanken erkennen lassen, sie aber im Laufe der Zeit um weitere Aspekte erweitert wurden, können sie als zentral für das gesamte Fachgebiet angesehen werden. Um detailliertere Aspekte auch aus den neueren Themen des Datenmanagements zu berücksichtigen, wurden zusätzlich solche Werke hinzugezogen, die sich speziell auf jüngere Themen beziehen und diese intensiver betrachten. Es wurde daher folgender Literaturkanon ausgewählt:

- *Elmasri und Navathe (2009)*: Grundlagen von Datenbanksystemen
- *Kemper und Eickler (2013)*: Datenbanksysteme
- *Kudraß (2015)*: Taschenbuch Datenbanken
- *Piepmeyer (2011)*: Grundkurs Datenbanksysteme
- *Unland und Pernul (2015)*: Datenbanken im Einsatz
- *Edlich et al. (2011)*: NoSQL
- *DAMA International (2010)*: The DAMA Guide to the Data Management Body of Knowledge

[34] Das Clustering wurde gegenüber der Methodik nach Mayring ergänzt, um einen besseren Überblick über die große erwartete Menge an Begriffen zu bekommen und einen Übergang zur zweiten Analysephase zu ermöglichen.

[35] Die Überprüfung der Vollständigkeit wurde gegenüber der Methodik nach Mayring ergänzt, um einen möglichst hohen Grad an Validität sicherzustellen.

Obwohl sich der Literaturkanon hauptsächlich aus deutschsprachiger Literatur zusammensetzt, kann eine Beeinflussung der Analyseergebnisse durch diesen Aspekt ausgeschlossen und im Gegenteil angenommen werden, dass die Literaturauswahl ausreichend ist, um eine internationale Perspektive zu berücksichtigen: Bei zwei der genutzten Lehrbücher handelt es sich um deutsche Ausgaben von englischsprachigen und international anerkannten Werken, die nur zur Vereinheitlichung der Analyse in ihrer deutschen Variante herangezogen wurden. Eine zusätzliche Absicherung findet jedoch dadurch statt, dass mit dem DAMA-DMBoK die internationale professionelle Sichtweise auf das Fachgebiet berücksichtigt wurde und dadurch, dass es sich auch bei den Autoren der anderen Lehrbücher um international renommierte Wissenschaftler handelt. Eine eingeschränkte bzw. verschobene Sichtweise ist daher nicht zu erwarten. Dies bestätigte sich auch während der durchgeführten Analyse: Es konnte nicht festgestellt werden, dass die rein deutschsprachigen Lehrbücher andere Schwerpunkte setzen, als die auch englischsprachig verfügbaren oder der DAMA-DMBoK.

Aufgrund des explorativen Charakters der durchgeführten Analyse war es außerdem nicht nötig, die unterschiedliche Art sowie den unterschiedlichen Umfang und Detailgrad der Literatur zu berücksichtigen, da eine quantitative Betrachtung der Anzahl an Codierungen für die Exploration des Fachgebiets nicht nötig bzw. hilfreich ist.

Festlegung der Kodiereinheit und der Klassifikationskriterien

Die qualitative Inhaltsanalyse nach Mayring sieht vor, Kodiereinheiten und Kriterien für die Kodierung von Textstellen festzulegen. Aufgrund des Ziels dieser Analyse, einen umfassenden Überblick über das Fachgebiet zu gewinnen, ist es sinnvoll, möglichst viele Aspekte aufzunehmen. Eine Konsolidierung und Zusammenfassung der Ergebnisse erfolgt in der zweiten Analysephase. Daher wurde entschieden, die Größe der Kodiereinheit möglichst offen zu lassen: Prinzipiell wurde es für zulässig erachtet, auch mehrere Begriffe als eine Einheit zu kodieren, insofern diese alleinstehend nicht aussagekräftig genug wären bzw. relevante Informationen verloren gehen würden. Es wurde jedoch grundsätzlich versucht, die Kodiereinheit minimal zu halten, um die Übersichtlichkeit des Kategoriensystems zu steigern und Redundanzen zu vermeiden. Mehrfachkodierungen derselben Einheit wurden aus demselben Grund zwar prinzipiell zugelassen, aber, durch den Versuch die Größe der Kodiereinheit minimal zu halten, soweit möglich vermieden.

Als Kriterium für die Kodierung bzw. Klassifikation wurde festgelegt, dass die betrachtete Einheit einen klar erkennbaren Bezug zum Datenmanagement aufweisen und somit über allgemeine Konzepte der Informatik klar hinausgehen musste. Entsprechend wurde vermieden, eher allgemeine und in vielen Bereichen der Informatik auftretende Begriffe in das Kategoriensystem aufzunehmen, sofern diese nicht im Datenmanagement eine spezielle Bedeutung bzw. Relevanz haben. Trotz dieses relativ weichen Selektionskriteriums wird eine negative Beeinflussung der Ergebnisse aber nicht angenommen: Während dieses Kriterium sicherlich für ein umfangreiches und nicht auf den ersten Blick klar überblickbares Kategoriensystem sorgt, stellt es gleichzeitig sicher, dass die Breite des Fachgebiets berück-

sichtigt wird. Das Ziel einer explorativen Analyse wird daher unterstützt, während eine Erhöhung der Übersichtlichkeit auf den zweiten Analyseschritt verlagert wurde. Der Kontext in dem eine Kodiereinheit auftritt, wurde gegebenenfalls genutzt, um zu entscheiden, ob ein Textabschnitt den geforderten Bezug zum Datenmanagement aufweist.

Entwicklung des Kategoriensystems

Die Kodierungsphase wurde in mehreren Schritten iterativ durchlaufen. Pro Durchgang wurde jeweils ein Buch aus dem Literaturkanon ausgewählt und analysiert. Dieses wurde, beginnend mit Inhaltsverzeichnis und Register, nach Aspekten durchsucht, die das Klassifikationskriterium erfüllen. Solche wurden, soweit noch nicht im Kategoriensystem vorhanden, in dieses aufgenommen: Falls ein neuer Begriff einen Unteraspekt bzw. eine Detaillierung eines bereits aufgenommenen Aspekts darstellte, wurde er diesem untergeordnet. Es wurde dabei angestrebt, eine möglichst detaillierte Einordnung vorzunehmen, d. h. neue Begriffe möglichst tief im Kategoriensystem anzuordnen, sodass die obersten Ebenen geringen Umfang hatten bzw. nur wenige direkte Kodierungen auf sich vereinen.

In der ersten Iteration wurde der DAMA-DMBoK betrachtet, da dieser einen guten ersten Überblick über das Fachgebiet liefert und somit die Vermutung war, dass dieser bereits einen rudimentären Eindruck des entstehenden Kategoriensystems bieten kann. In den darauffolgenden beiden Durchgängen wurde noch eine relativ große Anzahl weiterer Begriffe, auch auf den oberen Ebenen, ergänzt. Ab der vierten Iteration stellte sich jedoch eine Sättigung ein, es wurden nur noch wenige weitere Themen ergänzt. Stattdessen kamen insbesondere zusätzliche Details hinzu. Das Kategoriensystem wuchs ab diesem Zeitpunkt entsprechend eher in die Tiefe als in die Breite.

Ergebnis der Analyse war somit ein hierarchisch organisiertes Kategoriensystem, welches zentrale Themen des Datenmanagements inklusive diverser Details beschreibt. Zur Steigerung der Übersichtlichkeit des Systems schon während der Analyse wurde außerdem entschieden, diesen Analyseschritt mit dem darauffolgenden Clustering zu koppeln bzw. dieses parallel durchzuführen.

Clustering

Schon während der Entwicklung des Kategoriensystems wurde die sehr umfangreiche Menge an Begriffen in induktiv gebildeten Gruppen parallel zum Aufbau des Kategoriensystems reorganisiert, um einen besseren Überblick zu bekommen. Dieser Prozess wurde regelmäßig durch neu hinzugekommene Begriffe erneut angestoßen, sodass sich die induktiv gebildeten Gruppen regelmäßig veränderten und, mit zunehmender Analysedauer, schärften. Das Clustering wurde manuell durchgeführt, da keine Informationen zu den einzelnen Begriffen zur Verfügung standen, die eine automatische Clusterbildung ermöglicht hätten. Die gemeinsamen Ergebnisse von Analysedurchführung und Clustering sind in Abbildung 8.2 dargestellt.

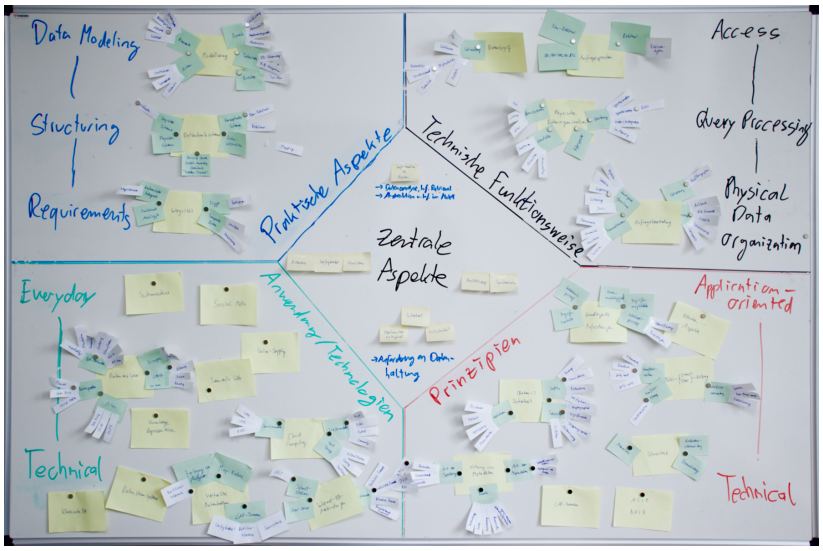


Abbildung 8.2: Ausschnitt des Zwischenergebnisses der explorativen Analyse und des manuellen Clusterings.

Im Rahmen des Clusterings konnten vier zentrale Gruppen ermittelt werden, die auch in Abbildung 8.2 erkennbar sind:

- *Modellierung* als eine der zentralen Tätigkeiten beim Umgang mit Datenmanagementsystemen, u. a. spezifiziert durch verschiedene *Datenschemata*, das Ziel der *Datenintegrität* oder verschiedene *Datenmodelle*
- *Implementierung*, d. h. eher technische Aspekte der Funktionsweise von Datenmanagementsystemen, wie beispielsweise den Systemen zugrundeliegende *Datenstrukturen* bzw. allgemein deren *physische Datenorganisation*, Grundlagen der *Anfragebearbeitung* sowie *Anfrageoptimierung*
- *Prinzipien*, die für Datenmanagementsysteme zentral sind, wie beispielsweise *Konsistenz*, Aspekte der *Datensicherheit* oder das *ACID-* und *BASE-Paradigma*
- *Beispielhafte Anwendungen* von Datenmanagement, wie beispielsweise *Datenbanken*, *Data Warehouses*, aber auch *Cloud Computing*, *Soziale Medien* und *Online-Shops*

Zusätzlich konnten Begriffe gefunden werden, die den Prinzipien zugehörig waren und besonders häufig genannt wurden. Um diese Auffälligkeit zu berücksichtigen, wurden sie in die speziell hervorgehobene Kategorie *zentrale Aspekte* eingeordnet. Dabei handelt es sich beispielsweise um *Synchronisation*, *Sicherheit*, *Mehrbenutzerbetrieb* und die im gesamten Fachgebiet zentrale *Unterscheidung von Information und Daten*.

Überprüfung der Vollständigkeit

Um schon in dieser ersten Analysephase einen möglichst hohen Grad an Vollständigkeit sicherzustellen und somit eine gute Basis für die weitere Analyse zu schaffen, wurde das

Ergebnis mittels einer teilautomatisierten Textanalyse validiert. Dazu wurde ein weiterer Dokumentenkorpus, der aus Vorlesungsskripten und -folien verschiedener Wissenschaftler aus dem Fachgebiet Datenmanagement bestand, hinsichtlich der in diesen am häufigsten vorkommenden Begriffe des Datenmanagements untersucht. Diese Arbeiten betrachten, wie auch der ursprüngliche Korpus, zum Teil das gesamte Forschungsgebiet, zum Teil auch nur Ausschnitte davon. Neben diesen Arbeiten wurde ein Großteil der Ausgaben des *Datenbank Spektrum* mit in den Korpus aufgenommen, da diese Zeitschrift der Fachgruppe *Datenbanken und Information Retrieval* der Gesellschaft für Informatik einen Einblick in die aktuellen Entwicklungen auf diesem Gebiet erlaubt. Um die in den Dokumenten am häufigsten vorkommenden Begriffe zu ermitteln, wurde wie folgt vorgegangen (vgl. Abbildung 8.3):

1. *Extraktion des reinen Texts der Dokumente:*
   Die als PDF vorliegenden Dokumente wurden zuerst, um die weitere Verarbeitung zu erleichtern, in reinen Text konvertiert (ggf. unter Nutzung von OCR-Techniken).
2. *Filterung von Stoppwörtern:*
   Es wurden alle Stoppwörter, d. h. Konjunktionen, Artikel usw., die in üblichen Listen vorkommen, herausgefiltert. Zusätzlich wurden jegliche Satzzeichen eliminiert.
3. *Aggregation der Worte nach Häufigkeit ihres Vorkommens:*
   Die Dokumente wurden zu einem Textdokument zusammengefasst, dessen Worte nach Häufigkeit ihres jeweiligen Auftretens aggregiert worden sind.
4. *Sortierung der Begriffe nach Anzahl der Nennungen:*
   Es erfolgte daraufhin eine Sortierung der Begriffe nach ihrer Auftretenshäufigkeit.
5. *Betrachtung der 300 häufigsten Begriffe:*
   a) *Zusammenführung verwandter Begriffe:*
      Da in der Begriffsliste Plural- und Singularformen sowie Abwandlungen desselben Wortstamms und Synonyme vorkamen, wurden diese zusammengefasst.
   b) *Filtern der Liste:*
      Daraufhin wurden Begriffe ausgefiltert, die keinen Bezug zum Feld Datenfeldmanagement aufweisen und somit das Klassifikationskriterium nicht erfüllen.
6. *Falls sich im vorherigen Schritt eine Änderung ergeben hat:* Rücksprung zu Schritt 3.

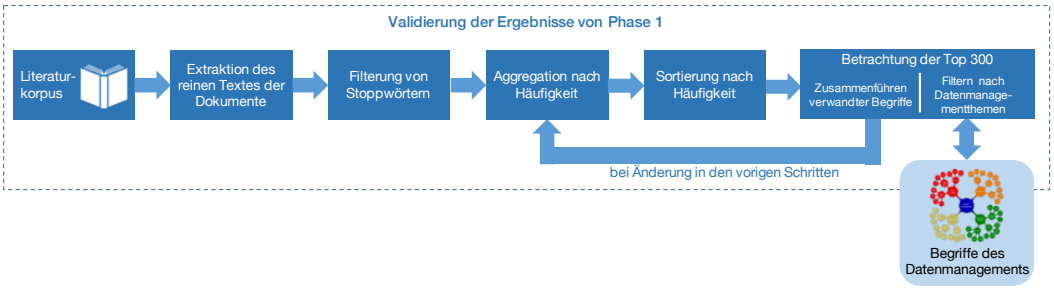


Abbildung 8.3: Ablauf der Validierung der Ergebnisse der ersten Phase.

Insgesamt wurden mit diesem Ansatz weitere 305 Dokumente mit 9.447 Seiten bzw. Folien analysiert, die zu über 1.800 verschiedenen Begriffen (nach Filterung von Mehrdeutigkeiten, Synonymen, Singular-/Pluralformen und Stoppwörtern) führten, von denen die 300 am häufigsten vorkommenden Begriffe detaillierter betrachtet wurden. Der Abgleich mit den vorherigen Ergebnissen führte dazu, dass zwar weitere Begriffe in das Kategoriensystem eingefügt werden konnten, es wurden dabei jedoch nur Ergänzungen in der Tiefe bzw. dem Detailgrad vorgenommen. Hingegen kam es zu keiner Erweiterung in der Breite. Somit kann ein hoher Grad an Vollständigkeit in der Breite angenommen werden. Trotzdem konnte durch diese Analyse eine weitere Verbesserung der Ergebnisse vorgenommen werden: Ein Vergleich der hinzukommenden Begriffe mit den bisherigen Clustern zeigte, dass eine Auflösung des Clusters „Zentrale Aspekte“ sinnvoll ist, da eine klare Einordnung in diese Gruppe kaum eindeutig möglich ist. Durch eine Neuordnung der Cluster konnten damit die in Abbildung 8.4[36] dargestellten Ergebnisse der ersten Analysephase gewonnen werden.

Abbildung 8.4: Ergebnisse der ersten Analysephase (dargestellt bis zur zweiten Ebene).

[36] Aus Gründen der Übersichtlichkeit stellt Abbildung 8.4 nur die Ergebnisse bis zur zweiten Ebene dar, sodass einzelne Begriffe die im Rahmen von Phase 2 als zentraler eingeschätzt wurden, hier nicht genannt, sondern nur durch übergeordnete Begriffe abgedeckt sind.

8.1.2 Phase 2: Ermittlung und Strukturierung der Schlüsselkonzepte

Durch die Ergebnisse der ersten Analysephase konnte zwar ein breiter Überblick über Datenmanagement gewonnen, aber noch keine prägnante Charakterisierung des Fachgebiets erstellt werden. Um dieses Ziel zu erreichen, wird nun in der zweiten Phase eine Strukturierung der Themen vorgenommen, um so die zentralen Schlüsselkonzepte des betrachteten Fachgebiets zu ermitteln. Die einzelnen Schritte, in die diese Phase aufgegliedert wurde, sind in Abbildung 8.5 dargestellt.

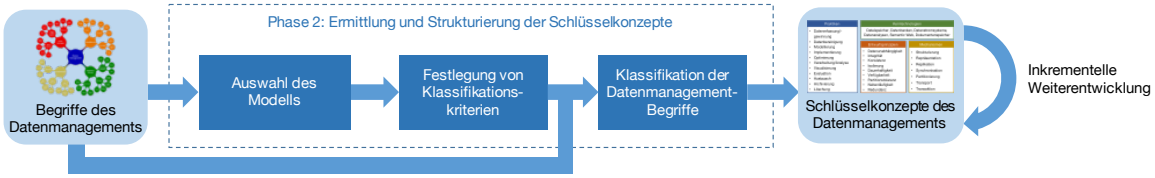


Abbildung 8.5: Ermittlung und Strukturierung der Schlüsselkonzepte.

Auswahl des Modells

Aufgrund des Ziels der ersten Phase, auf explorative Weise einen Überblick über das Fachgebiet zu geben, enthält das Zwischenergebnis dieser Phase Begriffe von unterschiedlichem Detaillierungs- und Abstraktionsgrad, die einen einfachen Überblick und eine übersichtliche Darstellung noch nicht erlauben. Um ein klares und verständliches Modell zu entwickeln, wurden daher die vier beim Clustering herausgebildeten Gruppen mit verschiedenen bereits etablierten Arbeiten zu Ideen, Prinzipien und Konzepten der Informatik verglichen: Insbesondere mit dem Framework der *Great Principles of Computing* konnten dabei große Gemeinsamkeiten erkannt werden, da dieses mit den Entwurfsprinzipien, Mechanismen, Kerntechnologien und Praktiken sehr ähnliche Bereiche berücksichtigt. Auch aufgrund der ähnlichen Zielsetzung dieser Arbeit, der fachlich geprägten Charakterisierung einer Wissenschaft bzw. eines ihrer Teilbereiche, scheint eine Adaption des den Great Principles zugrundeliegenden Modells für die angestrebte Charakterisierung von Datenmanagement sinnvoll, sodass die Entscheidung für diesen Weg getroffen wurde.

Festlegung von Klassifikationskriterien

Um eine nachvollziehbare, eindeutige und objektive Zuordnung der in der vorherigen Phase gefundenen Begriffe zu den vier Modellbereichen zu ermöglichen, mussten zuerst klare Kriterien für diese entwickelt werden. Dabei konnte auf Dennings textuelle Beschreibungen zurückgegriffen werden *(vgl. Denning, 2003b; Denning, 2004; Denning und Martell, 2015)*, die von ihm angelegten Kriterien wurden hingegen nicht publiziert. Entsprechend wurden die vier Bereiche wie folgt charakterisiert:

- *Kerntechnologien* stellen konkrete Anwendungen bzw. Technologien dar, die einen klaren Bezug zu Datenmanagement haben. Gleichzeitig repräsentieren sie auch zentrale Forschungsrichtungen des Fachgebiets.
- *Praktiken* sind Aktivitäten bzw. Methoden, die dem Datenmanagement zuzuordnen sind. Sie repräsentieren auch Kompetenzen, die bei der bzw. für die Nutzung und/oder Entwicklung von Datenmanagementsystemen notwendig sind.
- *Entwurfsprinzipien* müssen beim Entwurf von Datenmanagementsystemen berücksichtigt werden. Sie können jedoch auch bei der Auswahl eines Systems für einen konkreten Anwendungsfall genutzt werden, indem sie als Entscheidungskriterien herangezogen werden.
- *Mechanismen* stellen grundlegende Gesetze, Annahmen, Vorgehen oder Absprachen dar, die das Fachgebiet durchziehen und die für dieses fundamental sind. Sie beschreiben die grundlegende (technische) Funktionsweise von Datenmanagementsystemen.

Klassifikation und Auswahl der zuvor gefundenen Begriffe

Diese Beschreibungen konnten als Basis für die Klassifikation eingesetzt werden, die entsprechend des in Abbildung 8.6 abgebildeten Schemas ablief. Wie auch im Schema dargestellt, findet an dieser Stelle noch keine Auswahl bzw. Reduktion der Begriffsmenge statt (außer in Form der Validierung, ob ein Thema speziell dem Datenmanagement zugeordnet ist). Dies ist beabsichtigt, da eine solche mit Blick auf den einzelnen Begriff kaum valide möglich ist. Stattdessen findet eine Auswahl bzw. die Beschränkung auf zentrale Begriffe erst nach der eigentlichen Klassifikation der Begriffe innerhalb der jeweiligen Kategorien statt.

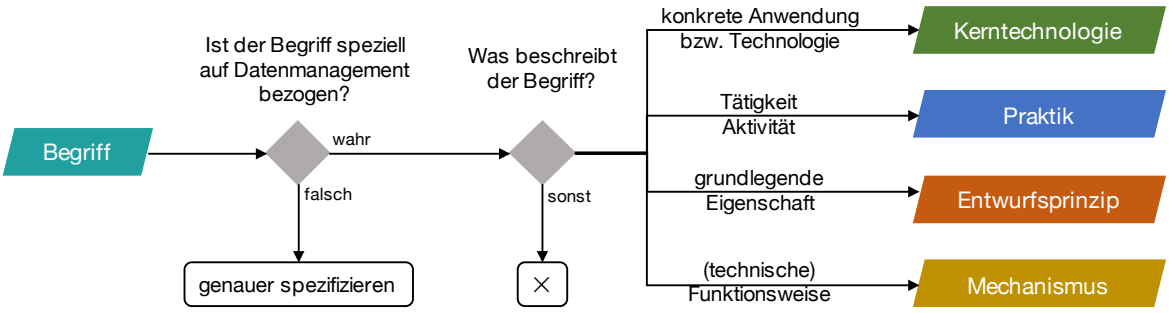


Abbildung 8.6: Zuordnung der gefundenen Begriffe zu den verschiedenen Kategorien.

Aufgrund des Ziels ein möglichst vollständiges, aber zugleich knappes und einfach erfassbares Modell zu erstellen, wurden Begriffe mit großer Überschneidung daraufhin zusammengefasst, um die Anzahl der vorkommenden Begriffe zu reduzieren. Entsprechend wurden auch Details vernachlässigt, die durch übergeordnete Begriffe ausreichend thematisiert wurden. Dies führte beispielsweise zu folgenden Vereinfachungen des Modells:

- Die drei Datenanalysemethoden *Klassifikation*, *Assoziation* und *Clustering* wurden unter der Praktik *Analyse* zusammengefasst.
- Die *Normalisierung* wurde, obwohl sie eine hohe Relevanz im derzeitigen Informatikunterricht genießt, nicht explizit aufgenommen, da es sich dabei um typische Möglichkeiten zur *Optimierung* eines Datenbankschemas, insbesondere hinsichtlich der Vermeidung von *Redundanz* bzw. zur Sicherstellung von *Konsistenz* handelt und somit durch diese Bereiche ausreichend abgedeckt wird. Entsprechend verhält es sich beispielsweise mit Möglichkeiten zur *Datenkompression*.
- *Funktionale Abhängigkeiten* wurden als Mittel zur Sicherstellung von *Konsistenz* eingeordnet und sind daher nicht direkt im Modell enthalten.
- *Metadaten*, *Datenmodelle* und *Primär-* sowie *Fremdschlüssel* wurden als Mittel zur *Strukturierung* betrachtet, die gleichzeitig auch zur Sicherstellung von *Integrität* und *Konsistenz* dienen.

Neben diesen Entscheidungen zur Vereinfachung blieb auch bei der Zuordnung der Begriffe zu den verschiedenen Modelldimensionen ein gewisser Entscheidungsspielraum: Obwohl versucht wurde, die Zuordnungskriterien möglichst klar und eindeutig zu definieren, konnte an verschiedenen Stellen eine eindeutige Zuordnung nicht direkt im ersten Zuordnungsversuch getroffen werden. Dies war beispielsweise bei der *Synchronisation* der Fall, die einerseits als Mechanismus interpretiert werden kann (im Sinne der Fragestellungen „Wie funktioniert Synchronisation?“, „Welche Probleme treten bei der Synchronisation von Daten auf?“ und „Wie werden diese Probleme (technisch) gelöst und/oder vermieden?“), aber andererseits auch als Entwurfsprinzip (mit dem Fokus darauf, wie Synchronisation in einem System auf technischer Ebene ermöglicht wird, welche Voraussetzungen nötig sind, wie Daten dafür strukturiert und vorbereitet werden müssen und welche Einflüsse Synchronisation auf andere Entwurfsprinzipien hat). Falls die Zuordnung in solchen Fällen nicht durch Überprüfung des Kontexts, in denen diese Begriffe in der Literatur vorkamen, geklärt werden konnte, wurde sie offengelassen, bis die anderen Zuordnungen getroffen waren. Dadurch konnten diese Begriffe gegebenenfalls unter Zuhilfenahme der erfolgten Kategorisierungen klarer eingeordnet werden. Für die *Synchronisation* konnte am Ende beispielsweise erkannt werden, dass deren Entwurfsaspekte bereits durch die Entwurfsprinzipien *Konsistenz*, *Integrität* und *Nebenläufigkeit* abgedeckt sind, während ihre funktionalen Aspekte noch nicht ausreichend durch andere Begriffe berücksichtigt waren. Eine Einordnung als Mechanismus war entsprechend sinnvoll, damit beide Sichtweisen auf die Synchronisation abgedeckt sind. In Fällen, in denen ein noch größerer Spielraum herrschte, wurde die Zuordnung mit anderen Wissenschaftlern diskutiert und diese gemeinsam vorgenommen, um eine möglichst hohe Objektivität zu erreichen.

8.1.3 Inkrementelle Weiterentwicklung

Da angestrebt wurde, dass das Modell der Schlüsselkonzepte des Datenmanagements das Fachgebiet nicht nur aus fachdidaktischer, sondern auch fachlicher Perspektive beschrei-

ben kann, ist die Validierung des Modells aus fachwissenschaftlicher Perspektive essenziell. Aus diesem Grund wurde entschieden, das Modell ausführlich mit verschiedenen Experten zu diskutieren und weiterzuentwickeln. Dazu wurde es im Rahmen von informatikdidaktischen Konferenzen vorgestellt, aber auch durch Einbeziehung von Experten aus dem Fachgebiet evaluiert. Als erster Schritt wurde dabei ein semistrukturiertes Experteninterview mit einem international renommierten Professor aus dem Fachgebiet durchgeführt. Dabei wurden insbesondere folgende Aspekte herausgegriffen:

- Vor der Vorstellung des Modells:
  - Wie kann Datenmanagement aus fachlicher Sichtweise zutreffend beschrieben werden?
  - Welche sind die zentralen Konzepte dieser Disziplin?
- Kurzvorstellung des Modells:
  - Kurze Skizzierung der methodischen Vorgehensweise
  - Vorstellung der vier Bereiche des Modells und ihrer Bedeutung
  - Stichpunktartige Erläuterung der Zuordnung einzelner Begriffe
- Nachdem das Modell vorgestellt wurde:
  - Welche Aspekte fehlen im Modell?
  - Enthält das Modell Widersprüche oder fachliche Fehler?
  - Wie gut wird Datenmanagement durch das Modell charakterisiert?
  - Ist das Modell auch aus fachlicher Sicht geeignet, um Datenmanagement zu beschreiben?

Der befragte Experte charakterisierte das Fachgebiet insbesondere durch das Streben nach Datenunabhängigkeit bzw. Datenabstraktion mit dem Ziel, Details der Speicherung zu verbergen und die Adaption des Systems an sich wandelnde Anforderungen zu ermöglichen. Als zentrale Aspekte wurden dabei insbesondere deklarative Sprachen und Abstraktion genannt. Die methodische Vorgehensweise zur Erstellung des Modells und der Aufbau auf bekannter Fachliteratur wurde als dienlich zur Erreichung der gesteckten Ziele erachtet. Insbesondere konnten alle Lehrbücher, die durch den Experten als besonders zentral für das Fachgebiet erachtet wurden, im genutzten Literaturkorpus wiedergefunden werden. Folglich konnten keine Verbesserungsvorschläge bzw. Ideen für den methodischen Ansatz benannt werden.

Auch zum Modell an sich war eine große Zustimmung erkennbar, der Experte erachtete dieses als passende Charakterisierung seines Fachgebiets. Insbesondere hinsichtlich der ermittelten Praktiken wurde dies deutlich: Diese wurden als konkret genug eingeschätzt, um wichtige Tätigkeiten aus dem Datenmanagement zu beschreiben. Es wurden jedoch

auch verschiedene Aspekte diskutiert, die zu geringfügigen Veränderungen des Modells führten:

- Es wurde diskutiert, ob weitere Praktiken, die sich mit Aspekten wie der Vergabe von Rechten beschäftigen, eine wertvolle Ergänzung darstellen könnten. Da alle diskutierten Kandidaten jedoch nicht spezifisch für Datenmanagement waren, wurde unter Beachtung des Zwecks des Modells entschieden, darauf zu verzichten.
- Die Kerntechnologie *Datenspeicher* wurde durch die beiden konkreteren Begriffe *Dateispeicher* und *Dokumentenspeicher* ersetzt, die auf gleicher konzeptioneller Ebene wie beispielsweise die schon vorhandenen *Datenbanken* angesiedelt sind. Auf diese Weise wird die Konsistenz des Modells erhöht, ohne dessen Charakter zu verändern.
- Die Relevanz des Konzepts *Transaktion*, das vor der Evaluation noch nicht explizit im Modell vertreten war, wurde diskutiert, da sie einerseits durch andere Konzepte im Modell mit abgedeckt wird (insbesondere *Nebenläufigkeit* und *Synchronisation*), andererseits jedoch als zentraler Aspekt des Themengebiets weitere Aspekte ergänzen kann. Somit wurde entschieden, die Transaktionen durch Aufnahme als Mechanismus stärker zu betonen.
- Weitere Veränderungen, die im Rahmen der Validitätsüberprüfung vorgenommen worden sind, konzentrieren sich auf die Anpassung einzelner Begriffe: Beispielsweise wurde über den Begriff *Konkurrenz* diskutiert, der zu diesem Zeitpunkt noch im Modell auftrat, der aber auch durch *Parallelisierung* oder *Nebenläufigkeit* hätte ersetzt werden können. Da hier keine klare Präferenz vorherrschte, blieb dieser hier jedoch vorerst bis zu einer weiteren Überprüfung der Ergebnisse bestehen.

Zu diesem veränderten Modell konnte in weiteren Diskussionen mit Personen aus dem Fachgebiet eine relativ hohe Zustimmung festgestellt werden. Die dabei immer wieder genannten Kritikpunkte beschäftigten sich insbesondere mit der Wahl von Begriffen und mit möglicherweise fehlenden Begriffen:

- Insbesondere wurde auch an dieser Stelle wieder die *Konkurrenz* genannt, die durch *Nebenläufigkeit* ersetzt werden sollte, wobei hier zwei nachvollziehbare Gründe genannt wurden, die zur Umbenennung führten: Einerseits ist die Nebenläufigkeit der im Fachgebiet üblichere Begriff, andererseits scheint Konkurrenz insbesondere eine ungünstige Übersetzung der englischen *concurrency* zu sein, die in der Anfangszeit des Fachgebiets verbreitet war, von der es sich aber mittlerweile gelöst hat.
- Auch die Praktik *Optimierung* wurde viel diskutiert, da Experten aus dem Fachgebiet diese insbesondere mit der Anfrageoptimierung in Verbindung setzen. Ein solches Begriffsverständnis ist jedoch sowohl bei Lehrerinnen und Lehrern als auch Schülerinnen und Schülern nicht zu erwarten, sondern eher die allgemeinere Interpretation von Optimierung, die hier auch angestrebt wird, sicherlich aber eine Anfrageoptimierung miteinschließen kann. Gleichzeitig wird durch Nutzung dieses allgemein bekannten und genutzten Begriffs auch kein Fehlverständnis hinsichtlich des Fach-

begriffs erzeugt, das es zu vermeiden gelten würde. Aus didaktischer Sicht kann der Begriff daher beibehalten werden. Gleichzeitig zeigt sich jedoch der Bedarf nach einer eindeutigen Klärung aller verwendeten Begriffe, wie sie im Folgenden und detaillierter in Anhang C stattfindet.

- Als fehlender Begriffe wurde beispielsweise *Atomarität* genannt, der im Modell jedoch durch andere Bereiche (insbesondere *Transaktionen*) mit abgedeckt wird. Auch eine explizite Berücksichtigung von *Performanz* wurde vorgeschlagen, andererseits ist diese jedoch nicht spezifisch für Datenmanagement, sondern ein übergeordnetes Ziel der Informatik, das wie *Sicherheit* und *Nutzbarkeit* (vgl. Abschnitt 8.3.3) durch andere Prinzipien ausgedrückt werden kann (insbesondere *Verfügbarkeit*, *Replikation*, *Synchronisierung*). Aus fachlicher Sicht ist diese mögliche Ergänzung klar nachvollziehbar, da sonst wesentliche Punkte nur relativ schwer erkennbar auftreten. Auch diesem Kritikpunkt wurde jedoch, anstatt das Modell durch weitere Begriffe zu ergänzen, durch die in Anhang C angehängte detailliertere Beschreibung begegnet, die solchen Aspekten Rechnung trägt. Ein bis dato fehlender Begriff wurde jedoch als Folge der Diskussion und aufgrund seiner klaren Bedeutung noch aufgenommen: die Praktik *Verarbeitung*.

Zusammenfassend beschreibt das Modell, obwohl es unter fachdidaktischen Gesichtspunkten entstand, auch die aus fachlicher Sicht zentralen Aspekte des Fachgebiets relativ gut. Um ein übersichtliches und einfach verständliches Modell erstellen zu können, mussten jedoch Einschränkungen in der abgedeckten Tiefe und Breite in Kauf genommen werden, denen durch die detaillierteren Beschreibungen in Anhang C begegnet werden soll. Trotz dieser Einschränkungen ist das entwickelte Modell der Schlüsselkonzepte von Datenmanagement sowohl aus fachdidaktischer als auch fachlicher Sicht zur Charakterisierung des Fachgebiets geeignet. Sicherlich sind verschiedene Aspekte des Modells diskutierbar und können nur im ausführlichen wissenschaftlichen Diskurs langfristig einbezogen werden, sodass auch zukünftig eine inkrementelle Weiterentwicklung des Modells geplant ist, die gegebenenfalls auch weitere Ideen des Fachgebiets mit aufgreifen kann.

8.2 Modell der Schlüsselkonzepte des Datenmanagements

Aus den beiden Analysephasen resultierte ein an das Framework der *Great Principles of Computing* angelehntes *Modell der Schlüsselkonzepte des Datenmanagements*, das in Abbildung 8.7 dargestellt ist. Um sowohl die Struktur des Modells als auch die enthaltenen Schlüsselkonzepte zu verdeutlichen, werden die vier Bereiche des Modells im Folgenden vorgestellt und kurz erläutert. Eine ausführliche Charakterisierung erfolgt an dieser Stelle jedoch aus Gründen der Übersichtlichkeit nicht, stattdessen wurde dieser Arbeit in Anhang C eine ausführliche Beschreibung beigefügt, in der zur besseren Nachvollziehbarkeit der Begriffsinterpretationen alle beschriebenen Schlüsselkonzepte ausführlich erläutert und deren Einordnung begründet werden. Zusätzlich sind dort zu jedem Schlüsselkonzept *Kernaussagen* enthalten, die dieses weiter explizieren und greifbarer machen.

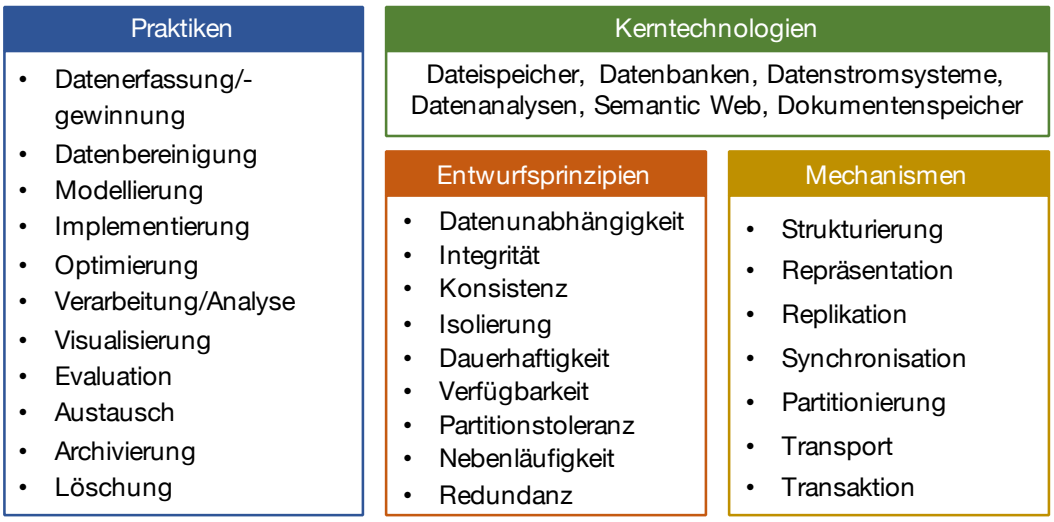


Abbildung 8.7: Modell der Schlüsselkonzepte des Datenmanagements.

8.2.1 Kerntechnologien

In den im Modell genannten *Kerntechnologien* spiegeln sich die Fortschritte im Datenmanagement gegenüber dem ursprünglichen Fachgebiet Datenbanken wider. Diese Kerntechnologien stellen heute gleichzeitig zentrale Forschungsrichtungen des Fachgebiets dar. Wie auch bei Denning, dessen Auswahl an Kerntechnologien sich im Laufe der Zeit verändert und insbesondere auch erweitert hat, ist zu erwarten, dass auch im Datenmanagement zukünftig weitere Kerntechnologien hinzukommen bzw. aktuelle an Bedeutung verlieren. Die derzeit relevante Auswahl an Kerntechnologien des Datenmanagements beläuft sich auf Folgende:

- Unter *Dateispeicher* werden klassische dateibasierte Datenspeicher verstanden, die Daten in Form von Dateien und mithilfe von Dateisystemen verwalten. Typischerweise werden zu jeder Datei in einer zusätzlichen Struktur verschiedene Metadaten gespeichert. Der Zugriff auf die Daten erfolgt dateibasiert, ohne dass das System über deren innere Struktur Bescheid weiß bzw. wissen muss.

- *Datenbanken* können unterteilt werden in relationale und nicht-relationale Datenbanksysteme: Die Nutzung des relationalen Modells setzt eine klare Strukturierung der Daten voraus und erzwingt deren Einpassung in ein (meist anwenderdefiniertes) Schema. Auf diese Weise kann die Erfüllung des ACID-Prinzips (vgl. Abschnitt 3.2), das die traditionellen Anforderungen an Datenbanken berücksichtigt, sichergestellt werden. Bei nichtrelationalen Modellen werden hingegen beispielsweise Integritätsanforderungen weniger stark gewichtet. Im Gegenzug lassen nichtrelationale Modelle jedoch verschiedene andere Strukturierungsmöglichkeiten für Daten zu und eröffnen mehr Freiräume.

- *Datenstromsysteme* sind, im Gegensatz zu den anderen angesprochenen Systemen, nicht für die dauerhafte Speicherung von Daten, sondern für eine schnelle und effiziente Analyse optimiert. Durch Ausnutzung des Datenstromprinzips erlauben sie eine sofortige Analyse neuer Daten, während gleichzeitig wenig Overhead entsteht, der die Analyse ausbremst, sodass die Ergebnisse üblicherweise sofort bereitstehen.
- *Datenanalysen* bzw. die verschiedenen dahinterstehenden *Datenanalysetechnologien* sind heute zentral für Arbeit und Forschung auf dem Gebiet Datenmanagement. Insbesondere im Zusammenhang mit *Big Data* und *Data Mining* entstanden und entstehen verschiedene Ansätze zur systematischen, schnellen und erkenntnisreichen Analyse von Daten.
- Ein weiterer zentraler Forschungsbereich des Datenmanagements ist das *Semantic Web*, bei dem es sich um die systematische Anreicherung von im Web zugreifbaren Daten handelt, mit dem Ziel diese einfacher auffindbar, analysierbar und navigierbar zu machen. Das Semantic Web stellt einen beispielhaften Anwendungsbereich von Metadaten dar.
- *Dokumentenspeicher* speichern Daten stärker strukturiert als Dateispeicher. Im Gegensatz zu diesen sind hier auch die inneren Strukturen der Dokumente zumindest zum Teil zugänglich und relevant, sie können zur Strukturierung und zum Zugriff auf die Dateien verwendet werden. Die Struktur wird jedoch nicht vorher (beispielsweise in Form eines Schemas) definiert, außerdem wird keine einheitliche Struktur für alle Dokumente erzwungen, sodass eine geringere Strukturierung als bei relationalen Datenbanken vorliegt.

8.2.2 Praktiken

Die *Praktiken* entsprechen den Aktivitäten und Methoden, die im Datenmanagement üblicherweise genutzt werden und dort nötig sind. Dabei sind insbesondere die folgenden zentral:

- Die *Datenerfassung/-gewinnung* beinhaltet alle Tätigkeiten, die zu Beginn des Verarbeitungs- bzw. Analyseprozesses stattfinden und Daten für weitere Verarbeitungsschritte verfügbar machen. Dies kann die Erfassung neuer Daten, beispielsweise mit Sensoren, die Zugänglichmachung von Daten durch geeignete Strukturierung oder Konvertierung aus anderen Formaten, aber auch die Recherche nach und Abfrage von geeigneten bereits existierenden Datensätzen sein. Gegebenenfalls können Daten mehrerer Quellen bereits an dieser Stelle zusammengeführt werden.
- Eine *Datenbereinigung* wird nötig, falls die vorliegenden Datensätze ungültige (z. B. Auslesefehler von Daten eines Sensors oder Werte außerhalb des definierten/zulässigen Wertebereichs), falsche (z. B. erkennbare Messfehler) oder ungeeignet formatierte Daten (z. B. Datumsangaben als Klartext oder im falschen Format) enthalten, die gefiltert und/oder korrigiert werden müssen.

- *Modellierung* wird insbesondere zur klaren und verständlichen Strukturierung von Daten und zur Verdeutlichung von Zusammenhängen zwischen verschiedenen Datensätzen genutzt, aber auch um einen Überblick über bereits existierende Datensätze und deren Struktur zu bekommen.
- Die *Implementierung* des Datenmodells in einem realen Datenmanagementsystem ermöglicht die Nutzung und Speicherung von Daten und ist damit grundlegend für die folgenden Praktiken.
- Die *Optimierung* umfasst beispielsweise die Anreicherung von Daten durch Metadaten zur Erreichung eines schnelleren Zugriffs (z. B. Indizierung), die Kombination von Daten, aber auch alle anderen Ansätze, die darauf abzielen, die Speicherung von und den Zugriff auf Daten möglichst effizient zu gestalten.
- Die *Verarbeitung/Analyse* von Daten umfasst insbesondere Aggregation von Daten, aber auch die Erzeugung neuer Informationen aus Daten unter Nutzung verschiedener Datenanalysemethoden, wie zum Beispiel *Clustering*, *Assoziation* und *Klassifikation*.
- Techniken zur *Visualisierung* von Daten werden genutzt, um die Analyseergebnisse verständlich und gut erfassbar für den Menschen aufzubereiten.
- Die *Evaluation* der Ergebnisse umfasst, neben der Beurteilung der eigentlichen Ergebnisse, auch die Einschätzung der Qualität des ursprünglichen Datensatzes und des Analyseansatzes.
- Der *Austausch* von Daten kann die Analyseergebnisse aber auch die Originaldaten umfassen und auf verschiedenen Wegen stattfinden.
- Die längerfristige *Archivierung* von Daten wird genutzt, um diese für zukünftige (oft noch nicht vorhersehbare) Zwecke zu nutzen. Durch die Archivierung wird die weitere Nutzung von Daten für eine gewisse Zeit unterbrochen, sie werden aber für mögliche spätere Nutzungen weiterhin vorgehalten.
- Die *Löschung* der Daten kann aus verschiedenen Gründen erfolgen: Neben der Löschung zur Gewinnung von Speicherplatz, kann sie beispielsweise auch nötig sein, um das Persönlichkeitsrecht von Personen zu wahren. Durch die (sichere) Löschung wird, im Gegensatz zur Archivierung, eine spätere Verwendung der Daten unterbunden.

Obwohl die Liste an Praktiken des Datenmanagements relativ umfangreich ist, ist eine weitere Reduzierung nur geringfügig möglich: Es könnten allenfalls die Praktiken *Archivierung* und *Löschung* zusammengefasst werden. Selbst diese Reduktion würde jedoch zu einer weniger deutlichen Ausprägung der unterschiedlichen Rolle dieser beiden Praktiken bezüglich der zukünftigen Verwendung der Daten führen. Aus diesem Grund wurde auf diese Reduktion verzichtet. Bei allen anderen Praktiken treffen ähnliche Argumente noch deutlicher zu, sodass ein Weglassen oder Zusammenfassen einzelner Praktiken zu einem Verlust relevanter Aspekte führen würde und allenfalls zur Schwerpunktsetzung, nicht aber im Allgemeinen, geschehen kann.

8.2.3 Entwurfsprinzipien

Während die Praktiken beschreiben, wie der Umgang mit Daten und Datenmanagementsystemen erfolgt, konzentrieren sich die *Entwurfsprinzipien* auf die Beschreibung von Eigenschaften, die Datenmanagementsysteme erfüllen sollen und die somit bei der Entwicklung und Auswahl von Datenmanagementsystemen zentral sind. Als Entwurfsprinzipien des Datenmanagements wurden die folgenden identifiziert:

- *Datenunabhängigkeit:*
  Die Arbeit mit Daten ist ohne Kenntnis ihrer internen Speicherung möglich, da Details der Implementierung durch Abstraktion vom Nutzer verborgen werden.
- *Integrität:*
  Der im System gespeicherte Datenbestand bleibt unter allen Umständen unversehrt.
- *Konsistenz:*
  Der Datenbestand weist keine logischen Widersprüche auf.
- *Isolierung:*
  Parallele Anfragen an das Datenmanagementsystem können sich nicht gegenseitig zu beeinflussen.
- *Dauerhaftigkeit:*
  Einmal durchgeführte Änderungen im Datenbestand bleiben dauerhaft erhalten, solange sie nicht explizit durch einen Anwender geändert werden.
- *Verfügbarkeit:*
  Der Zugriff auf Daten ist jederzeit schnell und effizient möglich.
- *Partitionstoleranz:*
  Das System kann selbst bei Ausfällen der Kommunikation zwischen Teilen eines verteilten Datenspeichers weiter genutzt werden.
- *Nebenläufigkeit:*
  Parallele Anfragen an das Datenmanagementsystem können gleichzeitig ausgeführt werden, solange sie keine gemeinsamen Ressourcen benötigen.
- *Redundanz:*
  Zur Vermeidung von Inkonsistenzen wird die mehrfache Speicherung derselben Daten vermieden, gleichzeitig kann durch diese aber die Ausfallsicherheit eines Systems erhöht werden.

Es ist offensichtlich, dass Datenmanagementsysteme nicht alle Prinzipien zugleich ideal umsetzen können. Stattdessen muss, je nach Zielsetzung und Priorisierung, bei der Entwicklung bzw. Auswahl eines Systems entschieden werden, welche der Entwurfsprinzipien im konkreten Szenario zentral sind. Dies zeigt sich beispielsweise an dem in Abschnitt 3.2 beschriebenen *CAP-Theorem*, das dieses Spannungsfeld für die drei nicht gleichzeitig erreichbaren Entwurfsprinzipien *Konsistenz*, *Verfügbarkeit* und *Partitionstoleranz* verdeutlicht.

8.2.4 Mechanismen

Die *Mechanismen* des Datenmanagements beschreiben die technische Funktionsweise von Datenmanagementsystemen. Dabei geht es insbesondere darum, wie der Zugriff auf Daten, aber auch deren Speicherung und Transport funktionieren. Gleichzeitig sind diese Mechanismen jedoch auch für die Einhaltung verschiedener Entwurfsprinzipien des Datenmanagements zentral, beide Bereiche des Modells in enger Beziehung zueinanderstehen. Das Fachgebiet Datenmanagement wird durch folgende Mechanismen charakterisiert:

- *Strukturierung* von Daten beschreibt Maßnahmen, die ergriffen werden, um die Suche nach und den Zugriff auf Daten zu ermöglichen und möglichst schnell und effizient zu gestalten. Dabei handelt es sich beispielsweise um die Anreicherung von Daten mit Metadaten, deren Einordnung in (beispielsweise hierarchische) Primärstrukturen, oder den Aufbau von Sekundärstrukturen wie Suchindizes.
- *Repräsentation* befasst sich mit Methoden und Techniken zur Speicherung von Daten, beispielsweise durch Nutzung (interner) Datenstrukturen.
- *Replikation* bezeichnet die redundante Speicherung derselben Daten auf verschiedenen Medien bzw. Datenspeichern, insbesondere mit dem Ziel, die Verfügbarkeit der Daten sowie die Toleranz des Gesamtsystems gegenüber Ausfällen eines Speichers bzw. Teilsystems zu erhöhen.
- *Synchronisation* bezeichnet die Koordination gleichzeitiger bzw. konkurrierender Zugriffe auf Daten. Andererseits wird darunter aber auch die Replikation von Daten verstanden, wenn diese nicht nur einfach durch Kopieren geschieht, sondern auch eine Konflikterkennung beinhaltet.
- *Partitionierung* befasst sich mit der verteilten Datenspeicherung über verschiedene Datenspeicher hinweg. Dabei werden die zu speichernden Daten, beispielsweise zur Erhöhung der Zugriffsgeschwindigkeit oder aufgrund zu großer Datenmengen, auf die verschiedenen Teilsysteme bzw. Datenspeicher verteilt, statt sie beispielsweise wie bei der Replikation in Kopie auf diesen zu speichern.
- *Transport* beschreibt die Übertragung von Daten innerhalb eines Systems und über die Systemgrenzen hinweg. Zentrale Aspekte stellen dabei beispielsweise die Sicherheit und Vertraulichkeit der Übertragung dar.
- *Transaktionen* bezeichnen eine Gruppe von Abfragen, die entweder alle zusammen (erfolgreich) durchgeführt oder alle abgebrochen werden. Sie werden insbesondere eingesetzt, um die Isolation nebenläufiger Anfragen zu ermöglichen und die Fehlertoleranz zu erhöhen. Gleichzeitig sind sie auch für die Erfüllung der ACID-Eigenschaften eines relationalen Datenbanksystems zentral.

Diese Mechanismen sind grundlegend für alle typischen Datenmanagementsysteme. Beispielsweise muss Daten immer, selbst in eher unstrukturierten Systemen, eine gewisse *Struktur* zugrunde gelegt werden, um einen Zugriff auf diese zu ermöglichen: Ohne jegli-

che Struktur kann ein geordneter Zugriff nicht erfolgen, da ohne diese ein Datensatz nicht einmal identifiziert werden kann. Gleichermaßen werden immer Techniken eingesetzt, um Daten effizient intern zu *repräsentieren* und einen sicheren und zuverlässigen *Transport* zu ermöglichen. Nicht immer sind jedoch alle Mechanismen gleichermaßen relevant: Nur wenn mehrere Datenspeicher in einem System genutzt werden, spielen beispielsweise *Partitionierung* und *Replikation* eine wichtige Rolle, während andererseits erst die nebenläufige Arbeit mit einem System auch *Synchronisation* und *Transaktionen* zu relevanten Konzepten macht.

8.2.5 Fazit zum Modell

Die Aufteilung des Modells in die vier Bereiche *Praktiken*, *Kerntechnologien*, *Entwurfsprinzipien* und *Mechanismen* sorgt für einen klar strukturierten und gut erfassbaren Aufbau und betont gleichzeitig die Vielseitigkeit des Themengebiets Datenmanagement. Die jeweilige Charakterisierung der in den vier Bereichen zusammengefassten Aspekte wird in Anhang D in Form eines Posters noch einmal zusammengefasst. Durch die vier Perspektiven auf das Themengebiet kann das Modell auf verschiedene Weise und für verschiedene Zwecke interpretiert und angewendet werden, beispielsweise bei der Planung von Informatikunterricht: Es betont nicht nur die zentralen Aspekte dieses Fachgebiets, welche für die geeignete Auswahl und den Umgang mit solchen Systemen, aber auch für ein grundlegendes Verständnis ihrer Funktionsweise nötig sind, sondern bietet auch verschiedene Anknüpfungspunkte zu anderen Bereichen der Informatik und kann somit auch dazu beitragen, die Konzepte des Datenmanagements mit anderen Themen des Informatikunterrichts in Verbindung zu setzen und im Wissen der Schülerinnen und Schüler zu verankern. Auch außerhalb des Unterrichts zeigt das Modell jedoch Potenzial, da es insbesondere genutzt werden kann, um zentrale Aspekte des Fachgebiets zu beschreiben und dieses auch aus fachlicher Sicht zu charakterisieren.

8.3 Anwendung für den Informatikunterricht

Die vier Bereiche des Modells spielen für den Unterricht eine unterschiedliche Rolle: Die Kerntechnologien sind eher werkzeugorientiert und stellen entsprechend im Unterricht ein Hilfsmittel und keinen spezifischen Lerninhalt dar. Die zentralen Inhalte sind hingegen insbesondere in den Entwurfsprinzipien zu finden, die die informatischen Prinzipien, die hinter Datenmanagement stehen, explizieren, aber auch in den Mechanismen, durch die die Realisierung der Entwurfsprinzipien erst möglich wird. Die Praktiken stellen hingegen eine wichtige Leitlinie für den Informatikunterricht dar. Diese vielfältige Anwendbarkeit wird im Folgenden beispielhaft skizziert: Zuerst wird gezeigt, wie anhand des Modells auf den *Lebenszyklus von Daten* und den Prozess, der im Rahmen von Datenanalysen durchlaufen wird, geschlossen werden kann und wie dieser Prozess als Leitlinie für den Unterricht genutzt werden kann. Dann wird gezeigt, wie exemplarische Themen des Datenmanagements

auf relevante Schlüsselkonzepte untersucht und das Modell so zur Unterrichtsplanung eingesetzt werden kann. Schließlich wird anhand der Beispiele *Sicherheit* und *Nutzbarkeit* verdeutlicht, wie das Modell genutzt werden kann, um zentrale Konzepte der Informatik aus Perspektive des Datenmanagements aufzugreifen.

8.3.1 Interpretation der Praktiken als Lebenszyklus von Daten

Eine erste Anwendungsmöglichkeit des Modells ist die Strukturierung von Unterricht. Dazu können beispielsweise die einem Themengebiet zugrundeliegenden Prozesse ermittelt werden, die sowohl Lehrerinnen und Lehrern als auch Schülerinnen und Schülern einen Überblick ermöglichen. Entsprechend werden von Lehrkräften bei der Planung und Strukturierung ihres Unterrichts häufig verbreitete Vorgehensmodelle, wie in der Softwareentwicklung das Wasserfallmodell oder die agile Softwareentwicklung, als Orientierung und Leitlinie für den Unterricht herangezogen, da somit ein Prozessmodell zur Verfügung steht, das den gesamten Fortgang des Unterrichts in diesem Bereich unterstützt und den Lernenden einen Überblick gibt. Im Datenmanagement existiert ein derartiges anerkanntes Prozessmodell bisher nicht. Es gibt jedoch verschiedene Vorschläge für Datenlebenszyklen oder Datenmanagementprozesse, die diese Rolle prinzipiell einnehmen könnten, deren Verbreitung im Fachgebiet aber oft nur gering ist und die sich teils auf spezielle Bereiche (wie Forschungsdatenmanagement) konzentrieren. Auch aus dem Modell der Schlüsselkonzepte des Datenmanagements kann durch geeignete Interpretation ein Datenlebenszyklusmodell abgeleitet werden: Aufgrund der praktischen Orientierung solcher Modelle, stellen die ermittelten Praktiken des Datenmanagements dafür den geeigneten Ansatzpunkt dar. Durch eine Sortierung der Praktiken unter Berücksichtigung ihrer gegenseitigen Abhängigkeiten[37] und durch Anordnung als Kreislauf dienen die Praktiken als Modell des Lebenszyklus von Daten und strukturieren somit auch die dem Datenmanagement zugrunde liegenden Prozesse. Das entstandene Datenlebenszyklusmodell ist in Abbildung 8.8 dargestellt.

Aufgrund der fachlichen Fundierung des Modells der Schlüsselkonzepte des Datenmanagements sowie des daraus abgeleiteten Datenlebenszyklusmodells, kann eine hohe Übereinstimmung mit anderen Lebenszyklusmodellen festgestellt werden. Beispielsweise beschreibt *Runkler (2015)* einen vierstufigen Data-Mining-Prozess (vgl. Abbildung 8.9), die *DAMA International (2017)* charakterisiert diesen durch sieben *Key Activities* (vgl. Abbildung 8.10) und *Chisholm (2015)* teilt ihn ebenfalls in sieben Phasen ein (vgl. Abbildung 8.11). Vergleicht man exemplarisch das Modell von *Chisholm (2015)* mit dem hier aus den Praktiken des Datenmanagements abgeleiteten Modell, ist eine klare Überschneidung erkennbar: Während die Phasen „data capture“, „data archival“ und „data purging“ im hier entwickelten Modell äquivalent auftreten, entspricht die *maintenance*-Phase bei Chisholm seiner Beschreibung folgend hier den Phasen *Bereinigung*, *Modellierung*, *Implementierung* und *Optimierung*, sodass das hier entwickelte Modell diese Phase weiter ausdifferenziert. Die *synthesis* befasst sich laut Chisholm insbesondere mit dem Ziehen logischer Schlussfolgerungen

[37] Die übliche Reihenfolge, in der die Praktiken zutage treten, wurde bereits in der Darstellung des Modells der Schlüsselkonzepte des Datenmanagements berücksichtigt.

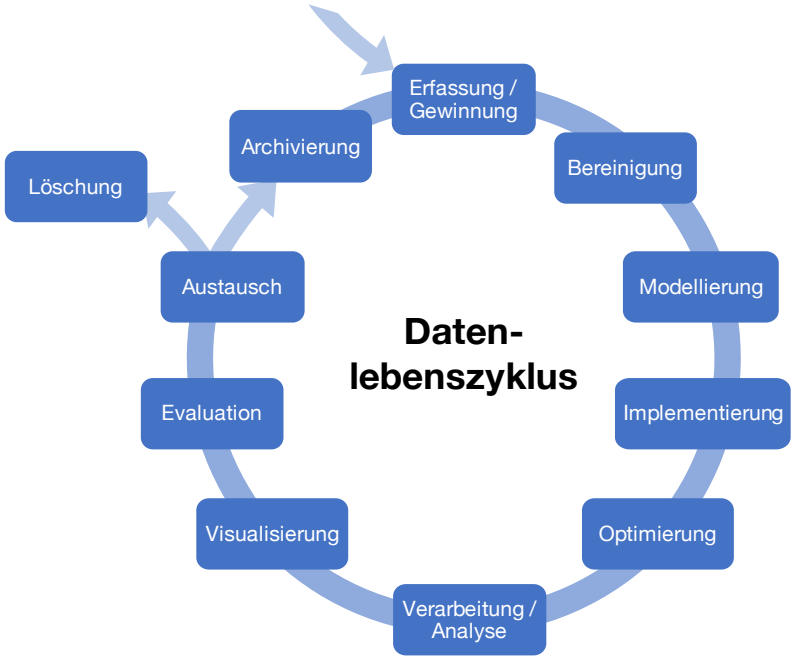


Abbildung 8.8: Interpretation der Praktiken des Datenmanagements als Datenlebenszyklus.

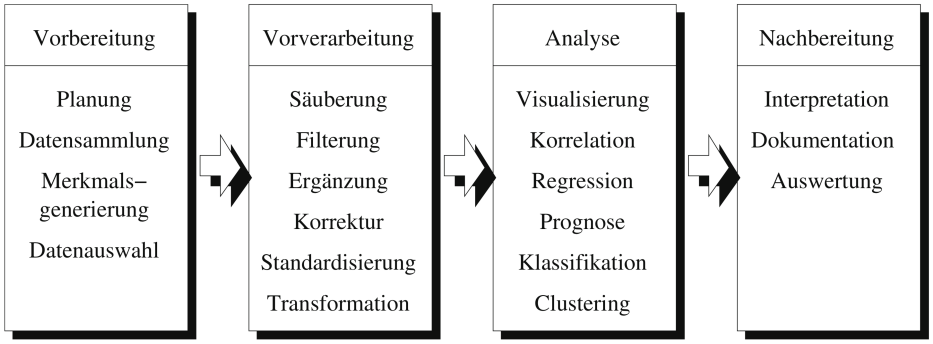


Abbildung 8.9: Vierstufiger Data-Mining-Prozess nach *Runkler (2015)*.

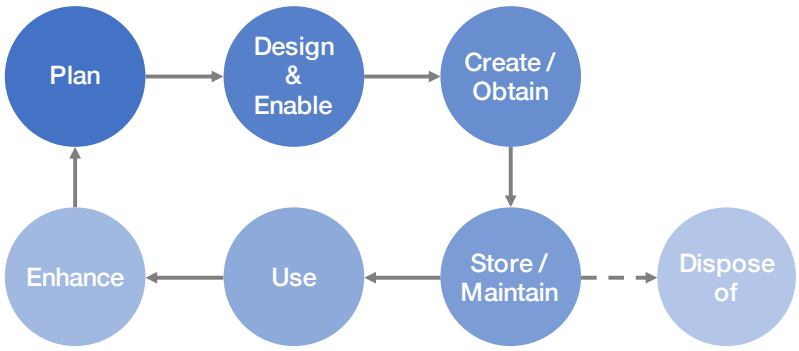


Abbildung 8.10: Sieben *Key Activities* des Datenmanagements nach *DAMA International (2017)*.

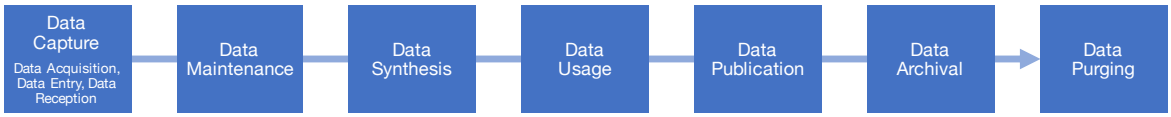


Abbildung 8.11: Sieben Phasen des Datenlebenszyklus nach *Chisholm (2015)*.

aus den Daten und ist daher hier, gemeinsam mit den zukünftigen Nutzungen von Daten und Analyseergebnissen, die Chisholm in der Phase „data usage“ vereint, in der Phase *Verarbeitung/Analyse* berücksichtigt, weist aber auch Bezüge zur *Visualisierung*, *Evaluation* und zum *Austausch* auf.

Das resultierende Lebenszyklusmodell deckt sich in vielerlei Hinsicht mit dem beispielhaft verglichenen Modell nach *Chisholm (2015)*, aber auch mit anderen aus der professionellen Nutzung von Datenmanagement oder aus der Forschung stammenden Varianten. Trotzdem zeigt sich eine andere Schwerpunktsetzung: Durch stärkere Ausdifferenzierung der für die schulische Thematisierung von Datenmanagement relevanten Phasen wird die Nutzung als Leitlinie für den Informatikunterricht wesentlich erleichtert und insbesondere durch detailliertere Charakterisierung der Nutzung von Daten auch konkretere Aspekte, die im Unterricht umsetzbar sind, hervorgehoben. Das entstandene Modell kann daher einen klaren Einblick in den Umgang mit und in die Verwendung von Daten geben, sowohl im privaten als auch im professionellen Umfeld. Gleichzeitig stellt es jedoch, sowohl für Lehrerinnen und Lehrer, als auch für Schülerinnen und Schüler, eine klare Leitlinie für den Datenmanagementunterricht dar. Es betont dabei die zentralen Phasen des Datenlebenszyklus, gleichzeitig sind aber nicht alle in allen Anwendungsfällen gleich relevant. Entsprechend wird eine Schwerpunktsetzung ermöglicht, zugleich aber das Bewusstsein für Aspekte erhöht, die sonst möglicherweise zu sehr vernachlässigt werden würden, wie zum Beispiel, dass Daten nicht einfach nur gespeichert, sondern eigentlich immer auch analysiert werden, aber auch dafür, dass die Entscheidung was mit Daten nach deren Nutzung passieren soll, nicht völlig außen vor gelassen werden sollte.

8.3.2 Untersuchung von Themen des Datenmanagements auf für diese relevante Schlüsselkonzepte

Eine weitere Möglichkeit zur Nutzung des Modells in der Praxis besteht darin, verschiedene Themen des Datenmanagements hinsichtlich ihrer Abdeckung von Schlüsselkonzepten zu untersuchen. Diese Betrachtungsweise trägt insbesondere dazu bei, auch als Lehrkraft ein grundlegendes Verständnis dieser Themen zu erlangen bzw. deren zentrale Aspekte zu erkennen, kann aber andererseits auch die Aufbereitung für den Informatikunterricht unterstützen, indem die Schlüsselkonzepte des betrachteten Themengebiets hervorgehoben werden, die auch im Unterricht die zentrale Rolle spielen sollten. Um an verschiedenen Beispielen die Bedeutung dieser Schlüsselkonzepte zu demonstrieren, werden im Folgenden verschiedene in diesem Fachgebiet zentrale und am Anfang dieser Arbeit in Abschnitt 3.2 beleuchtete Themen auf die jeweiligen Schlüsselkonzepte zurückgeführt bzw. durch diese

charakterisiert. Zusätzlich wird erläutert, inwiefern ein Verständnis der Schlüsselkonzepte dazu beiträgt, ein tiefgreifendes Wissen über die jeweiligen Themen aufzubauen. Dabei wird ein Schwerpunkt auf die Entwurfsprinzipien und Mechanismen gelegt, die für einen Einblick in die grundlegende Funktion der Themen zentraler sind als die Praktiken, die eher den Umgang mit Daten in verschiedenen Phasen des Datenlebenszyklus beleuchten. Diese Untersuchung der Themengebiete wird in Tabelle 8.1 zusammengefasst, indem dort ein Überblick über die Abdeckung der verschiedenen Entwurfsprinzipien und Mechanismen in den unterschiedlichen im Folgenden angesprochenen Themen gegeben wird.

Big Data

Big Data wird zumeist durch die sog. *drei V* charakterisiert, *volume, velocity* und *variety (Laney, 2001)*. Unter diesem oft als Modewort verwendeten Begriff wird häufig der gesamte Prozess, der mit der Analyse großer, vielfältiger und schnell wachsender Datenmengen in Zusammenhang steht, verstanden. Entsprechend spielen, je nach Auslegung, alle oder auch nur ein Teil der Praktiken des Datenmanagements eine wichtige Rolle, wobei oft ein Fokus auf die *Gewinnung/Erfassung*, *Modellierung*, *Speicherung/Implementierung* und *Verarbeitung/Analyse* erkennbar ist. Um Big Data zu verstehen, ist jedoch mehr als nur ein Blick auf die Praktiken notwendig: Da umfangreiche Daten gespeichert und möglichst schnell verarbeitet werden sollen, ist ein Verständnis von *Nebenläufigkeit* und den damit einhergehenden Möglichkeiten und Einschränkungen unabdingbar. Aber auch die Verteilung von Daten auf mehrere Datenknoten und die somit nötige *Partitionierung* und *Synchronisation* sind hier, genauso wie der *Transport* der Daten zentral, da im Falle großer Datenmengen eine Speicherung auf nur einem Datenknoten kaum mehr möglich ist. Entsprechend muss auch ein Verständnis darüber, was *Partitionstoleranz* bedeutet, erworben werden, damit die im Falle einer Partitionierung des Gesamtsystems auftretenden Probleme eingeschätzt und gegebenenfalls gelöst werden können. Ohne ein Verständnis der beim Transport notwendigen Maßnahmen und entstehender Probleme (beispielsweise der Verlangsamung einer Analyse durch den Transport und ggf. fehlerhafte Übertragungen), kann die Herausforderung Big Data zu analysieren kaum erkannt werden und wird, wie in der allgemeinen Wahrnehmung dieser Thematik erkennbar, rein auf die Speicherung großer Datenmengen beschränkt betrachtet.

	Big Data	Metadaten	Data Mining	Datenstrom-systeme	Datenbanken allgemein	Datenbanken relational	Datenbanken nicht-relational	Verteilte bzw. cloudbasierte Datenspeicher
Entwurfsprinzipien								
Datenunabhängigkeit					×	×	×	
Integrität		×	×			×		
Konsistenz			×	×		×		
Isolierung						×		
Dauerhaftigkeit					×	×	×	
Verfügbarkeit				×			×	×
Partitionstoleranz	×		×	×			×	
Nebenläufigkeit	×	×	×		×	×	×	
Redundanz						×	×	×
Mechanismen								
Strukturierung		×	×		×	×	×	
Repräsentation		×	×	×	×	×	×	
Replikation		×					×	×
Synchronisation	×	×					×	×
Partitionierung	×	×	×	×			×	×
Transport	×	×		×			×	×
Transaktion		×				×		

Tabelle 8.1: Übersicht über die unterschiedliche Relevanz bzw. Abdeckung der Schlüsselkonzepte für verschiedene Themen des Datenmanagements.

Metadaten

Metadaten sind für den Umgang mit Daten in allen Phasen des Datenlebenszyklus unabdingbar. Sie umfassen strukturierende Informationen (beispielsweise Beziehungen zwischen verschiedenen Datensätzen) und ergänzen die Daten durch Hintergrundinformationen (wie den Aufnahmeort eines Fotos) sowie aus administrativen Gründen notwendige Informationen (beispielsweise Zugriffsrechte). Entsprechend können Metadaten nicht unabhängig von ihrem jeweiligen Einsatzzweck betrachtet werden, im Gegenteil ist deren Bedeutung nur in diesem Kontext gegeben. Um zu verstehen, was Metadaten sind, müssen deren unterschiedliche Funktionen genauer betrachtet werden:

- Administrative Metadaten liefern Informationen darüber, wie am Datenspeicher *repräsentierte* Daten zu interpretieren sind. Solche administrativen Metadaten werden aber auch eingesetzt, um *Synchronisation, Replikation* und *Partitionierung* zu ermöglichen, den *Transport* von Daten abzusichern und *Transaktionen* zu verwalten. Sie stellen damit eine unabdingbare Basis für viele der Mechanismen des Datenmanagements dar.

- Strukturierende Metadaten werden hingegen hauptsächlich zur *Strukturierung* eingesetzt und sind insbesondere notwendig, um die Sicherstellung von *Integrität* und durch *Transaktionen* eine *nebenläufige* Arbeit mit einem Datenspeicher zu ermöglichen, indem diese die Struktur der Daten und somit ihre gegenseitigen Abhängigkeiten explizieren.

- Um Daten maschinell verarbeitbar zu machen, die in übliche Strukturen nicht ohne weiteres abbildbar sind, aber auch um zusätzliche Informationen über deren Bedeutung und Herkunft mitzuliefern, die über administrativ notwendige Informationen hinausgehen, werden deskriptive Metadaten eingesetzt. Sie können daher insbesondere zur adäquaten *Repräsentation* und zur *Strukturierung* eingesetzt werden.

Um zu verstehen, was Metadaten ausmacht, ist es essenziell, die möglichen Funktionen von Metadaten und entsprechend deren Notwendigkeit zur Umsetzung verschiedenster Prinzipien des Datenmanagements zu verstehen:

- Nur ein Verständnis über Repräsentation macht beispielsweise deutlich, warum Informationen zur Interpretation dieser gespeicherten Daten in Form von Metadaten zur Verfügung gestellt werden müssen.

- Ein Verständnis der Grundlagen des Transports ermöglicht es hingegen, zu erkennen, warum Metadaten notwendig sind, um beispielsweise das Ziel eines Datenpaketes zu kennzeichnen oder eine Überprüfung seiner Unversehrtheit durchführen zu können. Gleichzeitig hilft es auch, den Wert von Metadaten im Kontext von Datenanalysen zu verstehen und zu erkennen, warum diese Informationen im Gegensatz zu den eigentlichen Daten nur in Spezialfällen sicher verschlüsselt und so vor Angreifern versteckt werden können.

- Die Grundlagen der Strukturierung zu verstehen ermöglicht es hingegen, zu erkennen, warum Daten zusätzliche Informationen mitgeliefert werden, die diese beispielsweise durch Stichworte oder ähnliches charakterisieren, um eine schnelle Auffindbarkeit zu ermöglichen.

Entsprechend können die Vielfalt, Notwendigkeit und der Nutzen von Metadaten erkannt werden, indem diese durch die Brille der Entwurfsprinzipien und Mechanismen des Datenmanagements betrachtet werden. Der Umfang dieser Daten verdeutlicht aber auch, dass die andauernd und in allen Lebenssituationen erzeugten Metadaten gegebenenfalls spannende Quellen für Datenanalysen darstellen und somit – selbst wenn die eigentlichen Daten gut geschützt werden – zur Gefahr für die eigene Privatsphäre werden können.

Data Mining

Data Mining beschäftigt sich insbesondere mit der Gewinnung neuer Informationen aus einem großen Berg an Daten, die oft für andere Zwecke erfasst worden sind. Es handelt sich entsprechend um eine spezielle Form der Datenanalyse, die heute immer stärker an Bedeutung gewinnt. Um zu verstehen, wie solche Analysen funktionieren, ist es zum einen notwendig, den Unterschied zwischen Information und Daten zu verstehen und somit den Mechanismus *Repräsentation*, der die Abbildung von Informationen in Daten beschreibt und entsprechend den umgekehrten Prozess zur Analyse darstellt. Gleichzeitig muss aber auch Sinn und Aufgabe der *Strukturierung* verstanden werden, damit deutlich werden kann, wie verschiedene Datensätze zusammenhängen und wie deren Beziehungen zu Analysezwecken ausgenutzt werden können. Um eine hohe Analysequalität sicherzustellen ist es außerdem essenziell, *Integrität* und *Konsistenz* der analysierten Daten (bzw. allgemein deren Qualität) zu berücksichtigen und Möglichkeiten zur Realisierung aber auch Grenzen dieser beiden Entwurfsprinzipien zu verstehen. Da Data Mining außerdem üblicherweise mit sehr umfangreichen Datenmengen stattfindet, die in angemessener Zeit oft nicht mehr ohne Parallelisierung des Analyseprozesses analysiert werden können und die oft von mehreren Datenknoten abgefragt werden müssen, sind insbesondere auch Grundkenntnisse im Bereich der Entwurfsprinzipien *Nebenläufigkeit* und *Partitionstoleranz* bzw. des Mechanismus *Partitionierung* notwendig.

Datenstromsysteme

Datenstromsysteme befassen sich mit Datenanalysen, deren Ergebnisse in kurzer Zeit (idealerweise Echtzeit) verfügbar sein sollen. Für solche Analysen steht die *Verfügbarkeit* im Vordergrund, sodass ein Verständnis dieses Entwurfsprinzips eine zentrale Rolle spielt, um beurteilen zu können, welche Maßnahmen getroffen werden müssen, um diese zu erhöhen. Wie das CAP-Theorem verdeutlicht, steht diese Verfügbarkeit in Konkurrenz zu *Konsistenz* und *Partitionstoleranz*, weswegen Datenstromsysteme vermeiden, Daten dauerhaft zu speichern. Ein Verständnis dieser zentralen Entscheidung basiert daher auch auf

dem Wissen über diese Prinzipien, aber auch über die Mechanismen *Partitionierung* und *Transport*. Gleichzeitig muss jedoch hervorgehoben werden, dass eine Speicherung nie völlig vermieden werden kann, sodass *Repräsentation* auch in diesem Fall eine Rolle spielt, jedoch mit einem Fokus darauf, wie diese möglichst speicher- und zeiteffizient stattfinden kann.

Relationale und nichtrelationale Datenbanken

Datenbanken stellen in zentralen Aspekten das genaue Gegenteil zu Datenstromsystemen dar: Deren Ziel ist nicht die einmalige schnelle Analyse, sondern die dauerhafte und gut strukturierte, effiziente Speicherung großer Datenmengen. Entsprechend steht hier die sinnvolle und adäquate *Strukturierung* und *Repräsentation* der Daten klar im Vordergrund. Gerade wenn mehrere Datenbankmodelle, beispielsweise relational und nichtrelational (bspw. dokumentenorientiert), gegeneinander abgewogen werden müssen, sind Kenntnisse über die unterschiedliche Art der Strukturierung aber auch deren unterschiedliche Weise Daten im Endeffekt zu speichern, unvermeidbar. Eine charakterisierende Eigenschaft dieser Systeme ist die *Datenunbhängigkeit*, die im Sinne eines Schichtenmodells von den eigentlich gespeicherten Daten und deren Struktur abstrahiert, und auf höherer Ebene eine auf logischer statt technischer Ebene angesiedelte Strukturierung ermöglichen. Das Verständnis von Datenunabhängigkeit ist für den kompletten Umgang mit Datenbanken zentral, da nur dadurch weitere Möglichkeiten, wie beispielsweise ein *nebenläufiger Zugriff* und *Dauerhaftigkeit*, möglich werden. Nur durch Thematisierung dieser Entwurfsprinzipien, die in verschiedenen Datenbanksystemen unterschiedlich umgesetzt werden, kann ein klares Bild von Datenbanken gezeichnet und deren Sinn und Zweck deutlich gemacht werden.

Während diese Aspekte allen Datenbanken gemein sind, setzen relationale und nichtrelationale Datenbanken jeweils weitere unterschiedliche Schwerpunkte: So ist beispielsweise bei relationalen Datenbanken die *Konsistenz* und *Integrität* des Datenbestands zentral. Um Diese beiden Eigenschaften sicherzustellen wird versucht, *Redundanzen* zu vermeiden und es werden *Transaktionen* zur *Isolierung* verschiedener Abfragen eingesetzt. Nichtrelationale Datenbanken legen hingegen den Schwerpunkt auf die *Verfügbarkeit* der Daten, wozu insbesondere die *Replikation* beiträgt, wodurch auch deren *Transport* und *Synchronisation* notwendig werden. Gleichzeitig steigt aber auch die Bedeutung der *Partitionstoleranz*, wie bei allen verteilten Datenspeichern. Außerdem werden *Redundanzen* – im Gegensatz zu relationalen Datenbanken, die diese eher vermeiden – gezielt eingesetzt, um die Verfügbarkeit zu steigern, wodurch Konsistenz und Integrität eingeschränkt werden.

Verteilte und Cloud-Datenspeicher

Verteilte und Cloud-Datenspeicher sind allgemeiner ausgestaltet als Datenbanken und Datenstromsysteme und stärker auf die Datenspeicherung an sich fokussiert. Für diese Systeme ist es zentral, Daten nicht nur auf einem einzelnen Datenspeicher vorzuhalten, sondern

verteilt auf mehreren. Um die mit diesen Systemen verbundenen Phänomene, wie beispielsweise Synchronisationskonflikte, gleichzeitig aber auch deren Ursachen verstehen zu können, muss ein grundlegendes Verständnis der dahinterstehenden Mechanismen *Replikation*, *Synchronisation* und *Transport* erworben werden, die der Auslöser für die meisten auftretenden Phänomene sind. Entsprechend ist auch ein Verständnis verschiedener Entwurfsprinzipien zentral, insbesondere der *Partitionstoleranz*, die dann wichtig wird, wenn Teile eines verteilten Datenspeichers ausfallen. Zu wissen, wie sich ein genutztes System in solchen Fällen verhält, kann dazu beitragen, Datenverluste zu vermeiden oder rechtzeitige Vorsichtsmaßnahmen für diesen Fall zu ergreifen. Doch auch *Redundanz* ist beim Einsatz verteilter Systeme ein zentrales Prinzip, das teils Ursache für Probleme (beispielsweise Inkonsistenzen) ist, aber auch gezielt eingesetzt wird, um die Zugriffsgeschwindigkeit der Datenspeicher zu steigern und/oder die *Verfügbarkeit* bei Ausfällen von Teilsystemen sicherzustellen. Entsprechend betont die Betrachtung dieser Systeme insbesondere Aspekte, die beispielsweise auch relationale und nichtrelationale Datenbanken unterscheiden.

Zusammenfassung

Wie diese Beispiele zeigen, sind die ermittelten Schlüsselkonzepte des Datenmanagements zentral um verschiedene Themen und Entwicklungen aus dem Fachgebiet auf den jeweiligen Kern zu reduzieren und deren grundsätzliche Funktionsweise und damit einhergehende Phänomene einordnen und verstehen zu können. Durch die Explikation der Schlüsselkonzepte des Datenmanagementunterrichts kann somit ein konzeptorientierter Unterricht gefördert werden, der trotz der Thematisierung unterschiedlicher Themen ein regelmäßiges Wiederaufgreifen derselben Konzepte ermöglicht und somit im Sinne eines Spiralcurriculums einerseits auf das bereits von den Schülerinnen und Schülern erworbene Wissen aufbaut und dieses vertieft, andererseits aber auch zur Vernetzung des Wissens beiträgt. Entsprechend kann das vorgestellte Modell der Schlüsselkonzepte des Datenmanagements als Werkzeug bei der Planung und Entwicklung aber auch bei der Analyse von Unterrichtskonzepten und Curricula hilfreich eingesetzt werden und zur Entscheidung beitragen, welche Themen in welchem Umfang im Unterricht betrachtet werden.

8.3.3 Charakterisierung zentraler Konzepte der Informatik aus Perspektive des Datenmanagements

Die Schlüsselkonzepte des Datenmanagements sind jedoch nicht nur in Bezug auf Themen des Fachgebiets relevant, stattdessen tragen sie auch zur Sicherstellung der Erfüllung verschiedener übergeordneter Konzepte der Informatik bei, wie im Folgenden gezeigt wird. Bei der Charakterisierung des Begriffs *Computer Science* in der ACM Encyclopedia of Computer Science *(Denning, 2003a)* werden verschiedene Konzepte genannt, welchen in der Informatik im Allgemeinen eine übergeordnete Bedeutung zukommt: Beispielsweise beschäftigt sich Informatik in allen Fachgebieten mit *Nutzbarkeit*, *Zuverlässigkeit* und *Sicherheit*. Obwohl aufgrund ihrer Bedeutung für die gesamte Informatik erwartet wer-

den kann, dass sich diese Konzepte auch im Datenmanagement widerspiegeln, sind die drei Begriffe nicht explizit im vorgestellten Modell enthalten. Dies scheint insbesondere für Sicherheit unerwartet, da diese eines der zentralen Themen bei der Thematisierung von Daten im gesellschaftlichen Diskurs darstellt. Im Rahmen der durchgeführten Analyse konnten verschiedene Aspekte dieser Konzepte jedoch identifiziert werden, sodass sie auch in den Ergebnissen der ersten Analysephase enthalten waren. Während der Kondensation zum Modell der Schlüsselkonzepte wurden diese Begriffe jedoch durch andere subsumiert, da beispielsweise Sicherheit im Datenmanagement durch verschiedene Mechanismen und Prinzipien konkretisiert wird. Entsprechend haben diese Konzepte durchaus Bedeutung im Datenmanagement, sie manifestieren sich im betrachteten Feld jedoch durch spezifischere und auch im Modell enthalte Begriffe. Da gerade die übergeordneten Ziele der Informatik auch oft zur Begründung ihres Allgemeinbildungscharakters herangezogen werden, ist es für die praktische Nutzung des Modells daher eine interessante Aufgabe, zu untersuchen wie Datenmanagement zu den übergeordneten Zielen beiträgt und somit Aspekte des Fachgebiets zu identifizieren, die aus diesem Grund besonders hervorzuheben sind. Somit wird die Verankerung der identifizierten Schlüsselkonzepte in der Informatik expliziert und mögliche Anknüpfungspunkte hervorgehoben. Dies wird im Folgenden beispielhaft für *Datensicherheit* und *Nutzbarkeit* untersucht.

Sicherheit bezeichnet in der Informatik den Schutz von Systemen vor Manipulation, Ausfällen und Störungen. Um Sicherheit detaillierter zu betrachten, können die angestrebten Schutzziele bzw. Grundwerte der IT-Sicherheit betrachtet werden: Meist werden dabei mindestens drei Ziele definiert, *Vertraulichkeit*, *Integrität* und *Verfügbarkeit (vgl. Bundesamt für Sicherheit in der Informationstechnik, 2012)*. Integrität und Verfügbarkeit stellen dabei direkt Schlüsselkonzepte des Datenmanagements dar, da diese ein zentrales Ziel von und oft auch ein wichtiger Grund für die Nutzung von Datenmanagementsystemen sind, weswegen sie sich besonders deutlich in diesem Fachgebiet widerspiegeln. Das dritte Schutzziel ist hingegen nicht direkt im Modell erkennbar: Die *Vertraulichkeit* von Daten schließt den Schutz vor unberechtigtem Zugriff und Diebstahl von Daten ein, was im Datenmanagement nicht als eigenes Konzept realisiert, sondern auf Basis anderer Konzepte ermöglicht wird. Insbesondere die geeignete *Strukturierung* von Daten (damit eine sinnvolle Rechtevergabe und Sichteneinschränkung möglich wird) ist dabei genauso zentral wie die Abstraktion der dem Benutzer zur Verfügung gestellten Schnittstelle von den eigentlichen Daten (*Datenunabhängigkeit*), durch die eine Kontrolle erst ermöglicht wird. Gleichzeitig müssen – je nach angestrebtem Grad an Vertraulichkeit – auch beim *Transport* und der *Repräsentation* von Daten verschiedene Grundsätze beachtet werden, beispielsweise indem Möglichkeiten zur jeweils geeigneten Verschlüsselung der zu schützenden Daten genutzt werden. Diese sind jedoch keine originären Aspekte des Datenmanagements und werden daher im Modell nicht gezielt thematisiert. Auch zu den ersten beiden schon direkt im Modell ersichtlichen Schutzzielen tragen noch weitere Schlüsselkonzepte des Datenmanagements bei: *Integrität* zielt darauf ab, im System gespeicherte Daten korrekt und unverändert vorzuhalten, sodass keine fehlerhaften Daten zurückgeliefert werden. Dazu tragen auch jegliche Maßnahmen zur Sicherstellung von *Konsistenz* und *Dauerhaftigkeit* sowie *Replikation* von Daten bei. Bei

der *Verfügbarkeit* geht es hingegen darum, sicherzustellen, dass ein Datenmanagementsystem und die darin gespeicherten Daten den Anwendern jederzeit in nutzbarer Form zur Verfügung stehen. Dies wird im Datenmanagement insbesondere durch die *Replikation* von Daten und deren *Synchronisation* erreicht, wodurch eine höhere Anzahl an Rechen- bzw. Datenknoten im Datenmanagementsystems zur Verfügung steht und die Daten somit näher am Nutzer gelagert und schneller zur Verfügung gestellt werden können. Gleichzeitig tragen auch *Transaktionen*, *Isolation* und *Nebenläufigkeit* deutlich zur Steigerung der Verfügbarkeit bei, da dadurch erst ein Mehrbenutzerbetrieb ohne regelmäßige und lang andauernde Wartezeiten ermöglicht wird.

Somit zeigt sich eindeutig, dass das Konzept *Sicherheit* im Datenmanagement auf vielfältige Weise repräsentiert und somit auch im Modell der Schlüsselkonzepte des Datenmanagements enthalten ist. Die Zusammenhänge zwischen Sicherheit und den im Modell repräsentierten Schlüsselkonzepten werden zusammenfassend in Abbildung 8.12 dargestellt.

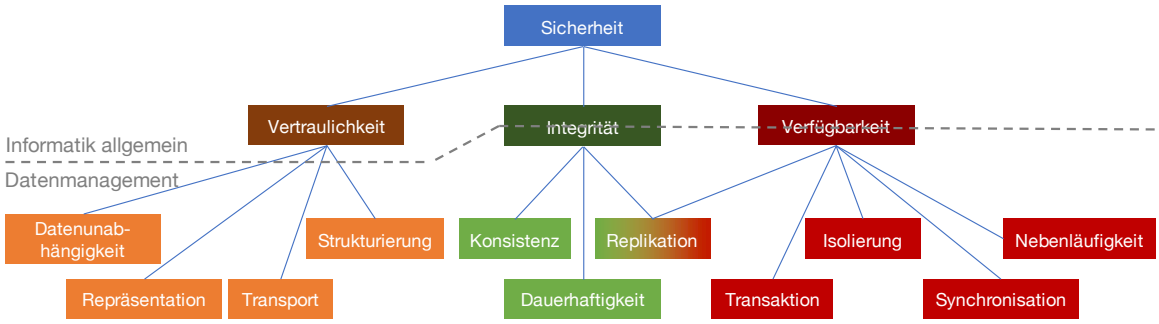


Abbildung 8.12: Das Konzept *Sicherheit* im Datenmanagement.

Nutzbarkeit befasst sich mit einer Verbesserung des Nutzungserlebnisses einer Anwendung oder eines Systems, mit dem Ziel Anwendern eine einfache, effiziente und verständliche Nutzung zu ermöglichen. Im Datenmanagement wird eine gute Nutzbarkeit insbesondere dadurch erreicht, dass der Zugriff auf bzw. die Speicherung von Daten möglichst einfach gehalten wird und somit eine flexible Nutzung der Datenmanagementsysteme ermöglicht wird. Dies wird insbesondere durch *Datenunabhängigkeit* erreicht, da somit die Anwenderschnittstelle unabhängig von der internen Repräsentation der Daten wird. Dies ermöglicht nicht nur einen Wechsel der technischen Infrastruktur ohne die Anwenderschnittstelle verändern zu müssen, sondern vereinfacht auch den Umgang mit dem System wesentlich, da der Nutzende keinerlei Informationen über die internen Eigenschaften des genutzten Systems haben muss. Gleichzeitig trägt jedoch auch die Erfüllung weiterer Entwurfsprinzipien zur Erhöhung der Nutzbarkeit bei: Insbesondere die *Verfügbarkeit* ist aus Nutzersicht eine zentrale Eigenschaft eines Systems, aber auch die *Konsistenz*, die es ermöglicht sich auf den gespeicherten Datenbestand zu verlassen und mögliche Probleme frühzeitig zu erkennen. Durch *Isolation*, *Nebenläufigkeit* und *Transaktionen* wird außerdem ein, heute in sehr vielen Fällen zentraler, Mehrbenutzerbetrieb ermöglicht und unerwünschte Seiteneffekte vermieden.

Auch die Nutzbarkeit spielt daher im Datenmanagement eine zentrale Rolle. Diese wird in Abbildung 8.13 visualisiert.

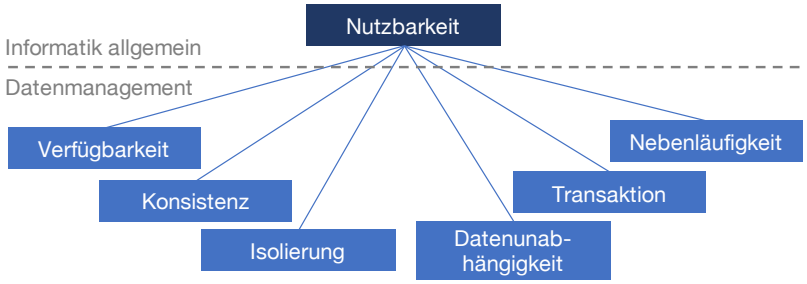


Abbildung 8.13: Das Konzept *Nutzbarkeit* im Datenmanagement.

Zusammenfassung Die Betrachtung von Datenmanagement aus der Perspektive zentraler Konzepte der Informatik zeigt, wie in den beiden Beispielen verdeutlicht, dass Datenmanagement auf verschiedene Weise zur Erreichung der mit diesen Konzepten einhergehenden Ziele beiträgt und diese Konzepte im Modell der Schlüsselkonzepte von Datenmanagement auch (implizit) berücksichtigt sind. Daher kann Datenmanagement durch Adressierung von Themen wie Sicherheit oder Nutzbarkeit aus einer weiteren Perspektive auch dazu beitragen, den Blick auf die zentralen Aspekte dieser Themen zu schärfen und ein besseres Verständnis zu fördern. Gleichzeitig kann aber auch das Wissen, das Schülerinnen und Schüler zu solchen übergeordneten Zielen mitbringen, genutzt werden, um Themen wie Integrität oder Verfügbarkeit entsprechend zu verankern und das Verständnis für solche Schlüsselkonzepte des Datenmanagements zu fördern.

8.4 Diskussion der Methodik und des entwickelten Modells

Das in diesem Kapitel entwickelte und vorgestellte Modell charakterisiert Datenmanagement durch vier Perspektiven, die *Kerntechnologien*, *Praktiken*, *Entwurfsprinzipien* und *Mechanismen*. Gleichzeitig werden durch diese vier Dimensionen Zusammenhänge zwischen den zentralen Sichtweisen auf Datenmanagement herausgestellt. Damit kann dieses einfach erfassbare Modell einerseits von fachlicher Seite einen wichtigen, klaren und strukturierten Überblick über Datenmanagement geben. Somit macht es dieses Fachgebiet auch für fach(gebiets)fremde Personen, wie beispielsweise Lehrkräfte, besser greifbar und zugänglich und verdeutlicht dessen Breite, betont durch die vier verschiedenen beleuchteten Bereiche aber auch dessen Tiefe. Im Vergleich zu älteren Arbeiten offenbart das Modell die deutliche Weiterentwicklung in den letzten Jahren und Jahrzehnten: Während Lockemann *Lockemann (1986)* noch die drei Aspekte *Konsistenz*, *Permanenz* und *Konkurrenz* als zentral für das damalige Gebiet Datenbanken hervorhob, kann Datenmanagement heute aufgrund seiner zunehmenden Breite nur noch durch eine wesentlich größere Auswahl an Schlüsselkonzepten ausreichend charakterisiert werden.

Das entwickelte Modell und dessen Diskussion mit anderen Wissenschaftlern zeigen, dass die gewählte Methodik zielführend auf das Fachgebiet anwendbar war. Dabei offenbaren sich verschiedene Vorteile dieses in der ersten Phase empirisch geprägten Ansatzes gegenüber anderen Arbeiten mit ähnlichen Zielen. Während die meisten Arbeiten im Bereich der Ideen, Konzepte und Prinzipien aufgrund ihres methodischen Vorgehens verschiedene Gütekriterien der empirischen Forschung nur eher eingeschränkt erfüllen können, trägt die hier angewandte Methodik zu höherer Objektivität, aber auch einer höhere Reliabilität und Validität bei. Der Grad an Erfüllung dieser drei Kriterien wird im Folgenden kurz skizziert:

Objektivität. Gerade in der ersten Analysephase wurden durch die systematische qualitative Inhaltsanalyse fachwissenschaftlicher Literatur und die zusätzliche semiautomatische Analyse einer Vielzahl weiterer Texte aus dem Fachgebiet, subjektive Einflüsse des Autors minimiert. Trotzdem können solche nicht komplett ausgeschlossen werden, da insbesondere in der zweiten Phase der Analyse eine manuelle Strukturierung und Einordnung der zuvor gefundenen Begriffe in das ausgewählte Modell und die Zusammenfassung verwandter Begriffe nötig war, wobei interpretative Aspekte teilweise nicht ausbleiben. Durch die Definition klarer Zuordnungskriterien für die verschiedenen Modellbereiche und die Diskussion im Expertenkreis wurde jedoch auch an dieser Stelle versucht, diese Einflüsse soweit wie möglich zu minimieren.

Reliabilität. Gleichzeitig sorgt der gewählte Ansatz für eine verhältnismäßig hohe Reliabilität. Dazu trägt im ersten Schritt die klare Vorgehensweise der qualitativen Inhaltsanalyse bei, aber auch wieder die teilautomatisierte Validierung der Ergebnisse. Einschränkungen der Reliabilität sind daher eher in der zweiten als der ersten Phase zu erwarten, wobei versucht wurde, diese auch an dieser Stelle durch die Nutzung geeigneter Kriterien zu vermeiden und Entscheidungen wie das Weglassen bzw. Zusammenführen von Begriffen nur bei klar möglicher Begründung zuzulassen. Nichtsdestotrotz lässt die Methodik gegebenenfalls unterschiedliche Ergebnisse zu, wodurch die Reliabilität eingeschränkt wird. Diese Ergebnisse müssen einander aber nicht widersprüchlich sein; stattdessen ist zu erwarten, dass es sich dabei (solange keine klaren Weiterentwicklungen im Fachgebiet stattgefunden haben) eher um eher geringfügige Abweichungen handelt, die kaum vermeidbar sind.

Validität. Die Validität des Modells wurde in mehreren Schritten überprüft und kann somit als relativ hoch angenommen werden. Insbesondere geschah dies durch die semiautomatisierte Analyse weiterer Literaturquellen in der ersten Analysephase, wodurch eine subjektive Beeinflussung des ursprünglichen Literaturkanons ausgeschlossen werden konnte. Auch die Diskussion der Ergebnisse der zweiten Phase mit Experten und die damit verbundene Einbeziehung der Perspektiven dieser Personen trägt zu einer Sicherung der Validität der Ergebnisse bei.

Insgesamt kann daher, insbesondere im Vergleich mit ähnlich ausgerichteten Arbeiten zur Charakterisierung der Informatik oder eines ihrer Teilbereiche, eine verhältnismäßig gute Erfüllung der Gütekriterien wissenschaftlicher Arbeit angenommen werden. Somit trifft einer der zentralen Kritikpunkte, der beispielsweise in der Diskussion um fundamentale Ideen genannt wurde, auf das hier entwickelte Modell nicht zu: *Baumann (1998)* kritisiert insbesondere die starke Subjektivität derartiger Ansätze. Aber auch die oft hierarchische Anordnung *(Baumann, 1998)* und die zu große Anzahl an Ideen *(Modrow, 2003; Modrow und Strecker, 2016; Baumann, 1998)* werden oft kritisiert. Auch diese Kritikpunkte werden im vorgestellten Modell aufgegriffen, indem einerseits eine Hierarchisierung nicht stattfindet und die verschiedenen Bereiche des Modells gleichwertig nebeneinanderstehen, andererseits auch, indem versucht wurde, die Menge der Begriffe durch Zusammenfassung auf ein Minimum zu reduzieren. Eine weitere Reduzierung würde, wie zuvor exemplarisch gezeigt, zu einem deutlichen Verlust an Details führen und dementsprechend die Verständlichkeit des Modells einschränken.

Das Modell weist aufgrund des gewählten Aufbaus klare Gemeinsamkeiten mit den *Great Principles of Computing* auf und kann als Spezialisierung dieses Modells auf das Fachgebiet Datenmanagement verstanden werden. Insbesondere wurde nicht nur der Aufbau des Modells übernommen, sondern es kann auch festgestellt werden, dass Teile der ermittelten Schlüsselkonzepte eine Spezialisierung der von Denning vorgestellten Great Principles darstellen: Beispielsweise wird das von Denning genannte Entwurfsprinzip *Sicherheit* im Datenmanagement durch die Entwurfsprinzipien *Integrität*, *Verfügbarkeit* und *Isolierung* konkretisiert. Gleichzeitig sind verschiedene andere hier ermittelte Schlüsselkonzepte des Datenmanagements Konkretisierungen der von Denning berücksichtigten *recollection*. Auch mit den fundamentalen Ideen der Informatik nach Schwill können gewisse Gemeinsamkeiten gefunden werden, da bei verschiedenen der hier vorgestellten Schlüsselkonzepte vermutet werden kann, dass sie bei Betrachtung der Kriterien, die *Schwill (1993)* an fundamentale Ideen der Informatik anlegt, zugleich prototypisch für fundamentale Ideen des Datenmanagements stehen:

- Aufgrund ihrer breiten Relevanz im Fachgebiet und sogar darüber hinaus wird bei allen Schlüsselkonzepten das *Horizontalkriterium* erfüllt.
- Wie in den Kapiteln 11 bis 12 noch gezeigt werden wird, können verschiedene der ermittelten Konzepte auf verschiedenen intellektuellen Niveaus vermittelt werden, sodass das *Vertikalkriterium* erfüllt wird.
- Durch den Alltagsbezug verschiedener Konzepte wird das *Sinnkriterium* adressiert (vgl. auch Beschreibungen in Anhang C).
- Obwohl das Modell der Motivation entstammt, die umfangreichen Veränderungen im Fachgebiet Datenmanagement greifbarer zu machen, ist mindestens ein Großteil der ermittelten Schlüsselkonzepte schon längerfristig im Fachgebiet relevant und wird es vermutlich auch zukünftig sein, sodass das *Zeitkriterium* erfüllt wird.

- Das *Zielkriterium* wird dadurch adressiert, dass alle Schlüsselkonzepte jeweils auf die Erreichung eines konkreten und oft idealisierten Ziels hin ausgerichtet sind, dies wird insbesondere bei den Entwurfsprinzipien deutlich.

Trotzdem unterscheiden sich die dargestellten Schlüsselkonzepte von Ideen: Im Sinne der platonschen Ideenlehre stellen Ideen die (nur geistig vorhandenen) Urbilder dar, die realweltliche Phänomene beschreiben und aus denen solche entstehen. Damit sind die hier vorgestellten (Schlüssel-)Konzepte auf einer deutlich konkreteren und weniger idealisierten Ebene angesiedelt.

Während Informatikunterricht im Bereich der *Daten* sich aus traditionellen Gründen bisher eher auf Datenbanken konzentriert, verdeutlichen die Schlüsselkonzepte die Vielfalt des diesem Unterricht zugrundeliegenden Fachgebiets. Durch die strukturierte Darstellung kann das Modell einen deutlichen Beitrag zur Curriculums- und Unterrichtsgestaltung liefern, da es eine Charakterisierung aus fachlicher Sicht ermöglicht, Lücken zwischen der fachwissenschaftlichen Sicht und der Umsetzung in Bereich der Unterrichtsplanung deutlich macht und einen klaren Fokus auf die Schlüsselkonzepte des Fachgebiets nahelegt. Auf der anderen Seite kann das Modell, wie beispielsweise auch die fundamentalen Ideen nach *Schwill (1993)*, als Relevanzfilter eingesetzt werden. Dieser kann insbesondere genutzt werden, um zu verhindern, dass eher unwichtige und/oder zeitlich unbeständige Aspekte, die auf den ersten Blick möglicherweise attraktiv scheinen, eine zu hohe Bedeutung im Informatikunterricht erlangen. Gleichzeitig ergeben sich aber noch verschiedene weitere unterrichtliche Vorteile: Durch das Modell wird das Fachgebiet strukturiert und der Kompetenzerwerb konkretisiert, die Schlüsselkonzepte liefern aber auch konkrete Begriffe, die den fundierten Gebrauch der Fachsprache unterstützen. Auch die drei Gründe, aus denen laut Denning die Great Principles interessant sind *(Denning, 2004)*, können auf das Modell der Schlüsselkonzepte des Datenmanagements übertragen werden:

- Die *Verständlichkeit* wird durch die Fokussierung auf die Schlüsselkonzepte des Datenmanagements erhöht. Zugleich wird das Fachgebiet greifbarer und es wird einfacher möglich, einen klaren Überblick über dieses zu bekommen.

- Das Modell hebt nicht nur Prinzipien des Fachgebiets, sondern insbesondere auch die Praktiken des Datenmanagements deutlich hervor. Zusätzlich gibt es damit auch einen klaren *Einblick* in die professionelle Speicherung, Verwaltung und Nutzung von Daten.

- Gleichzeitig trägt das Modell auch dazu bei, das *Bild der Informatik* im Allgemeinen und von Datenmanagement im Speziellen zu prägen. Es verdeutlicht damit die Breite von Informatik und die Möglichkeiten, die Informatik und Datenmanagement heute bieten. Dabei stellt es insbesondere einen der Bereiche der Informatik in den Vordergrund, der zwar in der gesellschaftlichen Wahrnehmung immer wichtiger wird und unser Leben immer stärker durchdringt, der aber im Vergleich zu anderen Bereichen, die eher nahe an der Softwareentwicklung angesiedelt sind, weniger oft mit Informatik assoziiert wird.

Die vorgestellte Methodik zur Ermittlung der Schlüsselkonzepte konnte damit am Beispiel des Fachgebiets Datenmanagement erfolgreich angewandt werden. Durch die Möglichkeit, ein Fachgebiet relativ flexibel zu explorieren und zu strukturieren, scheint das Modell auch bezogen auf andere Gebiete der Informatik und darüber hinaus interessant und erfolgversprechend zu sein.

9 Entwicklung eines Data-Literacy-Kompetenzmodells

Wie im Rahmen dieser Arbeit auf vielfältige Weise gezeigt wurde, stellen Datenmanagement und Daten im Allgemeinen ein wichtiges Feld der Informatik dar, aus dem heute jeder mindestens grundlegende Kompetenzen benötigt, um mit Daten und datengetriebenen Technologien selbstbewusst und fundiert umgehen zu können. Entsprechend wird immer häufiger die Forderung nach einer *Data Literacy* laut: Jeder soll heute grundlegende Kompetenzen im Umgang mit und der Verarbeitung von Daten erwerben, um entsprechend fundiert mit diesen umgehen und aktuelle Entwicklungen einschätzen zu können. Obwohl der Begriff ursprünglich im Hochschulkontext geprägt wurde und meist insbesondere aus dieser Perspektive betrachtet wird, kann der Wert der dahinterstehenden Idee auch aus Sicht der Allgemeinbildung erkannt werden: Insbesondere bei Betrachtung des oft genannten Welterklärungsarguments[38], kann der Wert einer Data Literacy heute kaum mehr bestritten werden, sodass der Begriff immer häufiger auch außerhalb der Hochschulen Anwendung finden wird.

Als letzter theoretischer Abschnitt dieser Arbeit wird daher in diesem Kapitel der Begriff Data Literacy aus informatikdidaktischer Perspektive und mit Fokus auf allgemeinbildenden Sekundarschulunterricht aufgegriffen. Hierzu reicht der Rückgriff auf existierende Ansätze zur Charakterisierung der Data Literacy nicht aus, da diese einerseits stark auf Hochschulen fokussiert sind, meist aber auch eine stark interdisziplinäre Sichtweise einnehmen, die den Blick auf die dahinterstehenden informatischen Grundlagen nur eingeschränkt zulässt. Im Folgenden wird daher, basierend unter anderem auf Grundlage der zuvor beschriebenen Aufarbeitung des Datenmanagements, ein Data-Literacy-Kompetenzmodell entwickelt, das den Schwerpunkt auf allgemeinbildende Schulbildung legt. Vor der eigentlichen Entwicklung werden existierende Ansätze zur Beschreibung der Data Literacy und ihrer Kompetenzen zusammengefasst und argumentiert, warum diese im hier betrachteten Kontext nicht ausreichend sind. Um die Validität des entwickelten Modells zu überprüfen, wird abschließend das resultierende Modell zu bereits existierenden Ansätzen zur Charakterisierung und Konkretisierung einer Data Literacy kontrastiert und zur Betonung der Praxisrelevanz auch die Nutzung im unterrichtlichen Kontext verdeutlicht. Darauf basierend wurde eine Unterrichtssequenz entwickelt und erprobt, die in Kapitel 12 beschrieben wird.

9.1 Existierende Ansätze zur Charakterisierung der Data Literacy

Obwohl Data Literacy bisher kaum aus allgemeinbildender Perspektive betrachtet wurde, können in verschiedenen Arbeiten Indizien dafür gefunden werden, dass den unter diesem

[38] Vgl. beispielsweise *Bussmann und Heymann (1987)*, die den „*Aufbau eines Weltbildes*“ und die „*Anleitung zum kritischen Vernunftgebrauch*“ als Kriterien für Allgemeinbildung sehen.

Begriff zusammengefassten Themen und Kompetenzen ein allgemeinbildender Wert attestiert wird: Beispielsweise führte *Weintrop et al. (2016)* bei der Adaption des *computational thinking (vgl. Wing, 2006)* für den mathematischen und naturwissenschaftlichen Unterricht den Begriff *data practice* ein, der mit den Unteraspekten „*collecting, creating, manipulating, analyzing and visualizing data*“ *(Weintrop et al., 2016)* klare Überschneidungen mit Data Literacy aufweist. Von *Weintrop et al. (2016)* werden Daten dabei als essenziell erachtet: „*Data lie at the heart of scientific and mathematical pursuits. They serve many purposes, take many forms, and play a variety of roles in the conduct of scientific inquiry.*“ Auch in anderen Dokumenten können verschiedene Aspekte der Data Literacy erkannt werden, ohne dass dieser Begriff konkret genannt wird: Beispielsweise greifen die *Computational Thinking Teacher Resources (CSTA und ISTE, 2011)* den Umgang mit und die Analyse von Daten auf, indem beispielsweise Schülerinnen und Schüler in den Klassen 9–12 sich mit der Entwicklung einer Umfrage zu einer konkreten wissenschaftlichen Fragestellung beschäftigen, mit der sie daraufhin Daten erheben und der Fragestellung unter Zuhilfenahme statistischer hypothesenprüfender Verfahren nachgehen sollen. Aber auch die *K–12 Computer Science Standards* der *CSTA (2017)* greifen entsprechende Aspekte auf: Unter anderem sollen Schülerinnen und Schüler, mit dem Ziel Vorhersagen zu treffen, Muster in Datenvisualisierungen erkennen und beschreiben.

Trotz dieser erkennbaren Relevanz Data-Literacy-orientierter Aspekte für den allgemeinbildenden Schulunterricht, wird Data Literacy bisher eher aus Sicht der (oft fächerübergreifenden) Hochschulbildung beleuchtet. Dieser Fokus kann insbesondere dadurch begründet werden, dass *data-intense scientific discovery* (oftmals auch als *eScience* bezeichnet), inspiriert durch die Vision von *Hey, Tansley und Tolle (2009)*, immer häufiger als neues Wissenschaftsparadigma angesehen wird, das den wissenschaftlichen Erkenntnisprozess insbesondere dadurch verändert, dass anstatt durch gezielte Beobachtung ausreichende Datenmengen über einen Sachverhalt zu erfassen, heute möglichst große Datenmengen über diesen erfasst werden können, die zum Teil sogar größeren Erkenntnisgewinn ermöglichen sollen. Da klassische Methoden nicht mehr ausreichen, um solche größeren Mengen an (Beobachtungs-)Daten zu analysieren, besteht eine klare Notwendigkeit, in der Hochschulausbildung neben den klassischen Forschungsmethoden auch einen Schwerpunkt auf die Schaffung einer Data Literacy zu legen. Eine umfassende Untersuchung insbesondere der Inhalte und Kompetenzen, die im Rahmen einer Data-Literacy-Ausbildung thematisiert bzw. gefördert werden sollten, fand bisher jedoch auch in diesem Bereich nur eher eingeschränkt statt: Beispielsweise wurden in einer der anerkanntesten Arbeiten zur Data Literacy von *Ridsdale et al. (2015)* Strategien und Best-Practice-Beispiele zur Data-Literacy-Bildung untersucht und dabei sowohl wissenschaftliche Artikel als auch graue Literatur, z. B. Berichte und White-Papers, sowie informelle Literatur wie Blogbeiträge berücksichtigt. Auf diese Weise konnten verschiedene, insbesondere auch interdisziplinäre, Sichtweisen auf Data Literacy miteinbezogen werden. Eines der Ergebnisse dieser Studie stellt eine Sammlung von 23 Kompetenzen[39] dar (vgl. Abbildung 9.1). Diese Kompetenzen/Kompetenzbereiche

[39] Den von *Ridsdale et al. (2015)* definierten Kompetenzen liegt ein anderer Kompetenzbegriff zugrunde, als der hier verwendete nach *Weinert (2001)*. Entsprechend können die Kompetenzen nach *Ridsdale et al. (2015)* in dieser Arbeit eher als Kompetenzbereiche betrachtet werden.

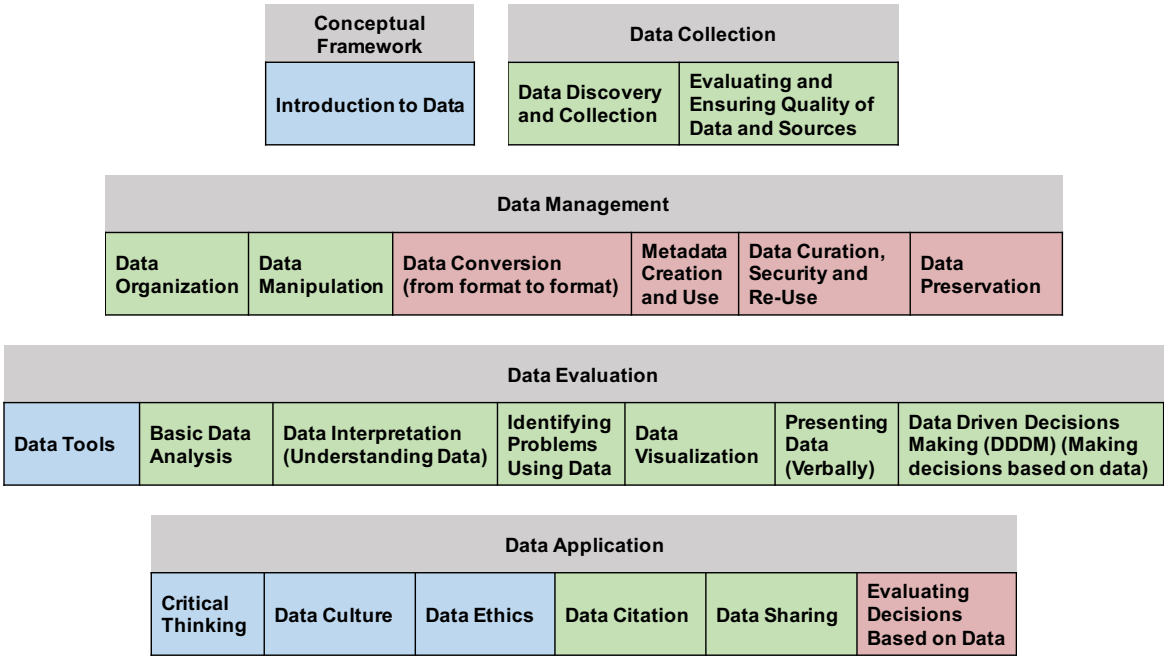


Abbildung 9.1: Kompetenzen der Data Literacy nach *Ridsdale et al. (2015)*.

werden durch 64 Aufgaben bzw. Fertigkeiten weiter charakterisiert, die eine deutliche Überschneidung zu Themenbereichen nicht nur des Datenmanagements, sondern auch der Data Science aufweisen: Beispielsweise werden *Datenermittlung und -sammlung* und *Datenmanipulation* als Kompetenzen sowie *nützliche Daten identifizieren*, *Daten bereinigen* und *anwenden von und arbeiten mit Daten in einer ethischen Art und Weise* als Aufgaben/Fertigkeiten genannt.

Obwohl diese Arbeit und andere ähnliche Ansätze verschiedene Charakteristika der Data Literacy aufzeigen und grundsätzlich geeignet sind um diese zu charakterisieren, können Sie aus informatikdidaktischer Perspektive (insbesondere mit Fokus auf Sekundarschulen) nicht als ausreichend angesehen werden, um als Grundlage für die weitere didaktische Forschung und für einen durch Data Literacy geprägten Informatikunterricht zu dienen. Dies ist einerseits durch die Verwendung eines nicht näher spezifizierten und im Fachgebiet unüblichen Kompetenzbegriffs begründet, insbesondere aber auch durch die gewählte auf Hochschulen fokussierte Sichtweise, den oft interdisziplinären Schwerpunkt und die überwiegende Betrachtung nichtfachlicher Quellen, die eine fachliche Fundierung nicht ausreichend sicherstellen können. Um diese Lücke aufzugreifen, wird im Folgenden auf Basis der bereits vorgenommen Charakterisierung von Datenmanagement sowie basierend auf einer inhaltlichen Beschreibung der Data Science die Entwicklung eines fachlich fundierten Data-Literacy-Kompetenzmodells beschrieben.

9.2 Fachliche Fundamente der Data Literacy

Um die angestrebte fachliche Fundierung sicherzustellen, müssen die hinter diesem Begriff stehenden informatischen Grundlagen ermittelt und berücksichtigt werden. Verschiedene Definitionen der Data Literacy legen dabei meist nur marginal unterschiedliche Schwerpunkte: Beispielsweise beschreiben *Ridsdale et al. (2015)* Data Literacy als „*ability to collect, manage, evaluate, and apply data, in a critical manner*“, während das *Hochschulforum Digitalisierung (2017)* „*die Kompetenzen, Daten zu erfassen, erkunden, managen, kuratieren, analysieren, visualisieren, interpretieren, kontextualisieren, beurteilen und anzuwenden*“ hevorhebt, die sich insbesondere durch Konkretisierung des Anwendungsaspekts auszeichnen. Der Vergleich dieser Charakterisierungen der Data Literacy mit dem bisherigen Schwerpunkt dieser Arbeit, dem Datenmanagement, zeigt, dass eine Datenkompetenz große Überschneidungen mit diesem Fachgebiet aufweist, gleichzeitig aber darüber hinausgeht. Dabei ist klar erkennbar, dass das Datenmanagement mit den zuvor ermittelten Praktiken bzw. dem in Abbildung 8.8 dargestellten Lebenszyklusmodell, alle in den beschriebenen Charakterisierungen von Data Literacy relevanten praktischen Aspekte berücksichtigt und auch die weiteren Schlüsselkonzepte in dieser relevant sind. Zusätzlich stellt die Data Literacy aber, insbesondere mit der Analyse, Visualisierung und Interpretation, auch Bereiche in den Vordergrund, die im Datenmanagement eine weniger zentrale Rolle spielen und heute eher der Data Science zugeordnet werden. Neben diesen klar informatischen Aspekte, die entsprechend insbesondere durch Data Science und Data Management geprägt werden, weist Data Literacy üblicherweise auch einen Anwendungsbezug auf: Dies zeigt sich beispielsweise durch die Aspekte *kontextualisieren, beurteilen* und *anwenden*, die das *Hochschulforum Digitalisierung (2017)* betont. Üblicherweise wird Data Literacy dazu auf einen spezifischen Anwendungsbereich bzw. Kontext bezogen, aus dem Wissen mitgebracht werden muss, um mit Daten zielführend und angemessen umzugehen. Entsprechend kann die Data Literacy – wie in Abbildung 9.2 dargestellt – auf der Grundlage von drei Säulen beschrieben werden.

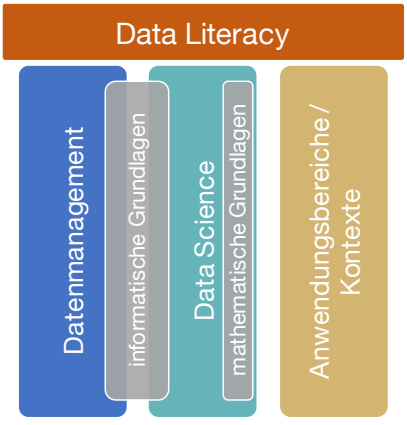


Abbildung 9.2: Datenmanagement, Data Science und Anwendungsbereiche bzw. Kontexte als Säulen der Data Literacy.

Für die fachlich fundierte Entwicklung eines Data-Literacy-Kompetenzmodells müssen daher die Grundlagen sowohl des Datenmanagements als auch der Data Science ausreichend berücksichtigt werden. Während dafür im Bereich des Datenmanagements, durch die bislang beschriebenen Arbeiten, bereits eine ausreichende Grundlage geschaffen wurde, war dies bis dato für die Data Science noch nicht der Fall. Aus diesem Grund wurden in einer qualitativen Studie alle deutschsprachigen Data-Science-Studiengänge hinsichtlich ihrer inhaltlichen Schwerpunkte analysiert und zu internationalen Studiengängen kontrastiert. Details zu dieser Studie, die hier nicht im Detail wiedergegeben wird, da sie einen Exkurs gegenüber den Schwerpunkten dieser Arbeit darstellt, sind in Anhang E sowie in *Grillenberger und Romeike (2018c)* zu finden. Als Ergebnis dieser Untersuchung stand die inhaltliche Charakterisierung der Data Science durch vier zentrale Themenbereiche.

- *Datenanalyse und Maschinenlernen* beinhaltet unter anderem den kompletten Prozess der Datenanalyse und die dabei notwendigen Schritte, insbesondere aber auch dabei genutzte Methoden, die zum Teil auch aus dem maschinellen Lernen stammen, aber in der Datenanalyse eine wichtige Rolle spielen (wie überwachtes und unüberwachtes Lernen).
- *Datenspeicher* befassen sich mit jeglichen Aspekten der Datenspeicherung und stellen somit die Schnittstelle zum Datenmanagement dar. Dabei wird nicht nur die Außenperspektive des Nutzers auf Datenspeicher betrachtet, sondern auch deren Funktionsweise.
- *Big Data* berücksichtigt die speziellen Aspekte des Umgangs mit großen und komplexen Datenmengen, insbesondere die dafür notwendigen Speicher- und Verarbeitungsarchitekturen, aber auch spezielle Algorithmen (z. B. Map-Reduce).
- *Datenschutz und Ethik* betrachten die Möglichkeiten der Data Science aus einer abstrakteren Perspektive und beziehen so auch ethische und datenschutzbedingte Überlegungen mit ein.

Diese vier Bereiche – mit den in der Studie ermittelten Unteraspekten – charakterisieren die Data Science aus informatischer Perspektive und können somit bei der Gestaltung des Data-Literacy-Kompetenzmodells zusammen mit den Schlüsselkonzepten des Datenmanagements zentrale Aspekte beisteuern.

9.3 Entwicklung des Data-Literacy-Kompetenzmodells

Auf Basis der Charakterisierung der Data Science durch ihre Inhalte sowie des Modells der Schlüsselkonzepte des Datenmanagements kann die theoretisch-argumentative Entwicklung eines Kompetenzmodells der Data Literacy erfolgen. In Anlehnung an andere Kompetenzmodelle, insbesondere diejenigen, die den *Empfehlungen der Gesellschaft für Informatik für Bildungsstandards Informatik für die Sekundarstufe I/II (Arbeitskreis Bildungsstandards, 2008; Arbeitskreis Bildungsstandards SII, 2016)* sowie den US-amerikanischen Mathematik-

Bildungsstandards *(*Principles and Standards for School Mathematics *2000)* zugrundeliegen, wird das angestrebte Kompetenzmodell in zwei Dimensionen aufgeteilt: *Inhaltsbereiche* sind auf die inhaltliche Perspektive bzw. die fachlichen Kompetenzen fokussiert, während *Prozessbereiche* den praktischen Umgang mit den Inhalten berücksichtigen. Diese Aufteilung scheint insbesondere auch für ein Data-Literacy-Kompetenzmodell zielführend, da auf diese Weise sowohl eine *fachliche Fundierung*, als auch der *praktische Umgang mit Daten* gleichermaßen betont werden und durch die enge Verknüpfung der beiden Modelldimensionen als voneinander nicht trennbar angesehen werden. Entsprechend müssen auch Kompetenzen aus dem Bereich der Data Literacy immer beide Sichtweisen berücksichtigen: Beispielsweise ist *Datenanalyse* ein potenzieller Prozessbereich, der einen wichtigen Anknüpfungspunkt zum Alltag darstellt, in welchem heute Datenanalysen und deren Ergebnisse immer häufiger thematisiert werden. Um jedoch nachvollziehen zu können, wie diese praktische Tätigkeit funktioniert, und entsprechend deren Auswirkungen zu beurteilen, müssen verschiedene Konzepte aus Datenmanagement und Data Science verstanden werden, die durch die verknüpften Inhaltsbereiche charakterisiert werden. Gleichzeitig reicht eine rein theoretische Thematisierung kaum aus, um die hohe Relevanz der Thematik zu vermitteln und beispielsweise zu zeigen, wie einfach solche Datenanalysen durchführbar sind. Somit wird deutlich, dass es zwar durchaus möglich ist einen, je nach Unterrichtsziel unterschiedlichen, Schwerpunkt auf die Prozess- oder die Inhaltsbereiche zu legen, aber gleichzeitig keiner der beiden Teile des Kompetenzmodells komplett außen vor gelassen werden kann.

In den folgenden beiden Abschnitten werden diese Teile des Modells unabhängig voneinander entwickelt, um deren unterschiedliche Ausgestaltung zu berücksichtigen. Daraufhin werden sie zu einem Kompetenzmodell der Data Literacy zusammengeführt, dieses diskutiert und zu existierenden Ansätzen kontrastiert.

9.3.1 Inhaltsbereiche der Data Literacy

Um die Inhalte der Data Literacy zu ermitteln, ist es notwendig, sowohl Aspekte der Data Science als auch des Datenmanagements miteinzubeziehen. Fasst man alle inhaltlichen Aspekte dieser beiden Charakterisierungen zusammen, führt dies zu sieben verschiedenen Bereichen, die als Grundlage für die Ermittlung der Inhaltsbereiche genutzt werden können: Aus Perspektive der Data Science wurden die vier der zuvor beschriebenen Charakterisierung entstammenden Bereiche *Datenanalyse und Maschinenlernen*, *Big Data*, *Datenschutz, Ethik* und *Datenspeicher* herausgegriffen. Im Bereich des Datenmanagements erfolgt hingegen eine Beschränkung auf die *Kerntechnologien*, *Entwurfsprinzipien* und *Mechanismen*, da die Praktiken einen anderen Fokus haben und nicht konkret inhaltlich beitragen. Um aus diesen fachlichen Inhaltsbereichen die Inhaltsbereiche des angestrebten Kompetenzmodells anzuleiten, ist eine weitere Ausdifferenzierung und Klärung notwendig, da diese Begriffe unterschiedliche konzeptuelle Ebenen beschreiben, zum Teil Überschneidungen aufweisen und sich auch hinsichtlich ihrer Konkretisierung stark unterscheiden. Um die Inhaltsbereiche abzuleiten, wurden diese Begriffe daher konsolidiert und dabei das Ziel

berücksichtigt, ein für Sekundarschulen geeignetes Kompetenzmodell der Data Literacy zu entwerfen, das den folgenden Kriterien genügen:

Jeder Inhaltsbereich...

- charakterisiert einen eigenen Teilbereich der Data Literacy.
- fasst eine Menge eng verwandter Konzepte/Ideen unter einem geeigneten Überbegriff zusammen.
- befasst sich mit Inhalten, die insbesondere für Data Literacy relevant sind, nicht nur für die Informatik im Gesamten.
- fokussiert die fachliche Sichtweise auf die jeweiligen Inhalte der Data Literacy.
- weist möglichst geringe Überschneidung mit anderen Inhaltsbereichen auf.

Zur Ermittlung der Inhaltsbereiche, wurden die zuvor genannten Themenbereiche zuerst, basierend auf ihren Definitionen bzw. Charakterisierungen, aber auch unter Betrachtung ihrer Überschneidungen und Unterschiede, detaillierter beschrieben. Dazu wurden sie zunächst auf eine längere Liste von spezifischeren Themen erweitert, um die Überschneidungen und Gemeinsamkeiten der verschiedenen Bereiche klarer erkennbar zu machen. Entsprechend konnte die folgende detaillierte Charakterisierung der zuvor gefundenen Begriffe gefunden werden:

- *Datenanalyse und Maschinenlernen:*
  - grundlegende Ideen der Datenanalyse, wie die Unterscheidung von Daten und Information, Informationsentropie, Korrelation vs. Kausalität
  - Methoden der Datenanalyse, wie Klassifikation und Clustering
  - datenbasierte Vorhersage
  - Lernen auf Basis von Daten, insbesondere überwachtes Lernen
  - Qualität von Daten und Analyseergebnissen
- *Big Data:*
  - korrelationsbasierte Datenanalyse
  - Techniken zur Verwaltung großer Datenmengen
  - Systeme zur Speicherung von Big Data
- *Datenschutz, Datenethik:*
  - Datenethik
  - Grundlagen der Datensicherheit
  - Datenschutz
  - persönliche bzw. personenbezogene Daten
- *Datenspeicherung:*

  - Systeme zur Speicherung und Verwaltung von Daten
  - Funktionsprinzipien von Datenspeichern
- *Kerntechnologien (des Datenmanagements):*
  - Systeme zur Speicherung und Verwaltung von Daten
- *Mechanismen (des Datenmanagements):*
  - Funktionsprinzipien von Datenspeichern
  - Repräsentation von Daten auf physikalischer Ebene
- *Entwurfsprinzipien (des Datenmanagements):*
  - Möglichkeiten zum Zugriff auf Daten
  - Anforderungen an Datenspeicher und Datenspeicherung

Um die in dieser Darstellung vorhandenen Überschneidungen zu vermeiden und ähnliche/verwandte Aspekte in einem Inhaltsbereich zu repräsentieren, wurden die dargestellten Aspekte neu gruppiert und (teils neuen) Begriffen, die die Inhaltsbereiche darstellen, untergeordnet. Durch die dabei erfolgte Reduktion konnten vier Inhaltsbereiche ermittelt werden, die sowohl die dargestellten inhaltlichen Aspekte berücksichtigen, als auch die oben dargestellten Kriterien erfüllen. Die erfolgten Zusammenfassungen sind in Abbildung 9.3, in der die Zuordnungen der einzelnen Aspekte zu den im Folgenden beschriebenen finalen Inhaltsbereiche (C1)–(C4), dargestellt.

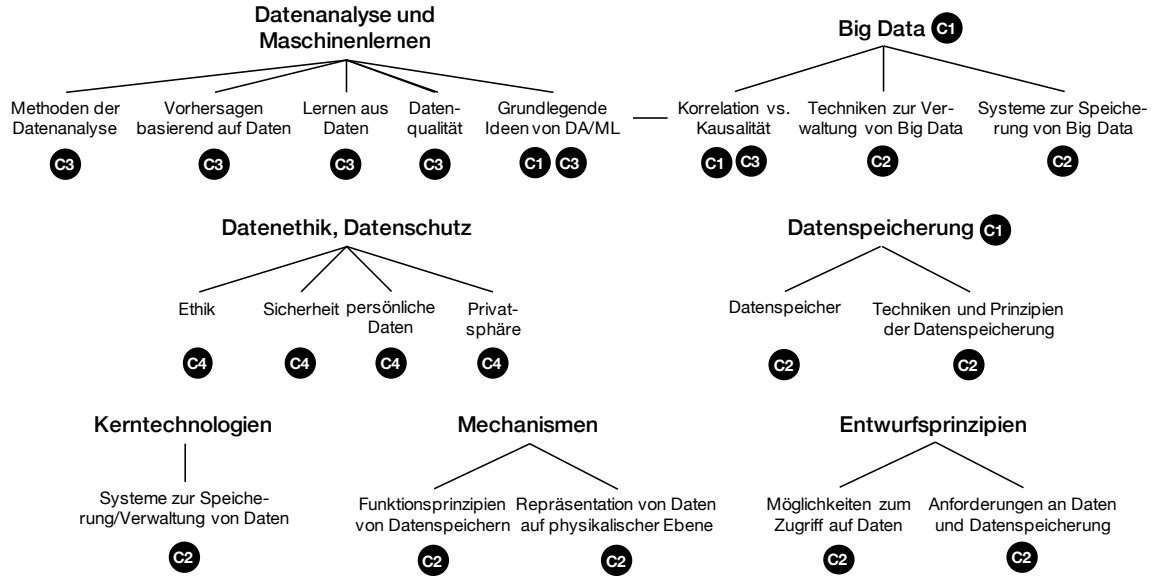


Abbildung 9.3: Bezüge der betrachteten Themenbereiche zu den Inhaltsbereichen (markiert durch C1–C4).

(C1) Daten und Information wurde als Inhaltsbereich neu eingeführt und vereinigt mehrere Aspekte: Er berücksichtigt beispielsweise Unterscheidung von Information und

Daten, die unterschiedliche Aussagekraft verschiedener Daten (im Sinne des Informationsgehaltes nach *Shannon (1948)*), Methoden zur Erfassung von Information in Form von Daten, aber auch zur Strukturierung von Daten durch die Anreicherung mit zusätzlicher Information beispielsweise in Form von Metadaten, sowie die Unterscheidung großer und kleiner Datenmengen bezüglich deren Aussagekraft/Informationsgehalt. Entsprechend enthält dieser Inhaltsbereich Aspekte, die ursprünglich den Bereichen *Datenanalyse/Maschinenlernen*, *Big Data*, *Datenspeicherung* und *Mechanismen* zugeordnet waren und gibt insbesondere einen Überblick über Daten im Allgemeinen.

(C2) **Datenspeicher und Datenspeicherung** berücksichtigt insbesondere Aspekte aus dem Bereich des Datenmanagements: Es beinhaltet dabei grundlegende Mechanismen von Datenspeichern, wie die Repräsentation von Daten bis hin zur physikalischen Ebene, deren Replikation, Synchronisation oder Transport. Gleichzeitig werden aber auch grundlegende Konzepte von Datenspeichern diesem Bereich zugeordnet, wie die Entscheidung ob Daten konsistent, dauerhaft und/oder hoch verfügbar gespeichert werden sollen. Somit beinhaltet dieser Inhaltsbereich insbesondere die eher statischen Aspekte des Umgangs mit Daten sowie deren Speicherung und die dabei relevanten Konzepte.

(C3) **Datenanalyse** beschäftigt sich mit allen Aspekten hinsichtlich der Gewinnung von (neuen) Informationen aus Daten. Dabei sind insbesondere Methoden, Prinzipien und Algorithmen der Datenanalyse relevant. Dies beinhaltet nicht nur traditionelle Methoden wie die systematische Aggregation bzw. deskriptive Methoden im Allgemeinen, sondern auch explorative, korrelationsbasierte und auf Vorhersagen ausgerichtete Datenanalysen, die oft in starkem Zusammenhang mit maschinellem Lernen stattfinden und Methoden aus diesem Bereich miteinbeziehen. Trotzdem wurde, gegenüber dem ursprünglichen Themenbereich *Datenanalyse/Maschinenlernen*, das Maschinenlernen hier außen vor gelassen, da der Fokus der Data Literacy insbesondere auf solchen Aspekten des Maschinenlernens liegt, die in starkem Zusammenhang mit der Datenanalyse stehen – und nicht auf dem Fachgebiet im Gesamten. Eine Miteinbeziehung hätte daher zwar zur Konkretisierung beigetragen, gleichzeitig aber auch eine verfälschte Wahrnehmung provoziert.

(C4) **Datenethik und Datenschutz** wurde direkt aus dem entsprechenden Thema abgeleitet und berücksichtigt alle ethischen, aber auch gesellschaftliche und teils dem Persönlichkeitsschutz entstammende Fragestellungen, die im Rahmen der Arbeit bzw. des Kontakts mit Daten auftreten. Dabei ist hervorzuheben, dass sich Datenschutz hier nicht nur auf den Schutz der Privatsphäre bezieht, sondern auch den Schutz persönlicher Daten beispielsweise durch Maßnahmen wie Verschlüsselung, explizit mit einbezieht.

Zusammenfassend berücksichtigen diese vier Inhaltsbereiche eine Vielfalt von Aspekten der Data Literacy aber auch generell der Informatik: Während der zweite Inhaltsbereich offensichtlich stark auf Aspekte des Datenmanagements fokussiert ist und der dritte auf

Data Science, greifen die anderen beiden allgemein relevante Aspekte auf, wie den Unterschied von Daten und Information. Diese vielfältigen Bezüge zu verschiedensten Themen der Informatik finden sich jedoch auch in den beiden eher spezialisierten Bereichen: Beispielsweise spielen im zweiten Inhaltsbereich bei der Thematisierung der Möglichkeiten zum Zugriff auf Daten auch Aspekte der Rechnerkommunikation eine nicht zu vernachlässigende Rolle, während im dritten Bereich hingegen klare Bezüge zur Algorithmik erkannt werden können. Entsprechend zeigen schon diese Inhaltsbereiche, dass Data Literacy kein eigenständiges Thema bzw. Fach darstellt, sondern sehr starke Wurzeln in der Informatik hat und vielfältige Bezüge zu anderen ihrer Themen aufweist, die im Kontext der Schaffung einer Data Literacy berücksichtigt werden müssen.

9.3.2 Prozessbereiche der Data Literacy

Um die Prozessbereiche zu erarbeiten, wurde der Datenlebenszyklus (vgl. Abschnitt 8.3.1) als Basis betrachtet: Dieser berücksichtigt, wie zuvor bereits angedeutet, den praktischen Umgang mit Daten im Allgemeinen und bezieht die zentralen Tätigkeiten sowohl aus Perspektive des Datenmanagements als auch der Data Science mit ein. Damit demonstriert dieser klar die praktischen Aspekte des Umgangs mit Daten und entsprechend auch der Data Literacy. Die Teilprozesse des Lebenszyklusmodells (vgl. Abbildung 8.8) können dabei als Kandidaten für Prozessbereiche der Data Literacy betrachtet werden: *Erfassung/Gewinnung*, *Bereinigung*, *Modellierung*, *Implementierung*, *Optimierung*, *Verarbeitung/Analyse*, *Visualisierung*, *Evaluation*, *Austausch*, *Archivierung*, *Löschung*.

Aus fachlicher Perspektive beschreiben diese elf Bereiche typische Aktivitäten, die für den Umgang mit Daten zentral sind. Gleichzeitig stellen sie diejenigen Tätigkeiten dar, die auch für die Schaffung einer Data Literacy essenziell sind, da nur die vollumfängliche Betrachtung dieser Praktiken die Möglichkeiten, die Daten heute bieten, offenlegt und nur so der fundierte Umgang mit diesen ermöglicht wird. Aus didaktischer Sicht ist eine derart umfangreiche Liste an Prozessbereichen, die eng miteinander verknüpft sind und zu deren eindeutigem Verständnis teils detailliertes Fachwissen notwendig ist, jedoch ungünstig, da dadurch die Verständlichkeit des entwickelten Modells eingeschränkt würde. Aus diesem Grund wurden diese Kandidaten für Prozessbereiche daher aus informatikdidaktischer Sicht konsolidiert, mit dem Ziel, deren Anzahl zu reduzieren, sie besser voneinander abzugrenzen und damit verständlicher darzustellen. Dazu wurden die zuvor genannten Kandidaten mit Informatiklehrkräften und Wissenschaftlern diskutiert, die unterschiedlich großes Vorwissen zu Datenmanagement und Data Science hatten. Um die Verständlichkeit und den Nutzen dieser potenziellen Prozessbereiche für den Unterricht zu diskutieren, wurde versucht in zwei Teilgruppen je eine Unterrichtssequenz zu skizzieren: Diese sollten jeweils möglichst alle Prozessbereiche miteinbeziehen und einen Schwerpunkt auf die Vermittlung grundlegender Kompetenzen im Sinne einer Data Literacy legen. In beiden Gruppen konnten dabei verschiedene Probleme identifiziert werden, von denen sich in einer nachfolgenden Diskussion folgende als Konsens herausgestellt haben:

- Manche der potenziellen Prozessbereiche hatten zu starke Überschneidungen und konnten kaum getrennt voneinander betrachtet werden: Insbesondere hängen *Implementierung* und *Optimierung* stark zusammen und gehen fließend ineinander über, aber auch *Archivierung* und *Löschung* können kaum voneinander getrennt werden, da eine Entscheidung für eine der beiden Praktiken gleichzeitig die jeweils andere ausschließt und auf derselben Überlegung fußt. Auch die *Erfassung/Gewinnung* von Daten sollte grundsätzlich gemeinsam mit der *Bereinigung* betrachtet werden, um Datenfehler möglichst frühzeitig erkennen und zu vermeiden. Diese drei Paare von Prozessbereichen wurden daher jeweils vereinigt, um (Verständnis-)Probleme, die durch diese im schulischen Kontext künstlich erscheinende Trennung entstehen, zu vermeiden, und somit die Verständlichkeit und Nachvollziehbarkeit des resultierenden Modells zu erhöhen.

- Der Bereich *Modellierung* weist starke Bezüge zu anderen Bereichen auf: Insbesondere ist es bereits bei der Erfassung der Daten essenziell, zu entscheiden, welcher Ausschnitt der Realität in Daten abgebildet werden soll. Aber auch bei der Implementierung, beispielsweise in Form einer Datenbank, sowie bei der Planung einer Datenanalyse ist Modellierung zentral. Diese beiden Rollen, die Modellierung hier einnehmen kann, Datenmodellierung und Prozessmodellierung, werden daher fortan getrennt betrachtet. Während Prozessmodellierung als den jeweiligen Bereichen inhärenter Aspekt betrachtet wird und somit nicht explizit betont wird, wird die Datenmodellierung aufgrund ihres übergreifenderen Charakters, der den gesamten Prozess beeinflusst, getrennt betrachtet. Um sie gleichzeitig hervorzuheben, gleichzeitig aber auch die Anzahl der Prozessbereiche gering zu halten, wurde *Modellierung* mit dem Bereich *Erfassung/Gewinnung und Bereinigung* zusammengeführt, da an dieser Stelle die zentralen Modellierungsgedanken, wie das Weglassen von irrelevanten Aspekten bzw. Daten, am stärksten zum Tragen kommen.

- Ähnlich wie *Archivierung und Löschung* stellt auch der *Austausch* eine Form des Umgangs mit vorhandenen Daten dar, der nicht für die konkrete Analyse relevant ist, sondern übergreifende Bedeutung hat. Dabei werden ethische Aspekte und externe Anforderungen (wie Vorschriften zur dauerhaften Vorhaltung von Daten und Datenschutzrichtlinien, die oft eine Löschung von Daten fordern und/oder deren Weitergabe einschränken) in allen drei Bereichen miteinbezogen. Aufgrund ihrer engen Verwandtheit wurden diese drei Bereiche zusammengeführt.

- Zusätzlich fehlte beiden Gruppen ein von ihnen als zentral erachteter Bereich: die *Interpretation*. Zwar kann diese als inhärenter Teil der *Analyse* sowie der *Visualisierung* angesehen werden, trotzdem wurde jedoch eine explizite Nennung als gewinnbringend erachtet, um ihre zentrale Bedeutung hervorzuheben. Entsprechend wurde diese ergänzt und mit der *Analyse* und *Visualisierung*, die typischerweise eng miteinander verknüpft sind, zusammengeführt.

- Die Praktik *Verarbeitung* wurde nicht explizit aufgenommen, da diese durch Analyse und Visualisierung konkreter gefasst wird und eine Beschränkung darauf keine als zentral erachteten Aspekte vernachlässigt.

Basierend auf dieser Argumentation konnte eine wesentlich kompaktere, übersichtlichere und leichter erfassbare Liste an Prozessbereichen erarbeitet werden:

(P1) erfassen, bereinigen und modellieren
Dieser Prozessbereich beschäftigt sich mit den frühen Phasen der Arbeit mit Daten. Mit dem Erfassen, Bereinigen und Modellieren von Daten werden drei Bereiche betrachtet, die nicht voneinander getrennt werden können: Bereits bei der Erfassung ist es notwendig, Daten in einer geeigneten Form zu strukturieren. Damit sind bereits hier zwei Modellierungsentscheidungen nötig: Welcher Ausschnitt der Welt soll in Form von Daten erfasst werden und wie werden diese strukturiert. Gleichzeitig muss bereits zu diesem Zeitpunkt eine Überprüfung der Datenqualität in der Hinsicht stattfinden, dass diese von Fehlern bereinigt werden, sodass später eine valide Verarbeitung und Analyse möglich wird. Entsprechend beschäftigt sich dieser erste Prozessbereich mit drei zentralen Fragestellungen: *Welche Charakteristika möchte ich in Form von Daten und auf welche Weise erfassen? Wie kann ich diese so speichern, dass ich sie später weiterverarbeiten kann? Sind die erfassten Daten für den angestrebten Zweck geeignet?*

(P2) implementieren und optimieren
Das Implementieren und Optimieren kann zu verschiedenen Zeitpunkten im Datenlebenszyklus stattfinden: Insbesondere betrifft es die Implementierung des Datenmodells in einem geeigneten Datenspeicher und die Speicherung der Daten in diesem. Andererseits ist – je nach gewählter Methode und Werkzeug – gegebenenfalls auch bereits bei der Datenerfassung oder später bei der Analyse eine entsprechende Implementierung (und auch Optimierung), beispielsweise von einfachen Algorithmen, nötig. Entsprechend können auch bei der Optimierung verschiedene Ziele verfolgt werden, die alle anderen Bereiche des Datenlebenszyklus betreffen können. Die zentralen Fragestellungen dieses Prozessbereichs sind daher: *Wie kann ich die Datenerfassung, -speicherung und Analyse praktisch realisieren? Wie kann ich das bisher erreichte hinsichtlich konkreter Ziele verbessern?*

(P3) analysieren, visualisieren und interpretieren
Das Analysieren von Daten beschäftigt sich mit der Extraktion neuer Informationen aus den gewonnenen und gespeicherten Daten. Dabei wird auf verschiedene Analysemethoden und Prinzipien zurückgegriffen. Um die gewonnenen Informationen, aber auch die ursprünglichen Daten, einfacher erfassbar zu machen und zur Unterstützung der Interpretation, werden häufig verschiedene Visualisierungstechniken eingesetzt. Teilweise kann auch die eigentliche Datenanalyse mit visuellen Methoden durchgeführt werden. Der dritte Prozessbereich beschäftigt sich entsprechend mit drei zentralen Fragestellungen: *Welche Informationen kann ich wie aus meinen Daten*

extrahieren? Wie kann ich den Menschen dabei unterstützen, das Wesentliche einfach zu erfassen? Welche Schlüsse kann ich aus meinen Analyseergebnissen ziehen?

(P4) austauschen, archivieren und löschen
Der vierte Prozessbereich ist während des gesamten Erfassungs-, Speicherungs- und Analyseprozesses von Bedeutung. Dieser betrachtet das Austauschen, Archivieren und Löschen von Daten, drei Tätigkeiten, die immer relevant sind, wenn mit Daten gearbeitet wird. Es muss beispielsweise entschieden werden, wer wie Zugriff auf Originaldaten oder Analyseergebnisse bekommt, welche Daten es Wert sind erfasst und ggf. langfristig gespeichert zu werden, welche Daten anonymisiert oder pseudonymisiert werden müssen, wie Daten, beispielsweise unterstützt durch Metadaten wieder auffindbar werden, etc. Gleichzeitig markiert die Löschung von Daten jedoch auch das einzige Ende des Datenlebenszyklus, sodass es bei dieser insbesondere darauf ankommt, zu entscheiden, wann Daten gelöscht werden sollen und wie dies auf angemessene und sichere Weise geschehen kann. Damit werden in diesem Prozessbereich die folgenden Fragestellungen angegangen: *Welche Daten möchte ich wie mit wem teilen? Welche Daten sollte ich langfristig archivieren? Wie können Daten langfristig archiviert werden? Wie kann ich meine Daten angemessen löschen?*

9.4 Das Data-Literacy-Kompetenzmodell

Durch Kombination der jeweils vier Prozess- und Inhaltsbereiche kann nun das Kompetenzmodell der Data Literacy betrachtet werden. Entsprechend seiner Konzeption werden auch in der grafischen Darstellung (vgl. Abbildung 9.4) diese beiden Typen von Bereichen eng miteinander verzahnt dargestellt.

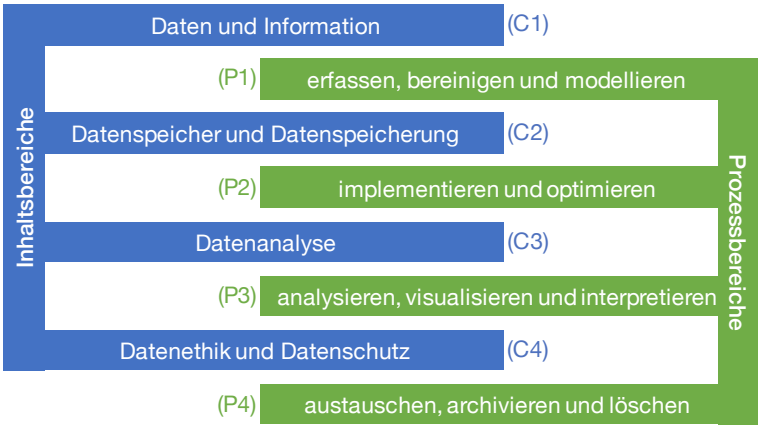


Abbildung 9.4: Das entwickelte Data-Literacy-Kompetenzmodell.

Durch das entwickelte Kompetenzmodell wird die Data Literacy auf inhaltlicher Ebene insbesondere durch Aspekte des Datenmanagements, der Data Science (mit Bezügen zum Maschinenlernen) und der Datenethik charakterisiert. Auf Prozessebene sind Bezüge zu

allen Aktivitäten bzw. Abschnitten des Datenlebenszyklus klar erkennbar. Diese Ausgestaltung des Modells ist konsistent zu verschiedenen Definitionen der Data Literacy, die häufig insbesondere die praktischen Aspekte des Umgangs mit Daten betonen, beispielsweise setzen *Deahl (2016)* den Schwerpunkt auf *„understand, find, collect, interpret, visualize, and support arguments using quantitative and qualitative data."* Im Allgemeinen setzen existierende Definitionen und Charakterisierungen der Data Literacy zwar verschiedene Schwerpunkte und beschreiben Data Literacy insbesondere aus praktischer Perspektive, dabei sind jedoch keinerlei Widersprüche zum hier entwickelten Kompetenzmodell erkennbar.

Im Sinne des hier entwickelten Data-Literacy-Kompetenzmodells können in diesem situierte Kompetenzen niemals nur Prozess- oder Inhaltsbereiche berücksichtigen, was auch durch die verschiedenen offensichtlichen Verbindungen im Modell deutlich wird: Beispielsweise weist der Prozessbereich *analysieren, visualisieren und interpretieren* offensichtlich deutliche Bezüge zum Inhaltsbereich *Datenanalyse und Maschinenlernen* auf. Jedoch können nicht nur die offensichtlich miteinander verknüpften Bereiche bzw. die in der grafischen Darstellung nebeneinander liegenden Bereiche miteinander verknüpft werden. Stattdessen sind für alle 16 möglichen Kombinationen der Prozess- und Inhaltsbereiche entsprechende Kompetenzen ermittelbar. Um diese Aussage zu untermauern, werden in Tabelle 9.1 für jede dieser Kombinationen drei exemplarische Kompetenzen dargestellt, die bisher jedoch nicht weiter evaluiert wurden. Insbesondere sind diese nicht speziell für ein bestimmtes kognitives Niveau angepasst, sodass sowohl Kompetenzen gelistet sind, die eher für die Sekundarschulinformatik geeignet sind, als auch solche, die erst im Studium thematisiert werden könnten. Entsprechend können und sollen diese Kompetenzen nur einen Eindruck über eine mögliche Ausgestaltung des entwickelten Kompetenzmodells geben. In der weiteren Forschung müssen diese Kompetenzen nicht nur tiefergehend evaluiert, sondern auch auf verschiedene kognitive Niveaus bezogen werden, die in einer weiteren Modelldimension berücksichtigt werden können. Entsprechend ist das Modell trotz seiner Entwicklung unter Berücksichtigung der Schulinformatik nicht auf diese beschränkt, sondern kann bei entsprechender Ausgestaltung auch in der Hochschulbildung eingesetzt werden.

Prozessbereich / Inhaltsbereich	**erfassen, modellieren und bereinigen**	**implementieren und optimieren**	**analysieren, visualisieren und interpretieren**	**austauschen, archivieren und löschen**
Daten und Information	- auswählen von Sensoren zur Erfassung der gewünschten Daten - strukturieren von Daten, sodass sie später analysiert werden können - überprüfen, ob die erfassten Daten die zugrundeliegende Information widerspiegeln	- implementieren von Algorithmen zur Erfassung der Daten - implementieren von Algorithmen um Daten von Web-APIs herunterzuladen - diskutieren von Möglichkeiten zur Optimierung der Datenerfassung und deren Grenzen	- kombinieren von Daten um neue Informationen zu gewinnen - verdeutlichen von Informationen durch Visualisierungen - interpretieren von Daten und Analyseergebnissen	- entscheiden, ob Originaldaten mit anderen geteilt werden sollen/dürfen - auswählen von Originaldaten, die archiviert werden sollen - auswählen einer Methode zur Löschung nicht mehr benötigter Daten
Datenspeicher und Datenspeicherung	- auswählen eines geeigneten Datenmodells - strukturieren von Daten, sodass sie gespeichert und wieder abgerufen werden können - darstellen des Datenmodells in einer geeigneten Form	- auswählen eines Datenspeichers und Speicherung der Daten - nutzen von Möglichkeiten zur Erhöhung der Effizienz beim Zugriff auf Daten - erhöhen der Speicherplatzeffizienz durch Nutzung von Kompression	- anbinden vorhandener Datenquellen an ein Analysewerkzeug - nutzen geeigneter Austauschformate zwischen Datenspeicher und Analysewerkzeug - speichern von Analyseergebnissen in angemessener Art und Weise	- bestimmen, wer zu welchem Zweck Zugriff auf Daten bekommt - bestimmen geeigneter Zugriffsrechte auf Daten - diskutieren von Problemen, die durch Löschung von Daten entstehen
Datenanalyse	- überprüfen, ob bestimmte Daten die Qualität der Analyseergebnisse beeinflussen - strukturieren von Daten in geeigneter Weise für deren Analyse - verbinden von Daten verschiedener Quellen für Analysezwecke	- implementieren von Analysealgorithmen - nutzen von Stellschrauben zur Beeinflussung der Qualität der Datenanalyse - optimieren von Datenanalysen um deren Qualität zu erhöhen	- auswählen geeigneter Analysemethoden - visualisieren von Daten und Analyseergebnisse - interpretieren der Ergebnisse der Datenanalysen	- entscheiden, mit wem welche Analyseergebnisse geteilt werden - diskutieren ob Originaldaten nach der Analyse archiviert werden - entscheiden, ob Informationen über den Analyseprozess geteilt werden
Datenethik und Datenschutz	- berücksichtigen ethischer Aspekte bei der Datenerfassung - entscheiden, ob es ethisch vertretbar ist, Daten verschiedener Quellen zu kombinieren - diskutieren der kontinuierlichen Datenerfassung hinsichtlich Ethik und Datenschutz	- diskutieren von Wegen zur Anonymisierung und Pseudonymisierung von Daten - ausschließen bestimmter Daten von der Speicherung auf Basis ethischer Überlegungen - vergeben von Zugriffsrechten auf Daten unter Berücksichtigung des Datenschutzes	- diskutieren ethischer Einflüsse der durchgeführten Analyse - entscheiden, ob Analyseergebnisse ausreichend anonymisiert sind - berücksichtigen der durch die Analyse hervorgerufenen Datenschutzprobleme	- diskutieren, ob Datenspeicherung für potentielle zukünftige Nutzung vertretbar ist - löschen von personenbezogenen Daten auf sichere Art und Weise - finden von Methoden zur angemessenen Anonymisierung von Daten und Analyseergebnissen

Tabelle 9.1: Exemplarische Kompetenzen aus allen Kombinationen von Inhalts- und Prozessbereichen des Data-Literacy-Kompetenzmodells.

Das Data-Literacy-Kompetenzmodell liefert daher für die Praxis einen wichtigen Beitrag, indem es dabei hilft, Unterricht hinsichtlich der Förderung grundlegender Data-Literacy-Kompetenzen zu evaluieren, solchen zu planen und die Abdeckung zentraler Aspekte der Data Literacy zu prüfen. Gleichzeitig trägt es auch zur weiteren Forschung im Bereich der Data Literacy, insbesondere mit Bezug auf die Schulinformatik bei: Das Modell kann zur Fundierung in diesem Bereich eingesetzt werden, da sowohl dessen Ursprung als auch Entstehung durch die theoretisch-argumentative Ableitung aus existierenden Arbeiten klar nachvollziehbar sind. Die empirische Basis trägt dabei zur Sicherstellung einer möglichst hohen Validität bei, während die offene Argumentation der Entstehung der Prozess- und Inhaltsbereiche zu einer hohen Nachvollziehbarkeit führt. Diese Entstehung wird in Abbildung 9.5 nochmals dargestellt.

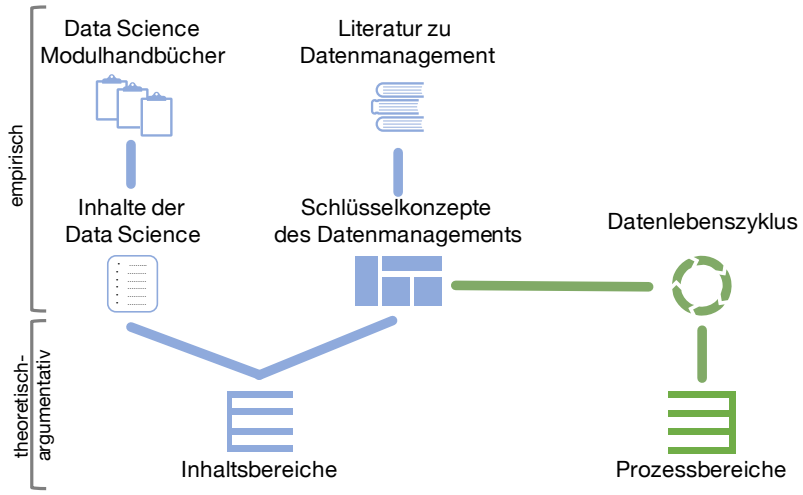


Abbildung 9.5: Abstammung bzw. Entwicklung des Data-Literacy-Kompetenzmodells.

9.5 Kontrastierung zu weiteren Data-Literacy-Kompetenzbeschreibungen

Wie einleitend erwähnt, ist die beschriebene Entwicklung eines Kompetenzmodells nicht der erste Ansatz zur Beschreibung der Data Literacy durch die ihr inhärenten Kompetenzen. Beispielsweise konnten *Wolff et al. (2017)* ausgehend vom Untersuchungsprozess (*problem – plan – data – analysis – conclusions*) verschiedene Kompetenzen ableiten, die von ihnen unter sieben *foundational competencies* zusammengefasst wurden (vgl. Abbildung 9.6). Diese Kompetenzen weisen eine große Überschneidung mit den hier entwickelten Prozessbereichen auf. Durch Nutzung eines detaillierteren Datenlebenszyklusmodell als Basis, geben die hier ermittelten Prozessbereiche jedoch einen tieferen Einblick. Nur ein Aspekt den Wolff et al. hervorheben, ist im hier entwickelten Modell nur implizit erkennbar: Der besondere Schwerpunkt auf die Kompetenz, Fragen ausgehend von Daten zu stellen. Dieser kommt sicherlich in der Praxis eine hohe Bedeutung zu, sie kann jedoch ausgehend von den hier

gewählten Grundlagen nicht fachlich begründet werden, spielt aber implizit insbesondere bei explorativen Analysen eine wichtige Rolle. In einer anderen Darstellung, vgl. Abbildung 9.7, verschieben auch Wolff et al. diesen Aspekt in den Realweltkontext, sodass an dieser Stelle kein Widerspruch der beiden Modelle, sondern allenfalls eine geringfügig andere Schwerpunktsetzung erkannt werden kann. Das hier entwickelte Modell fügt jedoch weitere Details hinzu und sorgt durch die explizite Betonung der Inhaltsbereiche, die bei Wolff et al. nicht enthalten sind, für eine stärkere fachliche Fundierung entsprechender Kompetenzen.

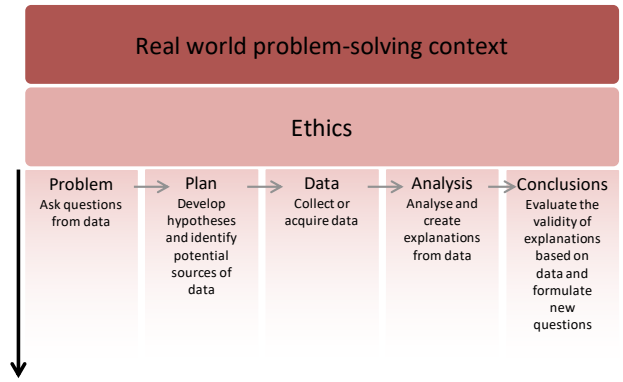


Abbildung 9.6: Data-Literacy-Kompetenzen nach *Wolff et al. (2017)*.

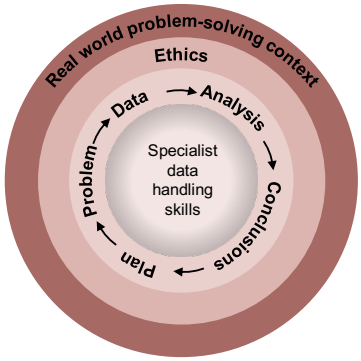


Abbildung 9.7: Data-Literacy-Pool nach *Wolff et al. (2017)*.

Auch in der von *Ridsdale et al. (2015)* durchgeführten Studie zur Data Literacy wurden verschiedene Kompetenzen ermittelt, jedoch nicht ausgehend von einer fachlichen Perspektive, sondern von interdisziplinären Best-Practice-Beispielen. Diese Kompetenzen entsprechen bei Vergleich mit dem dieser Arbeit zugrundeliegenden Kompetenzbegriff jedoch eher Kompetenzbereichen, sodass sie mit den hier ermittelten Inhalts- und Prozessbereichen verglichen werden können. Dabei kann eine hohe Übereinstimmung festgestellt werden: Obwohl sie anders strukturiert sind, werden alle hier ermittelten Prozessbereiche auch durch die von *Ridsdale et al. (2015)* ermittelten Kompetenzen abgedeckt. Ridsdale et al. fügen jedoch an manchen Stellen noch weitere Details hinzu, beispielsweise *metadata creation and use*, oder ergänzen Kompetenzen, die insbesondere aus interdisziplinärer Sichtweise relevant sind, wie *presenting data (verbally)*. Diese stehen jedoch nicht in Widerspruch zum hier entwickelten Modell, sondern sind dem unterschiedlichen Detailgrad und einer unterschiedlichen Zielsetzung geschuldet. Auch die fünf von *Ridsdale et al. (2015)* beschriebenen Wissensbereiche werden durch das hier präsentierte Modell gut abgebildet: Bis auf die *data application* stimmen diese größtenteils mit den hier entwickelten Inhaltsbereichen überein. Die Anwendung von Daten wird im hier entwickelten Kompetenzmodell einerseits auf den Bereich Datenethik und Datenschutz reduziert, der für alle Anwendungsbereiche gleichermaßen relevant ist, zusätzlich wird die Berücksichtigung verschiedener Anwendungsbereiche in den Kontext des jeweiligen Unterrichts ausgelagert. Entsprechend kann auch hier kein Widerspruch erkannt werden, obwohl Ridsdale et al. diese Bereiche, aufgrund der Fundierung ihrer Kompetenzen auf interdisziplinären Ansätzen, stärker betonen.

Zusammenfassend zeigt der Vergleich mit den beiden anderen Kompetenzbeschreibungen, dass der hier gewählte Ansatz, trotz eines stark unterschiedlichen methodischen Vorgehens, zu ähnlichen Ergebnissen führt: Alle drei Modelle berücksichtigen größtenteils dieselben Aspekte und weisen keine Widersprüche zueinander auf, unterscheiden sich jedoch geringfügig in ihrer Schwerpunktsetzung. Durch die klare Trennung der Inhalts- und Prozessbereiche liefert das hier entwickelte Modell einen klaren Beitrag zur weiteren Charakterisierung der Data Literacy und geht auf diese Weise über die in den beiden anderen Modellen dargestellten Kompetenzen hinaus. Aus praktischer Sicht ist insbesondere die Darstellung des Modells in Einklang mit den Bildungsstandards Informatik hervorzuheben, die den Lehrkräften im deutschsprachigen Raum oft bekannt ist, sodass dieses Modell für diese gut nutzbar und verständlich ist. Zusätzlich sorgt der methodisch klar nachvollziehbare Ansatz, der auf empirische Arbeiten zurückgeht und seine theoretisch-argumentativen Entscheidungen offenlegt, für eine bessere Nachvollziehbarkeit und Verständlichkeit des entstandenen Modells, das auf dieser Basis auch hinsichtlich verschiedener Zielgruppen spezialisiert werden kann.

Teil IV:

Datenmanagement und Data Literacy im Informatikunterricht

10 Datenquellen für den Informatikunterricht

Wenn die Schlüsselkonzepte des Datenmanagements nun im Informatikunterricht thematisiert bzw. in diesem entsprechende Data-Literacy-Kompetenzen erworben werden sollen, stellen sich verschiedene Herausforderungen, die von den unterrichtenden Lehrkräften bereits bei der Planung des Unterrichts berücksichtigt und gelöst werden müssen: Neben einer notwendigen didaktischen Reduzierung der fachlichen Inhalte, die zum Teil durch komplexe mathematische Grundlagen geprägt sind, stellt die Wahl geeigneter Werkzeuge und Datensätze für die Lehrerinnen und Lehrer oft ein Problem dar. Um diese Herausforderungen und auch die siebte Forschungsfrage aufzugreifen, werden zuerst in diesem Kapitel verschiedene Datenquellen für den Unterricht charakterisiert und diskutiert. Im nächsten Kapitel wird das Thema *Werkzeuge* aufgegriffen und ein Unterrichtswerkzeug zum Thema Datenstromsysteme konzipiert, implementiert, in einem Unterrichtsversuch und durch Diskussion mit Experten evaluiert und weiterentwickelt. Dieser praxisorientierte Teil der Arbeit wird daraufhin in Kapitel 12 mit der Entwicklung und Erprobung einer Data-Mining-Unterrichtssequenz abgeschlossen.

Das Problem, dass für den Unterricht geeignete Datensätze bzw. Datenquellen benötigt werden, ist nicht neu: Im Gegenteil steht der Datenbankunterricht schon länger vor der Herausforderung, dass zur geeigneten Thematisierung interessanter und alltagsnaher Themen geeignete Beispiele und Datenquellen nötig sind. Diese sollen einerseits ausreichend groß sein, um relevante Aspekte demonstrieren zu können, wie beispielsweise die zunehmende Signifikanz von Datenanalysen bei Vergrößerung der Datenbasis. Andererseits sollen sie aber auch an einen motivierenden Realweltkontext anknüpfen und Bezüge zu verschiedenen – möglichst schülernahen – Beispielen eröffnen. Während mit den häufig genutzten Datensätzen, z. B. aus fiktiven Schulbibliotheken oder Ergebnissen der Fußball-Bundesliga, zwar die Möglichkeiten von relationalen Datenbanken noch thematisiert werden können, trifft dies auf viele Möglichkeiten und Vorteile modernerer Bereiche des Fachgebiets Datenmanagement kaum mehr zu. Ohne geeignete Datenquellen kann der Unterricht gerade in Zusammenhang mit moderneren Kontexten seine potenzielle Stärke nicht ausspielen: Soll beispielsweise die korrelationsbasierte Datenanalyse (*Data Mining*) im Unterricht betrachtet werden, sind auf den ersten Blick umfangreiche Datensätze essenziell, die über das bisher üblicherweise genutzte Maß hinausgehen. Diese sind insbesondere dann notwendig, wenn es darum geht, die Fehlerquote der Analyse zu senken, um die Qualität der Analyse oder ggf. Prognose zu erhöhen – und somit die Mächtigkeit von Datenanalysen zu veranschaulichen. Auf diese Weise können (verhältnismäßig) große Datenmengen im Unterricht dazu genutzt werden, zu verdeutlichen, warum Unternehmen heute häufig versuchen, *alles* zu speichern und somit möglichst vollständige Datensätze zu bekommen *(vgl. Mayer-Schönberger und Cukier, 2013)*. Die Nutzung solcher Datenmengen im Unterricht stellt Lehrerinnen und Lehrer jedoch vor die Herausforderung, dass geeignete Datenmengen erfasst bzw. gefunden werden müssen. Die im schulischen Kontext sowieso gespeicherten

Datenmengen (z. B. Informationen über Schülerinnen und Schüler, Lehrpersonal, ...), wären dafür zwar durchaus interessant und ausreichend, eine Verwendung im Unterricht ist aber aufgrund des Daten- und Persönlichkeitsschutzes ausgeschlossen. Somit müssen alternative Datenquellen gefunden werden, wobei verschiedene Kriterien an die Auswahl der Datenquellen angelegt werden können:

- **Freie Verfügbarkeit:** Eine freie Verfügbarkeit der Daten ist nicht nur für die Nutzung im Unterricht praktisch, sondern ermöglicht es auch Schülerinnen und Schülern diese und ähnliche Daten für eigene Zwecke zu nutzen.

- **Nutzung von Realdaten:** Die Verwendung von Realdaten bringt gegenüber fiktiven Datensätzen den Vorteil, dass validere Aussagen getroffen werden können, sodass entsprechende Ergebnisse beeindruckendere Wirkung hinterlassen als künstlich erzeugte Daten, die speziell erstellt wurden, um bestimmte Ergebnisse zu ermöglichen.

- **Angemessene Größe:** Der ausgewählte Datensatz muss, je nach angestrebtem Unterrichtsziel, unterschiedlich groß sein. So ist beispielsweise für die Einführung in eine erste manuelle Datenanalyse oder das Thema Datenbanken ein kleinerer Datensatz genauso gut oder sogar besser geeignet als ein umfangreicherer. Für andere Zwecke, wie die Demonstration der Qualität und des Potenzials von Datenanalysen, ist hingegen ein größerer Datensatz zu bevorzugen – wobei auch hier „größer“ nicht zwingend mehrere Millionen Elemente bzw. einige Gigabyte Größe bedeutet, sondern anwendungsfallabhängig bereits bei wenigen hundert Elementen beginnen kann.

- **Alltagsbedeutung:** Indem ein für die Schülerinnen und Schüler bedeutsamer Alltagsbezug hergestellt wird, was bei vielen Datensätzen möglich ist, wird die Relevanz des Themas für das eigene Leben verdeutlicht.

- **Komplexität:** Als letztes Kriterium spielt sicherlich auch die Komplexität des Datensatzes eine Rolle, da sehr komplexe Datensätze selbst Ursache für Schwierigkeiten im Umgang mit den Daten darstellen können, obwohl sie möglicherweise bezüglich der anderen Kriterien besonders geeignet wären. Die Komplexität muss daher immer dem jeweiligen Unterricht bzw. der Unterrichtsphase angemessen gewählt werden.

Je nach Ziel des Unterrichts und den Anforderungen der Zielgruppe stehen unterschiedliche dieser Kriterien im Vordergrund, sodass es nicht möglich ist, eine einzige ideale Datenquelle zu benennen. Stattdessen stehen heute vielfältige Möglichkeiten zur Datengewinnung zur Verfügung, die die oben skizzierten Kriterien unterschiedlich erfüllen können. Diese können insbesondere in drei Kategorien eingeteilt werden: *Open Data*, *Programmierschnittstellen* verschiedener webbasierter Plattformen (*Web APIs*) und *Mikrocontroller* bzw. *Sensoren*. Diese drei Kategorien werden im Folgenden jeweils kurz beleuchtet.

10.1 Frei verfügbare Datensätze: Open Data

Der Begriff *Open Data* umfasst eine immer größer werdende Menge an frei verfügbaren und nutzbaren Datensätzen, die oft aus öffentlicher Hand stammen. Die auf diese Weise zur Verfügung gestellten Daten könnten vielfältiger nicht sein: von reinen Dokumenten über Auszüge relational organisierter Daten bis hin zu Geodaten und Ähnlichem. Diese Daten werden typischerweise mindestens für die freie Nutzung und Bearbeitung, beispielsweise unter „Creative-Commons"-Lizenzen, zur Verfügung gestellt. Üblicherweise ist auch deren Weitergabe nicht eingeschränkt. Für den Schulunterricht stellen sie eine wertvolle und rechtlich unproblematische Datenquelle dar. Konkrete Beispiele für derartige Plattformen sind das Open-Data-Portal des Bundes[40] (vgl. Abbildung 10.1) oder das Datenportal der Deutsche Bahn AG[41].

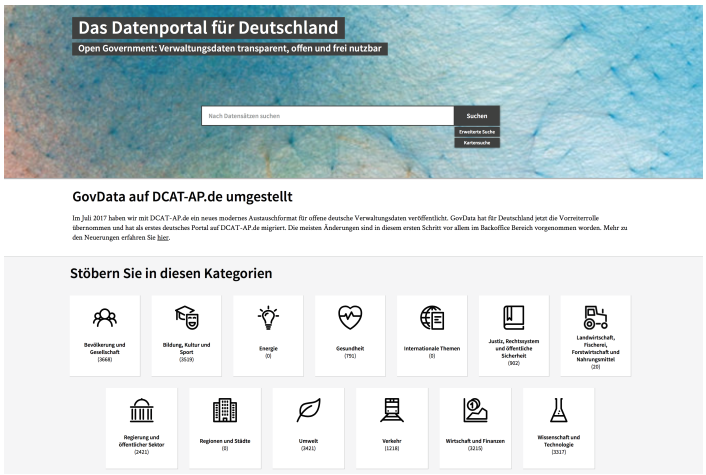


Abbildung 10.1: Open-Data-Portal GovData als Beispiel für vielfältige offene Datenquellen.

Ähnlich gelagerte Datensätze, die jedoch nicht aus öffentlicher Hand stammen und daher streng genommen oft nicht unter *Open Data* gezählt werden, sind heute außerdem im Kontext des Maschinenlernens und der Data Science verfügbar: Beispielsweise stellt die *University of California, Irvine* ein Archiv[42] von Daten für das Maschinelle Lernen zur Verfügung, die nach verschiedenen Kriterien gefiltert werden können. Eine alternative Plattform ist *kaggle*[43], die unter anderem auch eine große Menge verschiedenster Datensätze frei verfügbar anbietet. Diese beiden Webseiten sind nur beispielhaft für eine große Menge an verschiedenen Datensätzen zu sehen, die die zuvor genannten Kriterien unterschiedlich erfüllen, aber mindestens zum Teil für den Informatikunterricht direkt oder nach geringen Abwandlungen (bspw. streichen irrelevanter Attribute) einsetzbar sind. Ein Beispiel für die Nutzung eines solchen Datensatzes folgt in der im übernächsten Kapitel vorgestellten beispielhaften Unterrichtssequenz.

[40] `https://www.govdata.de`
[41] `http://data.deutschebahn.com/`
[42] UCI Machine Learning Repository: `https://archive.ics.uci.edu/ml/index.php`
[43] `http://www.kaggle.com`

10.2 Programmierschnittstellen von Web-Anwendungen: Web-APIs

Eine ähnlich große Vielfalt an Daten kann über Programmierschnittstellen (APIs) diverser Webplattformen oft kostenfrei abgefragt werden. Während die Daten bei Open-Data-Portalen meist direkt heruntergeladen oder in verschiedene Datenanalysewerkzeuge eingebunden werden können, muss der Zugriff auf Web-APIs typischerweise programmatisch erfolgen und ist daher im ersten Moment oft schwieriger. Er wird jedoch auch oft dadurch erschwert, dass häufig die Erstellung eines Nutzeraccounts und die Authentifizierung bei jedem API-Aufruf obligatorisch ist. Relativ einfache APIs, wie die der *OpenWeatherMap* [44] (vgl. Abbildung 10.2) oder die Search-API von Twitter [45] können, durch ihre Ausgestaltung als REST[46]-API und die daraus folgende Ansteuerung über Befehle in Form von Webadressen, trotzdem relativ problemlos zur Datengewinnung, auch live im Unterricht, verwendet werden. Diese Nutzung kann deutlich erleichtert werden, wenn das eingesetzte Werkzeug Unterstützung für standardisierte Zugriffsmöglichkeiten über REST-APIs bietet. Durch die großen und unterschiedlichen Datensätze, die dadurch zur Verfügung stehen bzw. oft kontinuierlich produziert werden, sind solche APIs somit spannende und vielfältige Datenquellen (nicht nur) für den Unterricht. Eine beispielhafte Nutzung der Twitter-API als Datenquelle erfolgt im in Abschnitt 11.2 entwickelten und vorgestellten Werkzeug.

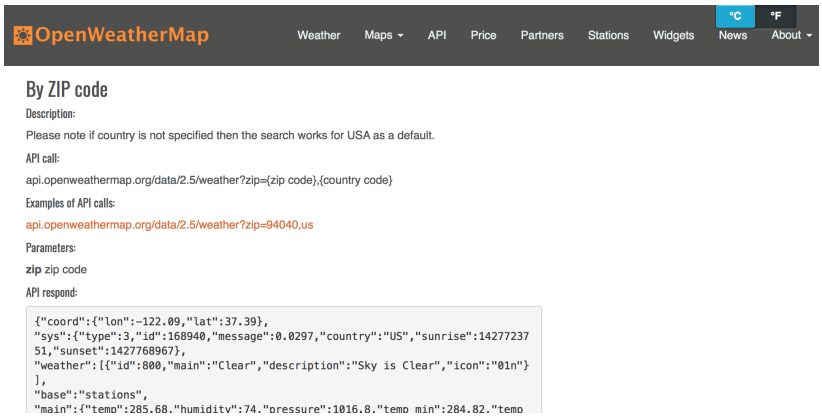


Abbildung 10.2: Beispielhafte Web-API: REST-API der OpenWeatherMap zur Abfrage der Wetterdaten einer bestimmten Postleitzahl.

[44] `http://openweathermap.org/api`
[45] `https://dev.twitter.com/overview/api`
[46] *Representational State Transfer* (REST) ist ein Paradigma, dem heute viele Web-APIs folgen. Die Schnittstelle wird dabei u. A. zustandslos implementiert und ist per HTTP erreichbar, was eine einfache und flexible Nutzung ermöglicht.

10.3 Daten selbst erfassen: Sensoren und eingebettete Systeme

Von den beiden vorher genannten Möglichkeiten unterscheidet sich die Nutzung von Sensoren im Informatikunterricht aus der Perspektive des Datenmanagements insbesondere dadurch, dass die Erzeugung und Erfassung der Daten live im Klassenraum stattfindet (Abbildung 10.3). Dies hat den Vorteil, dass der Unterricht weder von externen Datenquellen abhängig wird, noch ethische oder rechtliche Probleme einer Nutzung der Daten entgegenstehen, auf der anderen Seite jedoch natürlich der Aufwand vor der Nutzung solcher Daten meist höher ist als bei Nutzung anderer Quellen und nicht alle potenziell spannenden Informationen auf diese Weise erfasst werden können. Zur Erfassung von Sensordaten können heute preiswerte Mikrocontroller und Sensoren eingesetzt werden, sodass beispielsweise *Physical-Computing*-Projekte mit Datenmanagementthemen verknüpft werden können – wodurch auch die Verknüpfung dieser beiden für den Unterricht relativ neuen Themen deutlich wird. Durch die Vielfalt an Sensoren, die mit diesen Boards einsetzbar sind, ist die Erfassung verschiedenster Daten denkbar, beispielsweise von einfachen Aspekten wie Umgebungstemperatur oder Helligkeit über die Messung der Luftqualität bis hin zur Erfassung von Bewegungen durch Gyroskope oder unter Nutzung von GPS. Ein beispielhaftes Projekt aus diesem Bereich wird in Abschnitt 11.3 dargestellt.

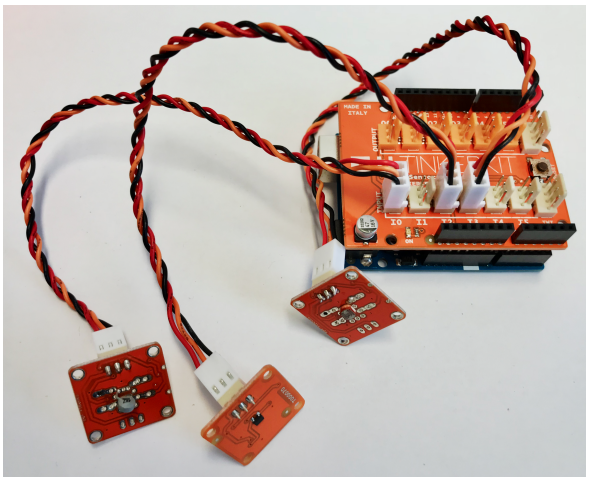


Abbildung 10.3: Erfassung von Sensordaten mithilfe eines Arduino TinkerKit.

11 Entwurf von Werkzeugen für den Informatikunterricht

Neben der Suche nach einer geeigneten Datenquelle stellt auch die Auswahl eines geeigneten Werkzeugs im Informatikunterricht häufig eine Herausforderung für die Lehrkräfte dar. Auch im Kontext des Datenmanagements wird diese Schwierigkeit als relevant erachtet, wie die Ergebnisse der Befragung in Abschnitt 6.1 zeigten. Obwohl für den Datenbankunterricht speziell für den Unterricht entwickelte Umgebungen existieren, z. B. *VideoCenter (Penon, 2013)* oder *FitnessCenter (Penon, 2017)*, werden in diesem häufig profesionelle Werkzeuge eingesetzt, die eine höhere Flexibilität erlauben, da sie nicht auf einen Kontext (beispielsweise Videothek oder Fitnessstudio) eingeschränkt sind. Häufig im Unterricht anzutreffende Vertreter sind insbesondere Kombinationen eines Datenbankservers wie *MySQL* mit einer Verwaltungsanwendung wie *HeidiSQL* oder *phpMyAdmin*. Aber auch Systeme, die keinen dedizierten Server benötigen, z. B. *Microsoft Access* oder *OpenOffice.org Base*, kommen zum Einsatz. Während für den Bereich Datenbanken solche bewährten Systeme zur Verfügung stehen, die trotz ihres professionellen Charakters bei Vernachlässigung einiger der zur Verfügung stehenden Funktionalitäten auch für den Unterricht geeignet sind, ist dies für andere Bereiche des Fachgebiets Datenmanagement weitaus eingeschränkter der Fall: Sowohl zu Datenstromsystemen, Content-Management-Systemen als auch moderneren Datenbanken existieren bisher kaum unterrichtliche Erfahrungen und nahezu keine Werkzeuge, die für den Unterricht geeignet scheinen. So gibt es zwar verschiedene moderne NoSQL-Datenbankserver (z. B. die dokumentenorientierte Datenbank *MongoDB*) und Werkzeuge zu deren Verwaltung (z. B. *Robo 3T*), gegen die bewährte relationale Systeme zwar theoretisch ausgetauscht werden könnten, jedoch sind sie aufgrund ihrer Komplexität und ihres eingeschränkten Bedienkomforts kaum sinnvoll im allgemeinbildenden Unterricht einsetzbar. Noch schlechter sieht es in anderen Bereichen, beispielsweise bei Datenstromsystemen aus, da hier nur wenige frei verfügbare Implementierungen existieren, diese aber zusätzlich eher zu Forschungszwecken entwickelt und eingesetzt werden (z. B. STREAM[47] und Aurora[48]). Aufgrund dieser Ausrichtung und der damit einhergehenden Komplexität und des eingeschränkten Bedienkomforts ist ein Einsatz im Informatikunterricht nicht denkbar. Ähnlich sieht es bei Themen wie Metadaten oder Synchronisierung aus, für die zwar (sogar im Alltag bewährte) Werkzeuge zur Nutzung existieren, aber keine, die den Blick auf die informatischen Grundlagen betonen und zu deren Verständnis beitragen.

Die Entwicklung von für den Unterricht geeigneten Werkzeugen zu verschiedenen Themen des Datenmanagements ist daher eine bisher größtenteils offene Aufgabe, die erst noch erfolgen muss, um ein Fundament für den Informatikunterricht zu diesen Themen zu schaffen. Dabei bietet es sich an, Kriterien an diese Werkzeuge anzulegen, wie sie beispielsweise von Papert bzw. Resnick *(vgl. Resnick und Silverman, 2005)* an konstruktionistischen Unterricht gestellt werden: Insbesondere sollen im Unterricht einsetzbare Werkzeuge eine

[47] `http://infolab.stanford.edu/stream/`
[48] `http://cs.brown.edu/research/aurora/`

niedrige Einstiegshürde (*low floors*) aufweisen, wodurch der Einsatz professioneller Werkzeuge oft ausgeschlossen wird. Gleichzeitig sollen sie aber auch für komplexere Aufgaben geeignet sein (*high ceilings*) und möglichst viel kreativen Freiraum lassen (*wide walls*). In diesem Kapitel wird diese Herausforderung aufgegriffen, indem ein solches Werkzeug für den Informatikunterricht zu einem der zentralen Themen des Datenmanagements entwickelt, in der Diskussion mit Lehrerinnen und Lehrern sowie einer kurzen Unterrichtserprobung evaluiert und daraufhin weiterentwickelt wird.

Als betrachtete Kerntechnologie des Datenmanagements wird hier die Datenstromanalyse herausgegriffen, die bisher im unterrichtlichen Kontext noch nicht betrachtet wurde und zu der es selbst im professionellen Bereich keine flexibel und einfach einsetzbaren Werkzeuge gibt. Um einen relevanten Alltagskontext aufzugreifen, wird das Werkzeug zunächst auf die Thematisierung von Echtzeitanalysen am Beispiel des Twitterdatenstroms hin optimiert.

Bevor jedoch die zugrundeliegenden Ideen des Werkzeugs sowie dessen Entwicklung, Erprobung und Weiterentwicklung dargestellt werden, werden an dieser Stelle die relevanten fachlichen Grundlagen zusammengefasst sowie zur Verortung des Themas Beispiele aus der professionellen Nutzung solcher Werkzeuge skizziert.

11.1 Fachliche Grundlagen

Eine wesentliche Herausforderung im modernen Datenmanagement stellt die schnelle Reaktion auf neue Daten bzw. Veränderungen dar: Beispielsweise ist diese bei einem System zur Tsunamierkennung und -vorwarnung basierend auf der Messung seismischer Wellen essenziell. In diesem Kontext gewinnen *Datenstromsysteme* heute an Bedeutung, die, im Gegensatz zu Datenbanken, Daten nicht dauerhaft, sondern allenfalls kurzfristig speichern. Durch eine sofortige Verarbeitung neu eintreffender Daten erzielen diese Systeme eine wesentlich höhere Geschwindigkeit und ermöglichen somit schnelle Reaktionen, sind durch die nicht notwendige Speicherung aber zugleich auch speichereffizient. Datenstromsysteme betonen daher die *velocity* der Verarbeitung von *Big Data*. Für eine detailliertere Darstellung der Funktionsweise dieser Systeme wird auf Abschnitt 3.2.2 verwiesen.

Neben der Technologie *Datenstromsysteme* stellen die grundlegenden Datenanalysemethoden ein zentrales Thema für die im Folgenden präsentierte Unterrichtsidee dar. Das Beispiel demonstriert, dass Grundlagen von komplexen und anfangs oft schwer erfassbaren Datenanalysen für den Unterricht didaktisch reduziert greifbar gemacht werden können. In der vorgestellten Unterrichtsidee können die drei wichtigsten Datenanalysemethoden *Klassifikation*, *Clusterbildung* und *Assoziation (Ester und Sander, 2000)* thematisiert werden:

- *Clusterbildung* dient dazu, Gemeinsamkeiten zwischen verschiedenen Daten zu finden und diese entsprechend geeigneter Merkmale zu Gruppen/Clustern zusammenzufassen. Für die automatisierte Clusterbildung werden bekannte Verfahren wie beispielsweise der *k-Means-Algorithmus* eingesetzt.

- *Klassifikation* bezeichnet das Einsortieren von Daten in Klassen anhand vorgegebener Merkmale. Von der Clusterbildung unterscheidet sich diese Methode daher dadurch, dass vordefinierte Klassen verwendet werden, anstatt diese erst aus den vorhandenen Daten induktiv zu ermitteln. Zur Klassifikation von Daten werden beispielsweise *Bayes-Klassifikatoren* oder *Entscheidungsbäume* eingesetzt.

- *Assoziationen* werden genutzt, um Zusammenhänge zwischen verschiedenen Merkmalen eines Datensatzes auszudrücken. Diese typischerweise in der Form *wenn–dann* formulierten Zusammenhänge werden häufig eingesetzt, um aus bekannten Merkmalen unbekannte vorherzusagen. Ein beispielhafter Algorithmus zur Assoziationsanalyse ist der Apriori-Algorithmus, den *Berendt et al. (2014)* in einem Unterrichtsbeispiel einsetzen.

Einen wichtigen Einsatzzweck von Datenstromsystemen stellt heute die Überwachung von Datenquellen („*Monitoring*“) dar, häufig mit dem Ziel, auf Veränderungen in den Daten oder auf die Erreichung von Schwellwerten zu reagieren. Gerade im Kontext des Internets der Dinge und speziell der Heimautomation ist dieses Prinzip allgegenwärtig, beispielsweise indem Beleuchtung und Heizung eines Gebäudes automatisch in Abhängigkeit der Anwesenheit von Personen geregelt werden oder Jalousien sich automatisch dem Sonnenstand anpassen. Gleichzeitig sind Datenstromsysteme oft auch eine effiziente Möglichkeit zur Analyse großer Datenmengen ohne diese dauerhaft speichern zu müssen, beispielsweise wenn die Datenmenge für den zur Verfügung stehenden Datenspeicher zu groß ist. Diese beiden Anwendungsfälle werden im Folgenden durch zwei Beispiele weiter charakterisiert.

11.1.1 Überwachung von (Sensor-)Datenströmen

Ein typischer Einsatzzweck ist die *Überwachung* von (Sensor-)Datenströmen. Dazu werden Datenquellen auf eine Annäherung an oder Überschreitung von Grenzwerten hin beobachtet und, beispielsweise durch Auslösung eines Alarms, darauf reagiert. Ein einfaches Beispiel stellt die Überwachung von Webseiten dar: Beispielsweise sollen Performanceeinbußen oder nicht erreichbare Webseiten erkannt und entsprechende Reaktionen ausgelöst werden. Auch aus Perspektive der Endanwender kann dieses Prinzip sinnvoll eingesetzt werden: Unterschreitet der Preis eines Produkts in einem Webshop einen festgelegten Grenzwert oder werden Änderungen an Webseiten durchgeführt, kann der Nutzer benachrichtigt werden. Diesen einfachen Beispielen ist gemein, dass die Daten als kontinuierlicher Datenstrom interpretiert werden, statt einzelne Zeitpunkte zu betrachten. Die meisten Datensätze dieses Datenstroms sind dabei uninteressant, beispielsweise der Preis eines Produkts solange er über dem Grenzwert liegt oder sich vom bisher ermittelten Preis nicht unterscheidet. Statt, wie bei Verwendung einer Datenbank, alle Daten erst einmal zu speichern und im Nachhinein die relevanten Daten im Analyseprozess auszufiltern, ist es in solchen Fällen sinnvoller, auf das Datenstromprinzip zurückzugreifen und den Datenstrom bereits zu filtern, bevor die Daten gespeichert und/oder analysiert werden.

Datenstromsysteme sind für diesen Anwendungsfall gut geeignet, denn sie erlauben es, kontinuierliche Abfragen durchzuführen, die aus allen am System eintreffenden Daten die für den aktuellen Anwendungsfall relevanten ausfiltern und weiterverarbeiten. Ein Beispiel, das Veränderungen des Inhalts einer Webseite erkennt, ist in Abbildung 11.1 dargestellt.

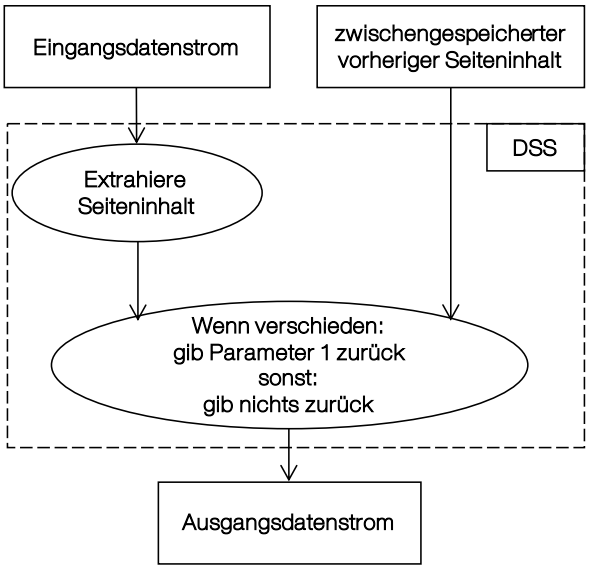


Abbildung 11.1: Analyse von Webseiten unter Nutzung eines Datenstromsystems (DSS).

11.1.2 Analyse des Twitterdatenstroms

Ein weiteres Beispiel stellen soziale Medien dar: Diese spielen heute einerseits eine wichtige Rolle im Leben von Jugendlichen, andererseits sind sie selbst für professionelle Datenanalysen heute eine interessante und viel genutzte Datenquelle. Basierend auf diesen Daten werden heute Trends, Verkaufszahlen von Produkten oder sogar Wahlergebnisse vorhergesagt – teils mit höherer Genauigkeit als mit traditionellen und oft wesentlich aufwändigeren Methoden (vgl. beispielsweise *Beauchamp (2017)*). Insbesondere Twitter kommt dabei eine zentrale Rolle zu: Aufgrund der eingeschränkten Länge der Tweets und der gleichzeitig großen Fülle an Metadaten sind diese meist relativ einfach analysierbar. Außerdem ist Twitter auch in verschiedensten Bevölkerungsschichten und Altersgruppen verbreitet, sodass, im Vergleich mit anderen sozialen Medien, für viele Analysen eine relativ repräsentative Stichprobe vorliegt.

Auf Twitter wurden 2013 pro Sekunde im Durchschnitt 5.700 Tweets veröffentlicht *(vgl. Krikorian, 2013)*. Es kann vermutet werden, dass diese Anzahl seitdem nicht abgenommen hat[49]. Auf die gesamte Datenmenge, die dabei erzeugt wird, kann über die sogenannte *Streaming API* von Twitter in nahezu Echtzeit zugegriffen werden, es ist jedoch nur ein Teil

[49] Verlässliche Zahlen, die die Anzahl der Tweets in einem bestimmten Zeitraum nennen, wurden in den letzten Jahren nicht veröffentlicht. Aktuelle Schätzungen gehen von ähnlichen Werten aus *(bspw. Omnicore Agency, 2018)*.

dieses Datenstroms kostenfrei verfügbar. Nur im kostenpflichtigen *Firehose* stellt Twitter die gesamte Datenmenge zur Verfügung, die frei nutzbaren *Sample Data Stream* und *Filter Data Stream* enthalten hingegen nur eine kleine Stichprobe der Tweets. Für die Schule ist die Größe dieser Stichprobe jedoch leicht ausreichend: in eigenen Tests konnten ca. 30–50 Tweets pro Sekunde abgerufen werden (genaue Zahlen werden hier von Twitter nicht genannt). Neben dem eigentlichen Inhalt enthält jeder Tweet ca. 150 weitere Attribute *(vgl. Dwoskin, 2014)*, die diesem als Metadaten mitgeliefert werden und auch in Datenanalysen einbezogen werden können: Neben einer eindeutigen ID liefert jeder Tweet beispielsweise die Sprache des Tweets mit, aber auch das Land aus dem dieser abgesetzt wurde und wenn möglich genauere Koordinaten des Ortes, aber auch Informationen über den Autor (wie Benutzername, Follower, ID) und sein Profil auf Twitter (wie beispielsweise die gewählte Hintergrundfarbe)[50]. Um die enorme Datenmenge, die auf Twitter zur Verfügung steht bzw. kontinuierlich generiert wird, nach unten abzuschätzen, kann angenommen werden, dass der reine Tweettext (unter der Annahme, dass dieser im Durchschnitt nur ca. 70[51] Zeichen lang ist und UTF-8-kodiert gespeichert bzw. übertragen wird) selbst ohne Metadaten mindestens 200 Bytes groß ist. Eine konservative Schätzung für den gesamten Tweet inklusive aller Metadaten und strukturierender Informationen des genutzten Datenformats beträgt daher sicherlich über 500 Bytes pro Tweet, sodass pro Stunde über 10 GB, pro Tag sogar über 250 GB an Daten entstehen – real wahrscheinlich deutlich mehr[52]. Selbst bei der für die Schule zugreifbaren Datenmenge von ca. 40 Tweets pro Sekunde kommen pro Stunde noch über 70 MB an Daten zusammen. Durch diese große Menge an zur Verfügung stehenden Daten mit umfangreichen Metadaten werden entsprechend interessante Datenanalysen möglich.

11.2 Blockbasierte Analyse des Twitterdatenstroms mit SnapTwitter

Um Datenstromanalysen nun im Informatikunterricht thematisieren zu können, ist, wie zuvor dargestellt, zwingend die Entwicklung eines geeigneten Werkzeugs notwendig, das im Gegensatz zu professionellen Werkzeugen eine für den Schulunterricht angemessene Komplexität besitzt. Diese Konzeption und Entwicklung des Werkzeuges wird im Folgenden vorgestellt, dazu wird als Zielgruppe der Informatikunterricht an allgemeinbildenden Schulen, ca. im Bereich der neunten bis elften Jahrgangsstufe, angenommen. Um eine mög-

[50] Eine komplette Übersicht der Metadaten eines Tweets ist mit Stand 2010 hier verfügbar: `http://online.wsj.com/public/resources/documents/TweetMetadata.pdf` (erstellt durch Raffi Krikorian, ehemaliger Vice President bei Twitter)

[51] Die Schätzung einer durchschnittlichen Länge eines Tweets auf 70 Zeichen entspricht der Hälfte der Maximallänge bis ins Jahr 2017. Da diese Länge mittlerweile verdoppelt wurde, kann davon ausgegangen werden, dass auch die mittlere Länge zunimmt – im Sinne der angestrebten unteren Abschätzung wird jedoch in der Rechnung weiter an den 70 Zeichen festgehalten.

[52] In einem eigenen Test wurden 68.094 Tweets erfasst, die mit strukturierenden Informationen (Attributnamen, JSON-Syntax usw.) knapp 104 MB belegten und somit im Mittel 1,5 KB belegten. Davon ausgehend, dass durch geeignete Speicherung eine mehrfache Speicherung von Attributnamen und ähnlichem vermieden werden kann, scheint die Schätzung auf mindestens 500 Bytes pro Tweet auch hier bestätigt.

lichst flexible Einsetzbarkeit des Werkzeugs zu gewährleisten wird außerdem versucht, möglichst geringe Vorkenntnisse für die Nutzung vorauszusetzen.

11.2.1 Konzeption und Entwicklung

Um die angestrebte möglichst geringe Einstiegshürde im Sinne des *low-floors*-Prinzips zu erreichen, gleichzeitig aber vielfältige Möglichkeiten und Wege bei der Datenanalyse zulassen (*wide walls*) und sowohl einfache als auch komplexe Analysen (*high ceiling*) zu erlauben, wurde entschieden nicht das komplette Werkzeug von Grund auf neu zu entwickeln. Stattdessen wird auf die blockbasierte Programmierumgebung Snap*!* aufgesetzt und diese so erweitert, dass einfache datenstrombasierte Datenanalysen ermöglicht werden. Die Wahl fiel auf dieses konkrete Werkzeug, da einerseits blockbasierte Programmierung heute vielen Schülerinnen und Schülern bereits bekannt und auch ohne Vorkenntnisse relativ einfach beherrschbar ist, sodass die Einstiegshürde sinkt. Andererseits zeichnet sich Snap*!* auch durch eine hohe Flexibilität und einfache Erweiterbarkeit aus, was zugleich die Werkzeugentwicklung unterstützt und vereinfacht, aber auch dafür sorgt, dass den späteren Nutzern vielfältige Möglichkeiten offenstehen. Prinzipiell kann das im Folgenden vorgestellte Konzept jedoch auch auf viele weitere blockbasierte Programmierumgebungen und auch auf die textuelle Programmierung übertragen werden.

Als Datenquelle wurde der Twitter-Datenstrom ausgewählt, da, wie zuvor erwähnt, da mit diesem eine kostenfreie Möglichkeit zum Zugriff auf ausreichend umfangreiche Datenmengen zur Verfügung steht, die live erzeugt werden, sich somit kontinuierlich verändern und entsprechend einen typischen Einsatzzweck für Datenstromanalysen darstellen. Aus den beiden Möglichkeiten zum Zugriff auf die Daten wurde der Filter-Datenstrom ausgewählt, da dieser es erlaubt, eigene Kriterien zu definieren, nach denen die übermittelten Tweets ausgewählt werden, sodass damit die Chance erhöht werden konnte, möglichst viele Tweets mit Standortdaten abzugreifen, die für die Analyse besonders spannend sind. Gleichzeitig lieferte dieser in eigenen Tests geringfügig höhere Datenraten als der Sample-Stream.

Um die notwendigen Funktionalitäten zum Zugriff auf Twitter und zur Analyse der Daten in Snap*!* zu implementieren, wurde ein Ansatz gewählt, der auch auf andere Datenquellen übertragen werden kann: Anstatt den eigentlichen Quellcode von Snap*!* zu modifizieren, wie bei verschiedenen anderen Snap*!*-Erweiterungen üblich (z. B. bei Snap4Arduino), wurden alle neuen Funktionen nativ implementiert, indem die von Snap*!* zur Verfügung gestellten Blöcke und insbesondere die Möglichkeit, JavaScript in Snap*!* zu nutzen, verwendet wurden. Auf diese Weise wird eine flexible Portabilität der Erweiterung auch auf andere Snap*!*-Varianten und eine hohe Kompatibilität zu zukünftigen Updates sichergestellt. Ein Hindernis bei der Umsetzung stellten jedoch die in allen bekannten Webbrowsern implementierten Maßnahmen zur Vermeidung von Cross-Site-Scripting-Attacken dar: Da es sich bei Snap*!* um eine JavaScript-basierte Browseranwendung handelt, unterliegt diese u. a. Schutzmaßnahmen zur Verhinderung von Cross-Site-Scripting, sodass der JavaScript-basierte Zugriff auf andere Webseiten bzw. deren Daten entsprechend eingeschränkt ist. Insbesondere wird ein direkter Zugriff aus Snap*!* auf die Twitter-API somit effizient unter-

bunden und ist innerhalb des Webbrowsers nicht direkt möglich. Um solche Zugriffe zu ermöglichen, müsste die angefragte Webseite (in diesem Fall Twitter) diese Zugriffe explizit erlauben, indem ein CORS[53]-HTTP-Header gesetzt wird, der definiert, welche konkreten Domains/IPs zugreifen dürfen. Da dies praktisch nicht umsetzbar ist, muss der seitenübergreifende JavaScript-Zugriff daher vermieden werden, indem eine lokale Anwendung auf den Client-PCs genutzt wird, die als Proxy fungiert, indem sie die Daten von Twitter abfragt und diese über eine HTTP-basierte REST-Schnittstelle unter Nutzung des entsprechenden HTTP-Headers zur Verfügung stellt, sodass Snap! darauf zugreifen kann (vgl. Abbildung 11.2). Um eine möglichst starke Plattformunabhängigkeit zu erreichen, wurde diese Proxyanwendung in Java implementiert. Sie hat zwei zentrale Funktionalitäten: Einerseits neben dem Abfragen und zur Verfügung stellen des gesamten Filter-Datenstroms von Twitter übernimmt sie zusätzlich die Authentifizierung, die für diesen Zugriff nötig ist. In größeren Gruppen wie Schulklassen ergibt sich dadurch noch ein weiterer Vorteil: Da es möglich ist, diese Anwendung auf einem zentralen Rechner zu starten, der dann als Proxy für die gesamte Gruppe fungiert, muss auch nur auf diesem eine Authentifizierung bei Twitter stattfinden, sodass entsprechend nicht für jeden Schülerrechner eigene Zugangsdaten notwendig sind. Gleichzeitig müssen nicht beliebig viele gleichzeitige Verbindungen zu Twitter aufrechterhalten werden, was die Internetverbindung deutlich belasten würde. Ein Screenshot der Proxyanwendung, die nach dem Starten keinerlei Relevanz für den Nutzenden mehr hat, aber während der Analyse dauerhaft im Hintergrund laufen muss, ist in Abbildung 11.3 dargestellt.

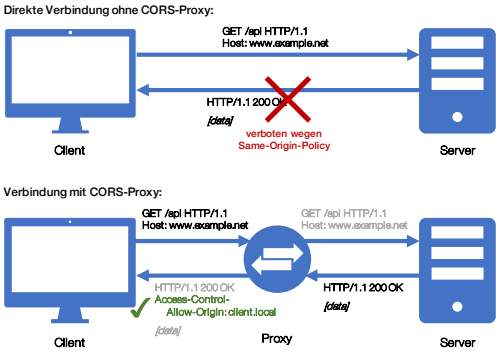


Abbildung 11.2: Schematische Darstellung der Aufgabe des SnapTwitter-Proxys.

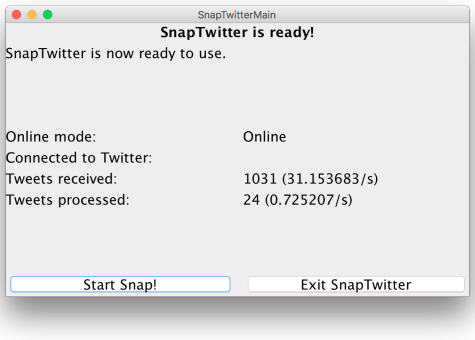


Abbildung 11.3: SnapTwitter-Proxy

Durch Nutzung der Proxyanwendung kann Snap! auf die im JSON[54]-Format vorliegenden Tweets unter Nutzung des standardmäßig zur Verfügung stehenden HTTP-Blocks problemlos zugreifen und diese mithilfe des JavaScript-Blocks interpretieren. Um diese Abläufe vor dem Nutzer zu verstecken, wurden sie in SnapTwitter in Blöcken gekapselt, die aufgrund der nativen Implementierung in Snap! jedoch prinzipiell durch jeden Nutzer

[53] *CORS* bezeichnet das *Cross-Origin Resource Sharing*, einen Mechanismus der es durch Setzen eines HTTP-Headers auf einer Webseite anderen Webseiten erlaubt, trotz der *Same-Origin-Policy* auf diese bzw. ihre Daten zuzugreifen.

[54] Die JavaScript Object Notation (JSON) ist ein strukturiertes Datenformat, das ursprünglich aus der Objektnotation von JavaScript kommt, heute aber vielfältig für den Datenaustausch verwendet wird.

einseh- und bearbeitbar sind, wodurch Erweiterungen und Modifikationen einfach möglich werden. Die implementierten Blöcke basieren insbesondere auf der Abfrage der Daten mittels des Snap!-HTTP-Blocks sowie der Interpretation der erhaltenen JSON-formatierten Tweets und der Extraktion relevanter Daten aus diesen. Dies ist am Beispiel des Blocks *read attribute from tweet* in Abbildung 11.4 dargestellt. Für den Zugriff auf die Twitterdaten sind insbesondere folgende Blöcke von Bedeutung:

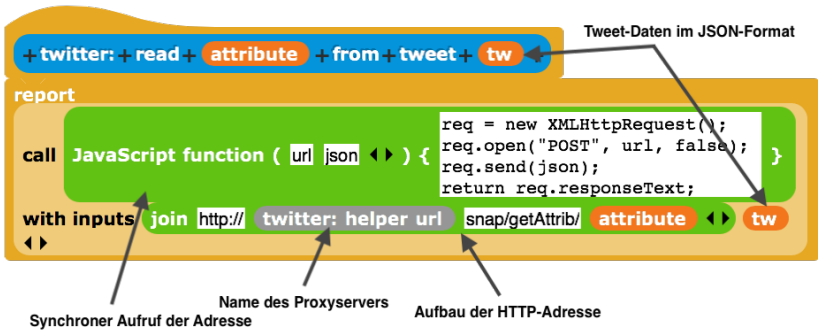


Abbildung 11.4: Implementierung des Blocks *read attribute from tweet* von SnapTwitter.

- twitter: prepare Initialisiert die Proxyanwendung und ruft Twitter-Login auf.
- twitter: connect to stream Stellt Verbindung zum Twitterdatenstrom her (analog: *disconnect*).
- twitter: for each tweet do Verarbeitung der empfangenen Tweets in einer Schleife.
- twitter: set language filter to Vorfilterung des Twitterdatenstroms auf Twitter-Seite nach Sprache (analog: *remove filter* sowie Geodaten- und Schlüsselwortfilter).
- twitter: read from tweet Auslesen eines Attributs des Tweets, wie beispielsweise Text.
- twitter: show tweet on map Anzeige eines Tweets auf einer Karte.

Aufgrund der Nutzung der Proxyanwendung, von der Snap! die Daten aktiv abfragt, weicht die Implementierung von SnapTwitter jedoch an einer Stelle vom Charakter typischer Datenstromanalysen ab: Ein Datenstrom wird normalerweise einmalig abonniert/angefragt, woraufhin die Daten von der Quelle kontinuierlich zum Empfänger gesendet werden, ohne dass dieser erneut aktiv werden muss. Dies ist zwischen Proxyanwendung und Twitter-API auch so der Fall, es kann jedoch kein dauerhafter Kommunikationskanal zwischen dem Proxy und Snap! aufgebaut werden. Entsprechend wurde die Datenabfrage so realisiert, dass die Daten durch Snap! regelmäßig angefragt werden und somit die aktive Kommunikationsrichtung im Vergleich zu Realanwendungen umgekehrt wurde, sodass die Implementierung an dieser Stelle dem Datenstromcharakter widerspricht. Dies ist jedoch für das Konzeptverständnis unproblematisch, da die Auswirkungen dieser Einschränkung rein auf technischer Ebene angesiedelt sind, da der Datenstromcharakter durch die Kombination von Snap! mit der Proxyanwendung so simuliert wird, dass dieser Unterschied für den Nutzer nicht erkennbar ist. Dazu wurde dazu die Proxyanwendung so implementiert, dass

sie die von Twitter empfangenen Daten nur eine minimale Zeitdauer zwischenspeichert und danach sofort wieder verwirft, sodass Daten nur innerhalb dieser Zeit von Snap*!* abgefragt werden und gehen ansonsten – wie in realen Datenstromsystemen, die gerade Daten verarbeiten und nicht bereit sind neue Daten zu empfangen – „verloren“ gehen.

11.2.2 Einsatzmöglichkeiten im Informatikunterricht

Das entwickelte Werkzeug kann auf unterschiedliche Weise und mit verschiedensten Zielen im Informatikunterricht eingesetzt werden. An dieser Stelle wird exemplarisch eine Möglichkeit vorgestellt, bei der es darum gehen soll, dass Schülerinnen und Schüler die Grundzüge von Datenstromanalysen kennenlernen und verstehen. Das Ziel im Unterricht ist daher insbesondere, eigene einfache Datenanalysen durchführen und Schlüsse aus den gewonnenen Ergebnissen ziehen zu können. Dies kann im Unterricht vielfältig kontextualisiert werden: Dabei sind verschiedene Datenanalysen gut geeignet, die Schülerinnen und Schülern heute bereits im Alltag begegnen, beispielsweise bei der Verwendung von sozialen Medien oder beim Einkauf im Online-Versandhandel. Andauernd werden unter anderem Personen hinsichtlich verschiedener Merkmale untersucht, anhand dieser in Klassen eingeordnet bzw. zu solchen zusammengefasst und basierend auf dieser Einordnung Schlüsse gezogen. Beispielsweise werden in Onlineshops Personen anhand ihrer zuletzt gekauften Produkte klassifiziert und dadurch versucht vorherzusagen, welche Produkte sie als Nächstes kaufen, sodass gezielt Werbung platziert werden kann. Andererseits werden aber auch Standortdaten verwendet, um beispielsweise Preisanpassungen für bestimmte Kundengruppen durchzuführen[55]. Um zu verstehen, wie solche Analysen und Vorhersagen überhaupt funktionieren, welches Potenzial und welche Risiken diese bergen und um fundiert entscheiden zu können, ob und in welchem Umfang man Anbietern, die solche Analysen nutzen, vertraut, sind grundlegende Kenntnisse und Erfahrungen in der Datenanalyse nötig. Dabei sind insbesondere die drei grundlegenden Datenanalysemethoden *Klassifikation*, *Clusterbildung* und *Assoziation* zentral, die bereits mit dem vorgestellten Werkzeug einfach und zielführend thematisiert werden können:

Klassifikation. Anhand einfacher Klassifikationsaufgaben können die Lernenden das der Klassifikation zugrundeliegende Prinzip erkennen: die Einteilung von vorliegenden Daten in zuvor definierte Kategorien. Beispielsweise kann dafür die Aufgabe gestellt werden, mit dem zur Verfügung gestellten Tool alle eingehenden Tweets anhand bestimmter Stichworte oder anhand der Sprache, in der sie verfasst wurden, zu klassifizieren. Als Ergebnis kann beispielsweise ein Balkendiagramm (z. B. wie in Abbildung 11.5) erstellt werden.

Die Diskussion der möglichen Aussagen, die aus der Klassifikation gewonnen werden können, offenbart den Schülerinnen und Schülern die Grenzen dieser Methode: Beispielsweise

[55] Das *dynamic pricing* wird regelmäßig in Onlineshops verwendet, ist aber nicht unumstritten, da es in diesem Fall nicht – wie im stationären Handel – alle Kunden gleichermaßen betrifft, sondern auf den Einzelnen abgestimmte Preise ausgerufen werden können (vgl. z. B. *Österreichisches Institut für angewandte Telekommunikation (2015)*).

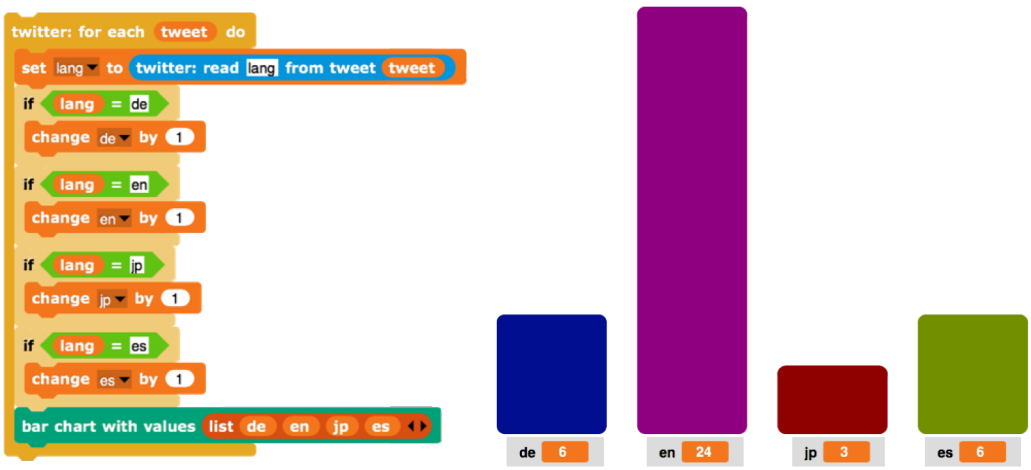


Abbildung 11.5: Klassifikation der Tweets nach Sprache, dargestellt als Balkendiagramm in Snap!.

kann durch Klassifikation problemlos analysiert werden, welche Produktgruppe beliebter ist als eine andere, woher die meisten Käufer kommen, und Ähnliches. Die zuvor beschriebenen Schlussfolgerungen, die heute aus Daten gewonnen werden können, gehen jedoch weit über dieses Beispiel hinaus: Es geht nicht nur darum zu entdecken, welches Produkt am beliebtesten ist, sondern darum, welches Produkt in welcher Region bevorzugt wird – es wird eine weitere Dimension eingeführt. Eine Kategorisierung nach mehreren Dimensionen wäre zwar möglich, die Anzahl der Kategorien explodiert dabei jedoch, da jede Kategorie mit jeder anderen kombiniert werden kann. Zusätzlich können in solchen Fällen nicht immer schon vor der Analyse die Kategorien festgelegt werden, beispielsweise muss die Region hier nicht zwingend administrativen Regionen entsprechen. Somit wird in solchen Fällen eine Klassifikation beliebig komplex beziehungsweise sogar unmöglich.

Clusterbildung. An dieser Stelle setzt die Clusterbildung an: Daten werden dort anhand ähnlicher Merkmalsausprägungen zu nicht vorher definierten Gruppen zusammengefasst. Trotz der komplexen mathematischen Grundlagen vieler Clusterverfahren, kann eine einfache Clusteranalyse bereits mit dem hier entwickelten Werkzeug bewältigt werden: Nach der Visualisierung einer bestimmten Eigenschaft auf der Karte, können die Cluster intuitiv optisch bestimmt und so erkannt werden, welche der Ausprägungen dieser Eigenschaft in verschiedenen Regionen vorherrscht. Obwohl insbesondere die Mächtigkeit dieses Vorgehens nicht der automatischen Clusterbildung entspricht, hilft diese intuitive Herangehensweise dabei, das Prinzip der Clusterbildung nachzuvollziehen, ohne die mathematischen Grundlagen verstehen zu müssen. In Abbildung 11.6 wird dieser Vorgang beispielhaft dargestellt. Bereits dieses einfache Beispiel zeigt dabei einige grundlegende Fragestellungen der Clusterbildung: *Sollen eher große und damit ungenauere Cluster erzeugt oder kleinere und genauere? Ab wann wird eine Ausprägung einer Eigenschaft als in einem Cluster vorherrschend charakterisiert? Welchen Fehlergrad bin ich bereit einzugehen?*

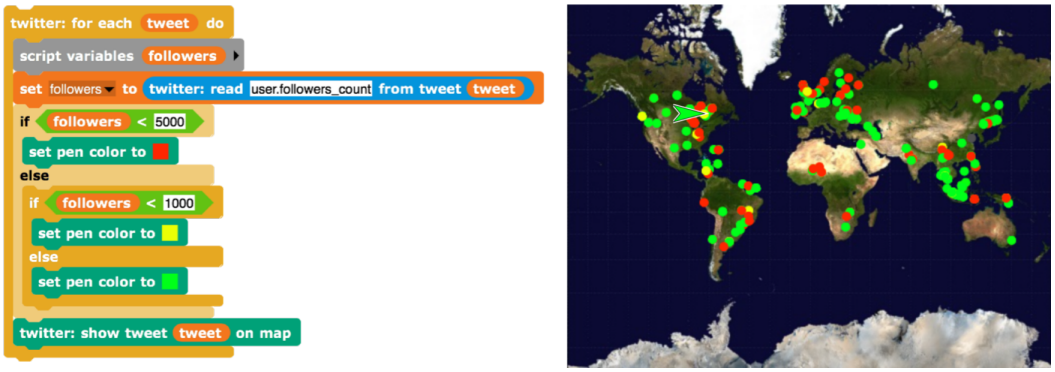


Abbildung 11.6: Visualisierung von Tweets auf einer Karte in Snap!. Die Farbe der Punkte entspricht der Follower-Zahl (rot: unter 500, gelb: mindestens 500, aber unter 1000, grün: über 1000).

Assoziation. Wenn anhand von Daten Entscheidungen oder Vorhersagen getroffen werden sollen, müssen für diese entsprechende Regeln definiert bzw. gefunden werden. Diese Regeln werden in Form von Assoziationen aufgestellt, die Korrelationen zwischen den Daten beschreiben. Eine Assoziationsanalyse kann sowohl manuell als auch automatisch stattfinden, wobei im letztgenannten Fall verschiedene Algorithmen eingesetzt werden können. Diese sind jedoch stark mathematisch geprägt und aufgrund ihrer Komplexität für Schülerinnen und Schüler nur eingeschränkt nachvollziehbar. Trotzdem kann auch die Assoziation mit den Lernenden thematisiert werden, indem potenzielle Zusammenhänge bzw. Regeln in Form von Assoziationen, die sich aus den obigen Betrachtungen ergeben, mit den Schülerinnen und Schülern auf ihre Aussagekraft hin untersucht werden: Mit Blick auf die in Abbildung 11.6 dargestellte Karte, scheint beispielsweise die Assoziation „Wer Twitter in Westeuropa, Südostasien oder den USA nutzt, hat mindestens 1000 Follower“ mit relativ geringem Fehler zutreffend zu sein. Diese Aussage vernachlässigt jegliche Kausalität (was im Sinne von korrelationsbasierten Analysen zulässig und üblich ist): Es wird nicht weiter überlegt, warum beispielsweise in Afrika, Australien oder Russland Twitter anscheinend kaum genutzt wird. Da es sich bei der Auswertung jedoch um eine nicht-repräsentative Stichprobe handelt (es wurden nur solche Tweets betrachtet, die Geodaten offenbaren) und diese auch sehr klein ist und einen Ausschnitt zu einer bestimmten Tageszeit zeigt, kann die Gültigkeit dieser Regel stark angezweifelt werden. Andere Assoziationen, beispielsweise ein Schluss von der Sprache des Tweets auf dessen Herkunftsland scheint jedoch in vielen Fällen auch bei längerer Überprüfung mit wenigen Ausnahmen (z. B. Weltsprachen wie Englisch) zutreffend zu sein und kann damit eine Analyse bereichern, indem diese Assoziation dabei hilft, das Herkunftsland auch in solchen Fällen berücksichtigen zu können, in denen diese Information ansonsten fehlen würde. Diese Grenzen der Aussagekraft von Datenanalysen müssen in diesem Zusammenhang den Schülerinnen und Schülern unbedingt bewusst werden, um falsche Schlüsse und somit auch das Vorurteil der Allwissenheit und Unfehlbarkeit von Big Data zu vermeiden, aber auch um die Bedeutung eines möglichst vollständigen Datensatzes bei Big-Data-Analysen nachvollziehen zu können.

11.2.3 Erprobungen des Werkzeugs und Erfahrungen

Erprobung im Informatikunterricht

Im Rahmen einer Erprobung durch einen Lehramtsstudenten konnten erste Erfahrungen mit dem implementierten Softwarewerkzeug gewonnen werden. Dabei waren aus Sicht der Forschung insbesondere folgende Fragestellungen interessant:

- Wie motiviert und interessiert erscheinen die Schülerinnen und Schüler bei diesem Thema?
- Ist das entwickelte Werkzeug für die Schülerinnen und Schüler intuitiv nutzbar?
- Treten Probleme bei der Verwendung des Werkzeugs im Unterricht auf?
- Inwiefern ist das Werkzeug geeignet, um im Unterricht moderne Datenanalysen zu thematisieren?

Da diese Erprobung jedoch keine detaillierte Evaluation des Werkzeugs im Unterrichtskontext darstellen sollte, sondern insbesondere die Machbarkeit eines solchen Unterrichts evaluieren und Ansätze für die Weiterentwicklung des Werkzeugs offenbaren sollte, wurde diese mit nur einer einzelnen Schulklasse und entsprechend einer Lehrkraft durchgeführt.

Rahmenbedingungen. Die Erprobung wurde in einer 11. Klasse eines Berliner Gymnasiums durchgeführt. Als Lehrkraft stand dabei ein Informatik-Lehramtsstudent einer Berliner Universität zur Verfügung, der keinen weiteren Bezug zu dem in dieser Arbeit beschriebenen Forschungsprojekt hat. Da der Unterricht aus organisatorischen Gründen in nur einer Doppelstunde stattfinden musste, wurde dieser in stark geraffter Form geplant, sodass dieser sicherlich nur einen vagen Einblick in die Thematik der Datenstromanalyse geben konnte. Der Fokus lag daher darauf, dass die Schülerinnen und Schüler Beispiele für die moderne Datenanalyse kennenlernen, die beiden Datenanalysemethoden *Klassifizierung* und *Clusterbildung* erklären und die durchgeführten Analysen aus ethischer Sicht kritisch hinterfragen sollten. Auf die explizite Thematisierung von Assoziationen und die Gewinnung von Vorhersagen wurde aus zeitlichen Gründen verzichtet, stattdessen wurden solche Aspekte nur im Rahmen der am Ende vorgesehenen Diskussion angeschnitten. Die Unterrichtseinheit wurde nach dem regulären Datenbankunterricht als Abschluss der Unterrichtssequenz angesiedelt, in der insbesondere relationale Datenbanken und SQL thematisiert wurden. Die Erfahrungen des Durchführenden wurden im Nachgang im Rahmen eines semistrukturierten Interviews erfasst und werden im Folgenden als Fallstudie wiedergegeben.

Unterrichtsablauf. Der geplante Unterricht wurde durch den Lehrer wie folgt geplant und durchgeführt:

1. **Einstieg:** Der Einstieg erfolgte über eine Eigeneinschätzung der Schülerinnen und Schüler zu den Fragen *„Wie wichtig ist Big Data für meine Zukunft?“* und *„Wie sehr interessiert mich Big Data?“*. Dazu wurden die Schülerinnen und Schüler aufgefordert, sich im Klassenraum in Diagrammform aufzustellen, wobei jede/-r sich als Datenpunkt in einem Koordinatensystem betrachten sollte, das durch die obigen Fragen als Achsen aufgespannt wurde (schematische Darstellung vgl. Abbildung 11.7).

2. **Begriffsklärung *Klassifikation* und *Clusterbildung*:** Das *lebende Diagramm* wurde direkt weiterverwendet, um die Begriffe *Klassifizierung* und *Clusterbildung* einzuführen und zu charakterisieren. Dazu wurde versucht Gruppen zu erkennen, d. h. die „Daten“ zu clustern, und die Schülerinnen und Schüler in diese einzuteilen (zu *klassifizieren*). So konnten die Lernenden bereits einen ersten Einblick in das Ziel dieser Methoden gewinnen und erkennen, welche Entscheidungen bei diesen beiden Methoden nötig sind bzw. wie diese manuell ablaufen können.

3. **Werkzeugeinführung:** Im Rahmen einer kurzen Lehrerdemonstration wurde den Schülerinnen und Schüler die Snap!-Erweiterung als Werkzeug vorgestellt. Dabei wurde an einem einfachen Beispiel demonstriert, wie Datenanalysen u. a. mit den schon kennengelernten Methoden selbstständig durchgeführt werden können. Diese Demonstration wurde bewusst sehr knapp gehalten, um den Schülerinnen und Schülern möglichst viel Zeit für die praktische Anwendung in den folgenden beiden Phasen geben zu können.

4. **Einstiegsaufgabe:** Ein vorgegebenes einfaches Einstiegsbeispiel, bei dem das Ziel war alle Tweets auf einer Karte darzustellen, sollte durch die Schülerinnen und Schüler im Rahmen der ersten Aufgabe so erweitert werden, dass die dargestellten Tweets je nach Sprache unterschiedliche Farben zugewiesen bekamen. Auf diese Weise konnten die Schülerinnen und Schüler erste Eindrücke der verwendeten Programmierumgebung gewinnen und deren Bedienung kennenlernen, während gleichzeitig die Einstiegshürde relativ gering gehalten wurde.

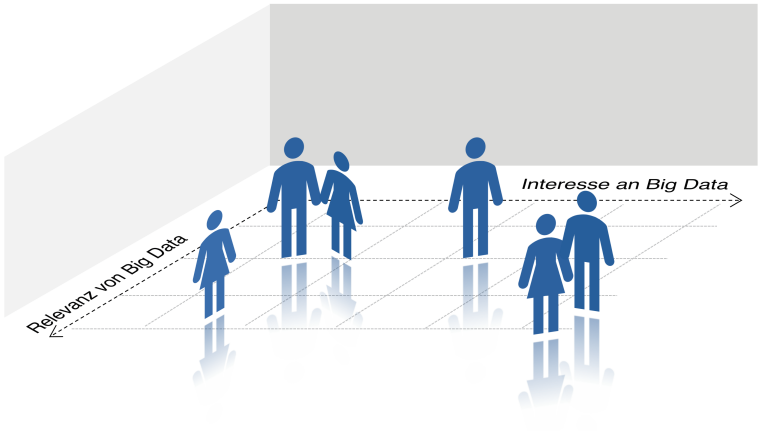


Abbildung 11.7: Erfassung des Interesses an und der Einschätzung der Relevanz von Big Data durch Aufstellung der Schülerinnen und Schüler als Datenpunkte in Diagrammform.

5. **Forschungsaufgabe:** Um den Schülerinnen und Schülern nun einen kreativeren Umgang mit dem Werkzeug zu erlauben, sollten sie sich selbst eine Forschungsfrage aus einem Katalog von Beispielfragen auswählen oder eine eigene Frage entwickeln, der sie im Anschluss unter Nutzung einer Twitterdatenanalyse nachgingen. Diese Fragen waren so auf die Zielgruppe ausgerichtet, dass sie von den Schülerinnen und Schülern zwar erfolgreich beantwortet werden konnten, sie aber dennoch forderten.
6. **Diskussion:** Zum Abschluss wurden die Möglichkeiten und Gefahren von Datenanalysen kritisch hinterfragt und diskutiert, wobei die Schülerinnen und Schüler insbesondere auch zu ethischen und moralischen Aspekten ihrer jeweiligen Forschungsfragen Stellung nehmen sollten.

Die erste bis dritte Phase nahmen dabei zusammen ca. 25 Minuten ein, für die abschließende Diskussion standen ca. 10 Minuten zur Verfügung, während die verbleibenden ca. 55 Minuten als aktive Arbeitszeit auf die Phasen vier und fünf entfiel. Die Arbeit an den beiden Aufgaben erfolgte unter Nutzung der vorgestellten Snap!-Erweiterung. Dazu wurde die Proxyanwendung auf einem der Computer eingesetzt, der als Server die Tweets von Twitter entgegennahm und allen Schülercomputern als Datenquelle offenstand. Dabei wurden keine Performanceprobleme o. Ä. verzeichnet. Da an der Schule eine sehr restriktive Firewall eingesetzt wurde, konnte jedoch kein Live-Zugriff auf die echten Twitterdaten erfolgen, stattdessen wurde mithilfe der Proxyanwendung ein Cache einer ausreichend großen Menge an Tweets erstellt und diese an Snap! als Simulation eines Livebetriebs ausgeliefert. Den Schülerinnen und Schülern wurde dieser Fakt vorerst nicht kommuniziert, sodass diese den Eindruck hatten, auf Livedaten zu arbeiten.

Erfahrungen und Einschätzung des Durchführenden. Allgemein konnte der Lehrer feststellen, dass die Schülerinnen und Schüler prinzipiell sehr am Themengebiet Datenanalyse und Big Data interessiert waren. Insbesondere bei der freien Bearbeitung der Forschungsaufgabe war eine hohe Motivation und der Versuch, mit den dabei auftretenden Herausforderungen zurechtzukommen, erkennbar. Die Arbeit mit den augenscheinlichen Livedaten eines bekannten sozialen Netzwerkes war für die Schülerinnen und Schüler ein besonderer Motivationsfaktor, was insbesondere auch erkennbar war, als am Ende der Stunde aufgeklärt wurde, dass es sich aufgrund der Firewallrestriktionen nicht um aktuelle Livedaten handelte. Im ersten Moment war dabei klare Enttäuschung wahrnehmbar, wodurch jedoch auch klar wurde, dass gerade der Zugriff auf echte Livedaten motivierend wirkt.

Das eingesetzte Werkzeug schien für die Schülerinnen und Schüler sehr gut verständlich und ohne große Einarbeitungsphase einsetzbar zu sein, obwohl sie bisher keine Erfahrungen mit Snap! hatten. Der Lehrer hatte, auch unter Berücksichtigung seiner vorherigen Erfahrungen mit dem Thema Big Data, den Eindruck, dass das Werkzeug für den Unterricht am Gymnasium geeignet didaktisch reduziert wurde, ohne zentrale Aspekte zu vernachlässigen. Gerade die Nutzung von Twitter, einem Standardbeispiel in der Big-Data-Analyse, schien für ihn dabei genauso reizvoll wie für die Schülerinnen und Schüler. Im Rahmen einer auf den Unterricht folgenden Feedbackrunde beklagten die Schülerinnen

und Schüler jedoch auch technische Aspekte, insbesondere dass das Werkzeug aufgrund der Verwendung von Blöcken statt Code teils als störend bzw. umständlich empfunden wurde. Auch wurde die Beschränkung auf ein Ausgabefenster (die Bühne von Snap!) als zu einschränkend empfunden.

Hervorzuheben war aber insbesondere das Ergebnis des Unterrichts: Trotz der sehr eingeschränkten Zeit meisterten die meisten Schülerinnen und Schüler die gestellten Aufgaben zumindest soweit, dass sie erste Ergebnisse erhielten und damit Einblicke in die Datenanalyse mithilfe von Datenstromsystemen gewinnen konnten. Besonders hervorzuheben war die selbstgewählte Forscherfrage eines Schülerpaars: Dieses hatte es sich zum Ziel gesetzt, eine wissenschaftliche Studie zu überprüfen, die besagt, dass die Menschen in den USA ihr soziales Leben wesentlich stärker ins Virtuelle verlagert haben, als beispielsweise in Asien[56]. Obwohl die Möglichkeiten zur Überprüfung der Studie insbesondere aufgrund der geringen Zeit im Unterricht natürlich eher beschränkt waren, haben diese Schüler einen guten Ansatz gefunden und dessen Schwächen diskutiert, sodass selbst diese kurze Unterrichtseinheit ihnen das Gefühl vermitteln konnte, dass sie selbst die Möglichkeit haben, Informationen nicht mehr nur zu glauben, sondern auch durch eigene Analysen zu überprüfen und kritisch zu hinterfragen, was ein wichtiges und positives Ergebnis ist.

Evaluation des Werkzeugs in Lehrerfortbildungen

Zur Ergänzung der in der dargestellten Erprobung gewonnenen Erfahrungen wurden außerdem die Einschätzungen von weiteren Lehrkräften gesammelt: In neun Fortbildungsworkshops im Rahmen verschiedener Veranstaltungen konnten die Erfahrungen und Einschätzungen von über 200 Lehrkräften unterschiedlicher Schulen und verschiedener deutscher Bundesländer gewonnen werden. Dazu wurde im Rahmen dieser Workshops eine Einführung in das Themengebiet Big Data und Datenmanagement gegeben und, als Idee zur Umsetzung im Unterricht, das Werkzeug SnapTwitter sowie Beispielaufgaben für die Schülerinnen und Schüler präsentiert und den Lehrerinnen und Lehrern die Möglichkeit gegeben, dieses Werkzeug selbst auszuprobieren.

In einer darauffolgenden Diskussion konnten verschiedene Ideen zur Weiterentwicklung gewonnen, gleichzeitig aber auch die Bedenken, welche die Lehrerinnen und Lehrer bei der Nutzung des Werkzeugs haben, identifiziert werden. Insbesondere haben verschiedene Lehrerinnen und Lehrer aus unterschiedlichen Gruppen unabhängig voneinander die Nutzung von Twitter als Datenbasis nicht nur als sehr interessant, sondern gleichzeitig auch als möglicherweise problematisch eingeschätzt: Neben der Notwendigkeit einer stabilen und nicht zu langsamen Internetverbindung für die Live-Analyse wurde grundsätzlich hinterfragt, ob die Verwendung der Twitterdaten für den Unterricht geeignet ist oder nicht. Insbesondere wurden dabei potenzielle Datenschutz- und allgemeine rechtliche Probleme gesehen, beispielsweise da Tweets nicht bzw. nur aufwendig vorgefiltert werden können und dadurch potenziell kritische Inhalte eines Tweets den Schülerinnen und Schülern

[56] Die entsprechende Studie konnte leider trotz umfangreicher Recherche nicht gefunden werden.

während des Unterrichts begegnen. Zusätzlich existieren in verschiedenen Bundesländern Richtlinien, die die Nutzung von sozialen Medien im Unterricht – auch wenn didaktisch begründet – stark einschränken[57]. Während dem ersten Kritikpunkt durch Schaffung einer Möglichkeit zur Offlinenutzung (wie sie auch im Rahmen der Unterrichtsevaluation bereits genutzt wurde) begegnet werden konnte, schränkt der zweite Kritikpunkt die Nutzbarkeit im Unterricht ggf. deutlich ein. Im Rahmen einer Entwicklung eines zweiten Werkzeugs wurde daher auch dieser Aspekt aufgegriffen und eine flexiblere Auswahl der Datenquelle ermöglicht.

11.3 Das weiterentwickelte Werkzeug Snap!DSS

Basierend auf den Ergebnissen der Erprobung wurde eine weitere Variante des Werkzeugs entwickelt, die insbesondere das Ziel hat, auch diesen zweiten Nachteil der Twitter-Datenstromanalyse mit Snap*!* aufzugreifen und somit eine in Bezug auf die Datenquelle flexibler einsetzbare Alternative zu bieten. Es wird daher die Festlegung und Optimierung auf eine konkrete Datenquelle aufgegeben. Stattdessen wird die Möglichkeit eröffnet, verschiedenste Datenquellen zu nutzen, solange diese durch *reporter*-Blöcke[58] in Snap*!* nutzbar sind. Zu diesen zählen unter anderem Web-APIs, die über REST-Aufrufe und somit unter Nutzung des HTTP-Blocks und ggf. zusätzlicher Verarbeitung abfragbar sind, aber auch Sensordaten, für die entsprechende Abfrageblöcke in speziellen Snap*!*-Derivaten wie *Snap4Arduino* zur Verfügung stehen. Durch die allgemeiner gehaltene Datenschnittstelle kann jedoch das vorher entwickelte und vorgestellte Werkzeug nicht komplett ersetzt werden, da es sich bei der Twitter-API um eine komplexere API handelt, die auch aufgrund des oben thematisierten *Cross-Origin-Sharing* nicht ohne die zwischengelagerte Proxyanwendung oder alternative ähnliche Lösungen abgefragt werden kann. Zusätzlich wird jedoch in der weiterentwickelten Variante auch der Charakter einer Datenstromanalyse besser gewahrt als bei der ursprünglichen Variante, indem auf die zwingende Verwendung einer Proxyanwendung verzichtet wird, die beispielsweise beim Zugriff auf Sensordaten nicht nötig ist.

11.3.1 Konzeption und Entwicklung

Neben den zuvor genutzten Twitterdaten existieren vielfältige weitere dynamische Datenquellen, die sich insbesondere in die beiden in Kapitel 10 erwähnten Kategorien *Web-APIs* und *Sensordaten* aufteilen[59]. Um beide Arten von Datenquellen im Unterricht flexibel einsetzen zu können, wird ein Werkzeug benötigt, das einerseits flexibel einsetzbar ist, andererseits aber trotzdem keine unnötig hohen Hürden bei der Nutzung aufweist. Daher wurde

[57] Beispielsweise in Bayern: *„Von einer unterrichtlichen Nutzung sozialer Netzwerke ist mit Blick auf die besondere Schutzbedürftigkeit der Schülerinnen und Schüler abzusehen." (Bayerisches Staatsministerium für Unterricht und Kultus, 2012)*

[58] Blöcke die Werte zurückliefern werden in Snap*!* als *reporter* bezeichnet.

[59] Die dritte in Kapitel 10 erwähnte Kategorie, *Open Data*, ist üblicherweise statisch oder in Form von Web-APIs zugreifbar und wird daher hier nicht weiter betrachtet.

auch hier wieder die Programmierumgebung Snap*!* als Basis gewählt. Das implementierte Werkzeug unterscheidet sich jedoch konzeptionell deutlich vom zuvor entwickelten SnapTwitter: Um einen flexiblen Zugriff auf verschiedene Datenquellen zu erlauben, wurde eine komplett neue und wesentlich universellere Implementierung des Datenstromsystems vorgenommen, die die Datenquelle soweit abstrahiert, dass jeder *reporter*-Block in Snap*!* genutzt werden kann. Zusätzlich implementiert das Snap*!*-Datenstromsystem Snap*!*DSS alle zentralen Funktionen von Datenstromsystemen und der üblichen Abfragesprache *CQL*. Im Vergleich zu SnapTwitter werden die Abfragen in dieser deklarativ statt imperativ formuliert, wodurch alle Anfragen als eigene Einheit betrachtet werden können, die auch ineinander verschachtelbar sind. Trotz dieser umfangreichen Möglichkeiten wird dieses Werkzeug jedoch, genau wie SnapTwitter, rein auf Funktionalitäten die Snap*!* standardmäßig mitliefert aufgebaut, sodass die Erweiterung prinzipiell in jedem beliebigen Derivat von Snap*!* einsetzbar ist. Dies wird insbesondere bei der Verwendung von Sensoren als Datenquelle wichtig: Während Snap*!* selbst keine Möglichkeit zur Kommunikation mit solchen mitbringt, wird dies beispielsweise in *Snap4Arduino* ermöglicht, mit dem die hier vorgestellte Erweiterung auch problemlos genutzt werden kann.

Das implementierte Werkzeug bietet daher folgende Kernfunktionalitäten:

- Nutzung mit jedem Derivat von Snap*!* möglich, das keine funktionalen Einschränkungen gegenüber Snap*!* vornimmt.
- Ermöglichung der zentralen Aspekte der Datenstrom-Anfragesprache *CQL*.
- Erzeugung eines Datenstroms aus einer beliebigen in Snap*!* zugreifbaren Datenquelle (Snap*!* reporter block).
- Kombination verschiedener Datenquellen bzw. Werte einer Datenquelle in einem Datenstrom, sodass mehrere Werte parallel betrachtet werden können.
- Ausführen von Anfragen auf Datenströmen unter Nutzung von Projektionen, Selektionen, Aggregationen und Schiebefenstern[60].
- Nutzung von Ausgangsdatenströmen einer Analyse als Eingangsdatenstrom für eine weitere (Verschachtelung von Anfragen).
- Kontinuierliche Auswertung der Anfragen im Hintergrund.

Die Implementierung dieser Funktionalitäten fand in Snap*!* in Form von *unevaluated blocks* statt. Dadurch konnte eine einfache Bedienung des Systems über relativ wenige Blöcke, die in ihrer Bezeichnung an die CQL angelehnt sind, erreicht werden. Die interne Struktur ist jedoch vergleichsweise komplex: Die drei Datenstrukturen *Datenstromsystem*, *Datenstrom* und *Abfrage* sind intern über mehrdimensionale Listen repräsentiert (vgl. beispielsweise Abbildung 11.8). Dabei wird ein Datenstrom intern als Liste gespeichert, die an definierten

[60] Bei *Schiebefenstern* (engl. *sliding windows*) handelt es sich um eine besonders bei der Datenstromanalyse verbreitete Technik, die es ermöglicht, in kontinuierlich ausgeführten Abfragen relative anstatt absoluter Selektionskriterien anzugeben. Dies ermöglicht flexiblere Anfragen, sodass beispielsweise eine Auswahl aller Datensätze der letzten Stunde möglich ist, ohne diese Stunde konkret benennen zu müssen.

Positionen Informationen über seine Datenquellen und die bisher auf dem Datenstrom durchgeführten Abfragen speichert. Eine Abfrage speichert alle ihre Parameter sowie die für diese Abfrage relevanten Werte des Datenstroms. Ein Datenstromsystem merkt sich den aktuellen Status (gestartet oder nicht) sowie die zugeordneten Datenströme und aktualisiert im Hintergrund regelmäßig (standardmäßig einmal pro Sekunde) alle Abfragen aller zugehörigen Datenströme, sodass die Werte der Abfragen immer auf dem aktuellen Stand bleiben, auch wenn sie gerade nicht aktiv durch Abruf des Ergebnisses einer Abfrage gelesen werden.

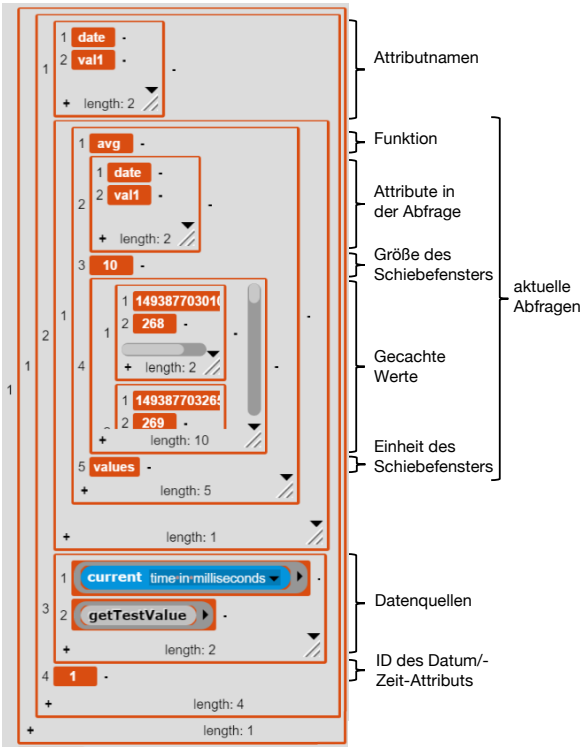


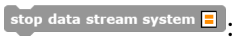
Abbildung 11.8: Interne Repräsentation eines Datenstroms in Snap!DSS.

Diese relativ komplexe Struktur wird durch die folgenden Blöcke für die Schülerinnen und Schüler zugänglich gemacht:

- create data stream system: Erzeugt ein neues Datenstromsystem und gibt einen Verweis auf dieses zurück.
- new data stream from ( ): Erzeugt einen neuen Datenstrom basierend auf den übergebenen *reporter*-Blöcken und gibt einen Verweis auf diesen zurück.
- add stream to data stream system: Fügt einen übergebenen Datenstrom als Eingangsdatenstrom zum Datenstromsystem hinzu.
- select ( ) from stream [last values]: Führt eine Abfrage auf einem Datenstrom unter Nutzung des Datenstromsystems aus und gibt dessen Ergebnisse als Verweis auf

einen Datenstrom zurück. Bei der ersten Ausführung wird die Abfrage im System gespeichert und ab diesem Zeitpunkt kontinuierlich im Hintergrund aktualisiert.

- 

: Startet das Datenstromsystem bzw. die Hintergrundverarbeitung.
- 

: Stoppt das Datenstromsystem bzw. die Hintergrundverarbeitung.

Durch die Implementierung in Snap! unter ausschließlicher Nutzung der Möglichkeiten, die dieses Werkzeug nativ bietet, kann das Snap!DSS trotz seiner umfangreichen Möglichkeiten durch Lehrerinnen und Lehrer aber auch Schülerinnen und Schüler individuell angepasst werden, beispielsweise indem die Blöcke umbenannt, versteckt oder in neuen Blöcken verwendet werden. Eine Anpassung an spezielle Datenquellen ist damit beispielsweise möglich, indem Funktionalitäten in neuen Blöcken gekapselt und so beispielsweise mit reduzierter Schnittstelle nach außen zur Verfügung gestellt werden.

Damit die Daten nicht nur textuell dargestellt, sondern auch visualisiert werden können, wurde beispielhaft noch die Möglichkeit implementiert, Daten in Graphvisualisierungen darzustellen. Für diesen Zweck wird die JavaScript-Bibliothek *plotly.js*[61] genutzt, die auch in verschiedenen professionellen Anwendungen verwendet wird. Um diese in Snap! einzubinden, wird sie über einen benutzerdefinierten JavaScript-Block automatisch im Hintergrund geladen und kann über drei dafür implementierte Blöcke genutzt werden (vgl. auch Abbildung 11.9): Der C-förmige *plot*-Block lädt die Bibliothek, initialisiert diese, bereitet die Darstellung der Graphen vor und stellt diese – nachdem die Blöcke innerhalb des „C“ bzw. der Klammer ausgeführt wurden – dar. Mit den inneren *plot*-Blöcken können mehrere verschiedene Graphen zugleich dargestellt und auch horizontale Linien, beispielsweise zur Kennzeichnung von Durchschnittswerten, eingefügt werden. Diese Implementierung ist dabei nur beispielhaft für die Einbindung weiterer JavaScript-Bibliotheken in Snap! zu sehen, wodurch die Funktionalität von Snap! und damit auch Snap!DSS flexibel erweitert werden kann.

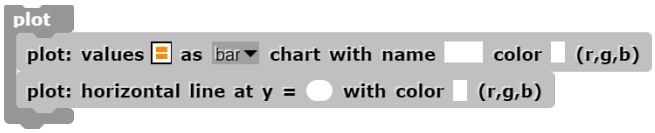


Abbildung 11.9: Blöcke zur Erstellung von Datenvisualisierungen unter Nutzung von *plotly.js*.

11.3.2 Einsatzmöglichkeiten im Informatikunterricht

Das entwickelte Werkzeug kann aufgrund der flexiblen Nutzbarkeit verschiedener Datenquellen im Unterricht in unterschiedlichen Kontexten eingesetzt werden. Durch die Möglichkeit auf Sensordaten zurückzugreifen, bietet sich beispielsweise die Nutzung in einem Physical-Computing-Projekt an, in dem eine eigene Wetterstation gebaut und programmiert wird. Diese Kombination von Physical-Computing mit der Analyse der Daten im Datenstromsystem spiegelt dabei ein wichtiges Funktionsprinzip moderner Techno-

[61] `https://plot.ly/javascript/`

logien wider, wie sie beispielsweise des im Internet der Dinge viel thematisiert werden. Im Rahmen eines solchen Projekts können Schülerinnen und Schüler die grundlegenden Funktionsprinzipien und Konzepte von Echtzeitdatenanalysen kennenlernen und eigene Analysen durchführen. Sie können dabei Daten verschiedener Datenquellen, z. B. mehrerer Sensoren oder APIs, miteinander kombinieren, um bessere Analyseergebnisse zu erhalten, und anhand dieser Ergebnisse einfache Vorhersagen erstellen. Durch Nutzung von APIs können in verschiedenen Fällen die eigenen Ergebnisse außerdem mit professionellen Ergebnissen abgeglichen und gleichzeitig ein Einblick in die unterschiedliche Funktionsweise verschiedener Datenquellen gewonnen werden.

Zur Generierung der Daten kann beispielsweise ein *Arduino-Uno*-Mikrocontroller-Board mit Temperatur-, Luftdruck-, Feuchtigkeits- und Lichtsensoren eingesetzt werden. Auf der Softwareseite kommt Snap4Arduino[62] mit der Erweiterung Snap!DSS zum Einsatz. Die Sensordaten werden als Datenstrom interpretiert und, nachdem sie dem Datenstromsystem hinzugefügt worden sind, ausgewertet. Hierbei bieten sich beispielsweise die kontinuierliche Ermittlung von Minimal-, Maximal- und Durchschnittswerten für verschiedene Zeiträume (beispielsweise die letzte Stunde) an – eine Aufgabe, die zwar einfach scheint, ohne Datenstromsystem aber eher ineffizient lösbar (Speicherung aller Werte) oder eher komplex (z. B. durch Nutzung eines Ringpuffers als Datenstruktur) wäre. In den Abbildungen 11.10 bis 11.12 wird als Beispiel die Erfassung der Daten eines Lichtsensors dargestellt.

Die Ergebnisse dieser Analysen können direkt in Snap! angezeigt, aber auch als Eingaben für weitere Analysen oder für jegliche andere Snap!-Blöcke genutzt werden. Damit ist es beispielsweise möglich, sowohl aktuelle Werte als auch die Ergebnisse der Analysen gemeinsam auf einem LCD-Display anzuzeigen. Da die Daten kontinuierlich im Hintergrund erhoben werden und diese somit völlig unabhängig von der Visualisierung sind, kann die Visualisierung jederzeit während der laufenden Analyse gewechselt werden, ohne die Analyse selbst zu beeinflussen.

Das Beispielprojekt zeigt damit einen einfach verständlichen und zu implementierenden Einstieg in die Welt der Datenstromsysteme, ohne dass dabei zentrale Konzepte vernachlässigt werden bzw. der Reduktion zum Opfer fallen müssen. Die Analysen werden gleichzeitig einfacher und nachvollziehbarer, als wenn beispielsweise Datenbanksysteme involviert sind, da in diesen Fällen oft eine Kommunikation zwischen verschiedenen Systemen nötig wird, während in diesem Beispiel die komplette Programmierung innerhalb desselben Softwarewerkzeugs stattfinden kann wie auch die Erfassung der Daten. Dieses Konzept lässt sich auch auf viele weitere Beispiele übertragen, beispielsweise den Bau eines Rauchmelders, der auf einer einfachen Stufe nur optisch funktioniert (wie die meisten üblichen Rauchmelder), zur Vermeidung von Fehlalarmen auf einer komplexeren Stufe des Projekts aber beispielsweise Temperaturdifferenzen miteinbeziehen kann.

[62] `http://www.snap4arduino.rocks`

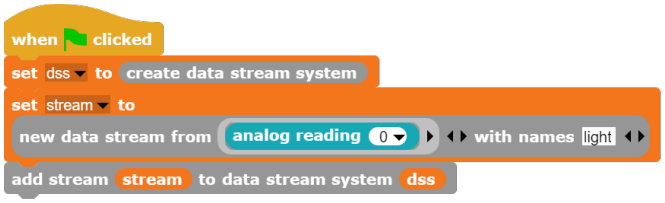


Abbildung 11.10: Blockcode zur Analyse eines Sensordatenstroms.

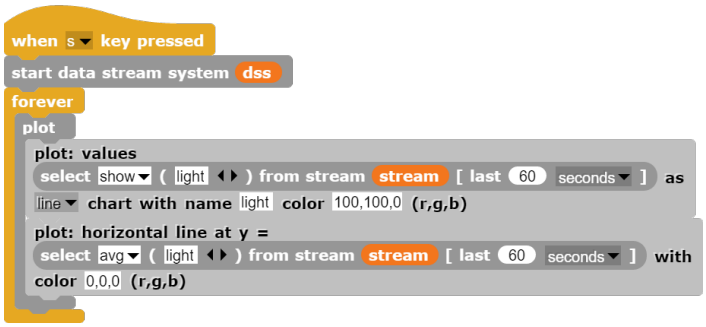


Abbildung 11.11: Blockcode zur Visualisierung von Sensordaten.

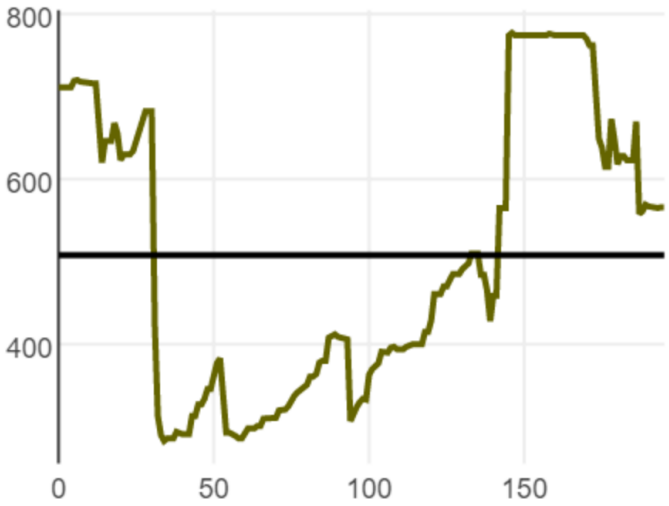


Abbildung 11.12: Ergebnis einer Auswertung des Datenstromsystems: Das Liniendiagramm repräsentiert die eigentlichen Werte, die der Sensor übermittelt hat. Die horizontale Linie kennzeichnet deren Durchschnittswert.

12 Erprobung und Evaluation einer Unterrichtssequenz zum Thema Data Mining

Während die zuvor vorgestellten Werkzeuge nur in einer 90-minütigen Unterrichtsstunde knapp evaluiert und/oder in Lehrkräftefortbildungsworkshops zur Diskussion gestellt wurden, fand zur tiefergehenden Evaluation der Möglichkeiten und Grenzen im Sekundarschulunterricht eine weitere unterrichtliche Erprobung statt. Statt die zuvor angesprochenen Themen weiter zu vertiefen, wurde in dieser Erprobung ein weiteres besonders spannendes und einflussreiches Thema des modernen Datenmanagements aufgegriffen, das *Data Mining* in Kombination mit der Vorhersage basierend auf Daten: Statt Daten für eine spezielle Analyse gezielt zu sammeln, wird heute versucht, aus auf Vorrat gesammelten Daten wertvolle Informationen zu gewinnen *(Dorschel, 2015)*. Trotz der hohen Bedeutung auch in verschiedenen Alltagskontexten, sind die Mächtigkeit und potenzielle Folgen solcher explorativen Datenanalysen für den Großteil der Bevölkerung kaum einschätzbar. Um jedoch Entscheidungen darüber treffen zu können, ob man konkrete Informatiksysteme und/oder Dienste nutzt, die Daten in großem Umfang sammeln und analysieren, ist es notwendig, das Potenzial und die Gefahren der Big-Data-Verarbeitung insbesondere auch am Beispiel des Data Mining einschätzen zu können und dabei gleichzeitig auch einen Einblick in die Charakteristika korrelationsbasierter Datenanalysen zu bekommen. Um diesen Anforderungen nachzugehen, wurde eine Unterrichtssequenz zum Thema „Data Mining – Schürfen nach dem Datengold“ konzipiert und in zwei Klassen erprobt, die es Schülerinnen und Schülern erlauben soll, erste Einblicke in die Gewinnung von Informationen aus Daten, die nicht für diesen Zweck erhoben wurden, zu erhalten.

Durch wissenschaftliche Begleitung der Erprobung dieser Unterrichtssequenz sollte insbesondere evaluiert werden, ob es möglich ist, dieses eher komplexe Thema adäquat im allgemeinbildenden Schulunterricht zu thematisieren. Die genaueren Fragestellungen der Untersuchung werden in Abschnitt 12.3.1 angegeben, nachdem zuvor im Folgenden die geplante Unterrichtseinheit und die dazu notwendigen didaktischen Vorüberlegungen beschrieben und begründet werden.

12.1 Didaktische Vorüberlegungen

Um den Schülerinnen und Schülern die Grundlagen des Data Mining nahezubringen, war es insbesondere notwendig, diesen sehr umfangreichen Bereich genauer zu spezifizieren bzw. einzugrenzen und gegebenenfalls an verschiedenen Stellen didaktische Reduktionen vorzunehmen. Da für den Unterrichtsversuch nur eine für ein solch umfangreiches Thema sehr knappe Zeit von drei Doppelstunden zur Verfügung stand, wurde der Fokus nicht auf einen tiefgehenden Einblick in verschiedene Analysemethoden gelegt, sondern stattdessen

zentrale Schritte der Gewinnung von neuen Erkenntnissen anhand von Daten mit den Schülerinnen und Schülern thematisiert, um ihnen einen Einblick in die Möglichkeiten zu gewähren und das Interesse an diesem Thema zu wecken. Dementsprechend wurde entschieden, zur Minimierung des zeitlichen Aufwands auf einem vorhandenen Datensatz aufzusetzen, statt selbstständig Daten zu erfassen. Die Schülerinnen und Schüler sollten jedoch die Möglichkeit bekommen, diesen selbstständig zu analysieren und basierend auf dem entwickelten Modell Vorhersagen zu treffen.

Bezug zum Data-Literacy-Kompetenzmodell und angestrebte Kompetenzen. Aus Sicht der Prozessbereiche des zuvor entwickelten Data-Literacy-Kompetenzmodells befasst sich die Unterrichtssequenz daher insbesondere mit dem *Analysieren, Visualisieren und Interpretieren (P3)* der Daten, wobei Visualisierungen der Ergebnisse nur soweit genutzt wurden, wie es für das Verständnis zwingend notwendig war, ohne jedoch die konkrete Aufbereitung der Ergebnisse näher zu thematisieren. Aus inhaltlicher Perspektive wurde hingegen angestrebt, eine breitere Abdeckung zu erreichen, indem mit Grundlagen aus dem Bereich *Daten und Information (C1)* aufgegriffen, Einblicke in die *Datenanalyse (C3)* gegeben und insbesondere Implikationen im Bereich *Datenethik und Datenschutz (C4)* diskutiert werden. Auf die explizite Thematisierung der Datenspeicherung wurde hingegen verzichtet, da diese für das Unterrichtsziel insbesondere daher verzichtbar war, als die angestrebte Zielgruppe bereits Erfahrungen mit Tabellenkalkulationssoftware aus dem Vorunterricht mitbringt und somit zumindest mit dieser Art der Darstellung der gespeicherten Daten umgehen kann. Aus diesen Inhalts- und Prozessbereichen konnten die im Folgenden dargestellten angestrebten Kompetenzen abgeleitet werden.

- Die Schülerinnen und Schüler erläutern, warum aus gespeicherten Daten verschiedene und ggf. neue Informationen gewonnen werden können (C1/P3).
- Die Schülerinnen und Schüler charakterisieren den Unterschied zwischen korrelations- und kausalitätsbasierten Zusammenhängen in Daten sowie deren jeweilige Aussagekraft (C1/P3, z. T. auch C4/P3)
- Die Schülerinnen und Schüler skizzieren den Ablauf einer (korrelationsbasierten und vorhersageorientierten) Datenanalyse (C3/P3).
- Die Schülerinnen und Schüler charakterisieren eine typische Analysemethode und erläutern das dieser zugrundeliegende Prinzip an einem geeigneten Beispiel (C3/P3).
- Die Schülerinnen und Schüler führen einfache Datenanalysen unter Nutzung einer üblichen Methode händisch sowie unter Nutzung eines geeigneten Softwarewerkzeugs aus (C3/P3).
- Die Schülerinnen und Schüler prognostizieren fehlende Attribute eines Datensatzes unter Rückgriff auf die von ihnen durchgeführte Datenanalyse (C3/P3).
- Die Schülerinnen und Schüler bewerten das Ergebnis ihrer Vorhersage und erläutern Ideen zu deren Verbesserung (C3/P3).

- Die Schülerinnen und Schüler reflektieren ihre Ergebnisse kritisch und diskutieren sie unter Berücksichtigung ethischer und gesellschaftlicher Gesichtspunkte (C4/P3).

Genutzte Methoden zur Datenanalyse, Vorhersage und Evaluation. Um diese angestrebten Kompetenzen erreichen zu können, ist es sinnvoll, nicht reine Anwendungsfertigkeiten in Form der Bedienung eines geeigneten Softwarewerkzeugs zu schulen, sondern den Schülerinnen und Schülern einen klaren Einblick in eine Datenanalysemethode zu geben und ein Verständnis des hinter dieser stehenden Prinzips zu ermöglichen. Damit ein möglichst einfacher Einstieg in die Datenanalyse auch ohne Vorkenntnisse möglich wird, bietet sich insbesondere die Thematisierung von *Klassifikationsbäumen* an: Diese Bäume sind informatisch betrachtet (nicht zwingend binäre) Entscheidungsbäume, die sich durch ein einfaches und klar verständliches Prinzip sowie die Möglichkeit, sie relativ übersichtlich grafisch darzustellen auszeichnen. Mit Ausnahme der Blätter entspricht dabei jeder Knoten des Baumes einer Entscheidung auf dem Weg hin zur Einordnung des analysierten Datensatzes in eine Klasse. Diese verschiedenen Möglichkeiten der Einordnung werden durch die Blätter des Baumes repräsentiert. Um einen Datensatz in einer Klasse einzuordnen – und basierend auf dieser Einordnung wiederum eine Vorhersage über unbekannte Attribute des Datensatzes zu treffen – wird der Baum von der Wurzel bis zum Blatt durchlaufen und an jedem inneren Knoten (einschließlich der Wurzel) anhand der vorliegenden Daten entschieden, ob dem linken oder rechten Ast gefolgt wird, bis ein Blatt – d. h. eine Klasse – erreicht wird. Anhand der gemeinsamen Eigenschaften aller dieser Klasse zugeordneten Datensätze können dann Vorhersagen getroffen werden, indem davon ausgegangen wird, dass wenn alle Datensätze einer Klasse eine bestimmte Eigenschaft erfüllen, und ein neuer Datensatz aufgrund seiner Eigenschaften in diese Klasse eingeordnet werden kann, er auch alle weiteren Eigenschaften der Klasse erfüllen müsste. Dieses Vorhersageprinzip ist selbst für jüngere Schülerinnen und Schüler leicht verständlich und vermittelt gleichzeitig – soweit nicht bereits vorhanden – ein grundsätzliches Verständnis über Bäume in der Informatik, die eine fundamentale Idee der Informatik *(Schwill, 1993)* darstellen.

Während dieser Teil des Analyseprozesses, die eigentliche Prognose, relativ einfach ist, stellt der Aufbau des Klassifikationsbaumes die schwierigere Aufgabe für die Lernenden dar: Hier müssen zuerst potenzielle Klassen ermittelt werden, was zwar relativ einfach wird, wenn genau ein unbekanntes Attribut mit bekannter und beschränkter Wertemenge existiert, aber auch beliebig komplex werden kann, wenn die konkreten Ausprägungen des analysierten Merkmals unbekannt sind. Um diese Aufgabe, die für große und komplexe Datensätze nur noch schwer überblickbar ist, nachvollziehen zu können, wird im geplanten Unterricht anfangs mit einem relativ kleinen und übersichtlichen Datensatz gearbeitet, zu dem ein entsprechender Klassifikationsbaum manuell erstellt und basierend auf dem entstehenden Klassifikationsbaum händisch eine Prognose getroffen werden kann. Für die Vertiefung der dabei erworbenen Kompetenzen und um einen Einblick in das hohe Potenzial, die einfache Durchführung und die Schnelligkeit einer solchen Datenanalyse zu bekommen, kann jedoch nicht bei diesem händischen Ansatz aufgehört werden. Stattdessen bietet es sich an, mit den Schülerinnen und Schülern auch einen umfangreicheren (aber

trotzdem noch relativ gut handhabbaren) Datensatz zu analysieren, bei dem die händische Analyse kaum mehr möglich ist. Spätestens bei dieser Analyse soll auch eine Evaluation der Analyseergebnisse stattfinden, wozu beispielsweise Konfusionsmatrizen, die die vorhergesagten Werte den Referenzwerten gegenüberstellen, gut geeignet sind, da diese auf übersichtliche Weise einen ersten Einblick in die Qualität der Datenanalyse bzw. der darauf basierenden Vorhersage geben.

Gewählter Datensatz. Zur Auswahl des im Unterricht genutzten Datensatzes wurden die in Kapitel 10 dargestellten Kriterien herangezogen. Insbesondere soll es sich bei den zu analysierenden Daten um reale Daten aus einem für die Schülerinnen und Schüler bedeutsamen Kontext handeln, um eine direkte Betroffenheit zu erreichen. Während beispielsweise in der öffentlichen Verwaltung oder in großen Unternehmen zwar große Datenmengen existieren, die zum Teil auch offen verfügbar sind, sind diese im Allgemeinen daher kaum für eine derart kurze Einführung in die Datenanalyse geeignet, sondern können eher bei längeren Unterrichtsequenzen ihr Potenzial offenbaren. Stattdessen wurde hier ein frei verfügbarer Datensatz ausgewählt, der den Schülerinnen und Schülern wesentlich näher liegt: Der im *UCI Machine Learning Repository* zur Verfügung stehende Datensatz *Student Performance Data Set*[63] basiert auf Realdaten zweier portugiesischer Schulen aus den Jahren 2005 und 2006 und beinhaltet nicht nur anonymisierte persönliche Daten von fast 650 Schülerinnen und Schülern, sondern auch deren Schulnoten in jeweils drei Leistungserfassungen (vgl. *Cortez und Silva (2008)*). Basierend auf diesen Daten konnten *Cortez und Silva (2008)* in einer wissenschaftlichen Untersuchung verschiedene Zusammenhänge dieser Attribute aufzeigen, die auch von Schülerinnen und Schülern entdeckt werden können. Aufgrund des kategorialen Charakters von Schulnoten eignen sich diese besonders gut als vorhergesagtes Attribut, sodass die zentrale Herausforderung im Rahmen der Unterrichtssequenz sein wird, aus den allgemeinen Daten, die über die portugiesischen Schülerinnen und Schüler bekannt sind sowie ihrer ersten beiden Noten die dritte „vorherzusagen“. Durch den klaren Bezug dieses Datensatzes zum Alltag der Lernenden kann dabei eine relativ hohe Betroffenheit erwartet werden, wahrscheinlich werden einige von der doch relativ guten Prognosequalität überrascht sein, sodass eine Diskussion über die ethische Vertretbarkeit einer derartigen Vorhersage im Unterricht angestoßen werden kann.

Für den Einstieg in die Unterrichtssequenz, bei dem eine händische Datenanalyse stattfinden soll, ist dieser Datensatz jedoch deutlich zu umfangreich – sowohl in der Anzahl der erfassten Attribute als auch der Anzahl der Datensätze. Entsprechend wird zu Anfang der Unterrichtssequenz ein weiterer Datensatz benötigt. Dazu wurde die Entscheidung getroffen, nicht eine kleine Teilmenge aus diesem Datensatz auszuwählen, sondern stattdessen einen zweiten Kontext zu wählen und hier einen fiktiven Datensatz zu nutzen, der sicherstellt, dass die Schülerinnen und Schüler schnell zu ersten Ergebnissen kommen: Einen fiktiven Auszug aus einer Bestelldatenbank eines Onlinehändlers, die in der Realität auch häufig Data-Mining-basierte Analyseverfahren einsetzen. Die Nutzung eines solchen künstlichen Datensatzes entspricht zwar nicht dem Kriterium der *Nutzung von Realdaten*,

[63] `https://archive.ics.uci.edu/ml/datasets/student+performance`

da dieser Datensatz jedoch nur zur Einführung und nicht für den Hauptteil des Unterrichts eingesetzt wird, und somit später die reale Nutzung verdeutlicht wird, wurde dieser Kompromiss bewusst eingegangen.

Gewähltes Softwarewerkzeug. Während der für den Unterrichtsanfang gewählte Datensatz händisch analysiert werden sollte und somit kein Softwarewerkzeug notwendig war, müsste für die Analyse des für den Hauptteil gewählten Datensatzes ein entsprechendes Werkzeug gefunden werden, das die Erstellung von Klassifikationsbäumen, die darauf aufbauende Datenanalyse und Vorhersage sowie zur Evaluation der Vorhersagequalität die Erstellung von Konfusionsmatrizen unterstützt. Da es sich bei allen diesen Methoden um klassische und bewährte Methoden der Datenanalyse handelt, werden diese von nahezu jedem üblichen Werkzeug unterstützt. Um den Schülerinnen und Schülern jedoch die Arbeit mit dem Werkzeug möglichst zu vereinfachen und möglichst direkt und ohne detaillierte Einführung in ein Programm oder gar eine Programmiersprache mit der Analyse starten zu können, fiel die Wahl auf das Analysewerkzeug *Orange 3*. Dieses für Nicht-Informatiker entwickelte Analysewerkzeug steht unter freier Lizenz bei der Universität Ljubljana[64] zur Verfügung. Es arbeitet mit einer grafischen Bedienoberfläche die den Datenfluss visualisiert, beinhaltet alle benötigten Funktionalitäten und kann ohne Vorkenntnisse in der Programmierung bedient werden (vgl. Abbildung 12.1). Im Vergleich zu anderen ähnlichen Werkzeugen (beispielsweise RapidMiner[65]) arbeitet es offline, ist portabel auf allen üblichen Betriebssystemen ausführbar und frei verfügbar, sodass eine flexible Nutzung ohne Einschränkungen im Schulunterricht möglich ist. Damit ermöglicht das Werkzeug sowohl für Lehrerinnen und Lehrer als auch Schülerinnen und Schüler einen einfachen Einstieg in die Unterrichtssequenz. Durch die offene Lizenz sind außerdem Anpassungen des Quellcodes möglich – für den Schulversuch wird dies genutzt, um nicht benötigte Funktionalitäten auszublenden, auf diese Weise das Werkzeug übersichtlicher zu machen und somit den Einstieg weiter zu vereinfachen.

12.2 Überblick über die Unterrichtssequenz

Die im Folgenden kurz vorgestellte Unterrichtssequenz wurde im Hinblick auf eine flexible Einsetzbarkeit in der Sekundarstufe I/II geplant. Dazu wurde soweit möglich darauf verzichtet, auf potenzielle Vorkenntnisse der Schülerinnen und Schüler aufzubauen. Stattdessen wurde die Sequenz so gestaltet, dass sie nahezu unabhängig vom vorhergehenden und nachfolgenden Unterricht einsetzbar ist, solange eine sinnvolle Verknüpfung herstellbar ist. Entsprechend bietet es sich beispielsweise an, diese Sequenz zum Einstieg in das Thema Datenbanken zu nutzen, d. h. um diese Systeme zu motivieren, oder auch als Ausblick am Ende dieses Themas. Die Unterrichtssequenz wurde auf eine Dauer von vier Doppelstunden ausgelegt, wobei drei Doppelstunden als Minimum angesehen werden, wenn entweder

[64] `https://orange.biolab.si`
[65] `https://rapidminer.com`

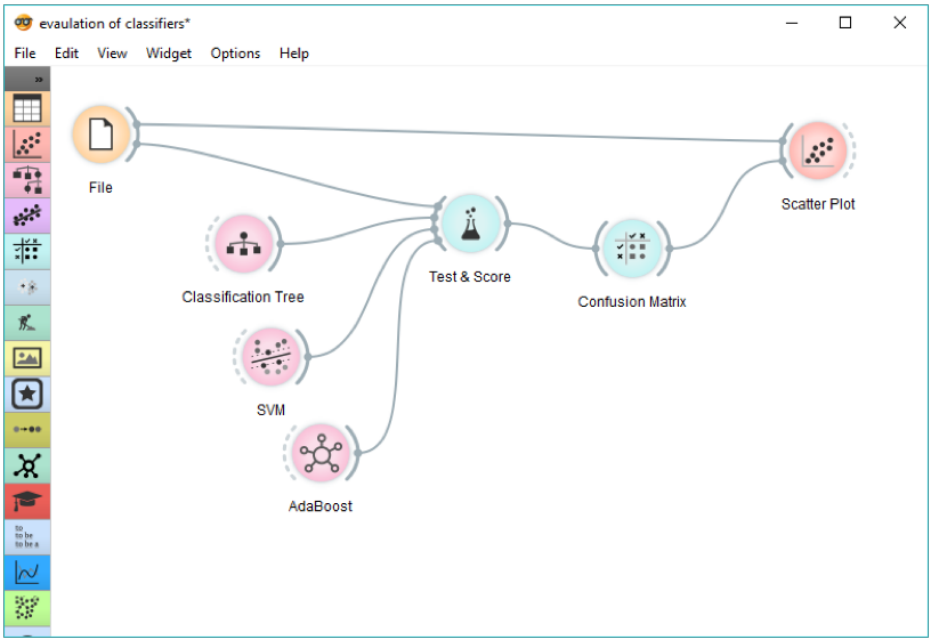


Abbildung 12.1: Das im Unterricht verwendete Datenanalysewerkzeug Orange 3.

die erste (Doppelstunden 1–2) oder letzte Phase (Doppelstunde 4) stark abgekürzt wird. Eine Maximaldauer kann hingegen kaum angegeben werden, da an vielen Stellen Erweiterungen möglich sind, beispielsweise indem weitere Analysemethoden thematisiert und einander gegenübergestellt oder Optimierungen diskutiert und erprobt werden. Die Unterrichtssequenz wird im Folgenden kurz skizziert, indem ein grober Überblick über die vier geplanten Doppelstunden gegeben wird, eine detaillierte Lehrkräftehandreichung mit den zugehörigen Arbeitsblättern kann Anhang F entnommen werden.

Doppelstunde 1: Motivation und Einführung grundlegender Begriffe. In der ersten Doppelstunde wird durch Präsentation eines Presseberichtes über eine Datenanalyse des US-Einzelhändlers Target das Thema motiviert. Anhand dieses Artikels und weiterer bekannter Analysen wird ein erster Einblick in die Datenanalyse im Einzelhandel gewonnen und darüber diskutiert. Der Wert verschiedener Daten für die Händler wird angerissen. Auf dieser Basis werden die Begriffe Kausalzusammenhang, Korrelation und Prognose eingeführt. Zum Abschluss wird der grundsätzliche Ablauf einer auf Prognosen ausgerichteten korrelationsbasierten Datenanalyse eingeführt und Ideen für potenzielle Zusammenhänge in Daten eines Versandhandels gesammelt.

Doppelstunde 2: Händische Datenanalyse und Einführung der Methodik. Die zweite Doppelstunde beschäftigt sich mit der manuellen Datenanalyse und der Einführung in die Methode der Klassifikation unter Nutzung von Klassifikationsbäumen. Dazu werden anhand konkreter Daten händisch Regeln für Schlussfolgerungen abgeleitet, die unter

Nutzung von Klassifikationsbäumen visualisiert und strukturiert werden, um daraufhin anhand dieser rein korrelativen Regeln Prognosen zu treffen. Durch nicht immer vorhandene Eindeutigkeit der Prognose bzw. auch der Prognoseregeln wird eine Diskussion über die Analysequalität angestoßen. Als Abschluss der Stunde findet eine Einführung in die Aufgabe der nächsten Doppelstunde statt, indem bereits der zu analysierende Datensatz und die Aufgabe vorgestellt werden und erste Vermutungen der Schülerinnen und Schüler über die Prognoseergebnisse und -qualität gesammelt werden.

Doppelstunde 3: Automatisierung der Datenanalyse am PC. Auf Basis der vorherigen Doppelstunden wird in der dritten Doppelstunde die Analyse des größeren und händisch kaum mehr analysierbaren Schülerdatensatzes am PC durchgeführt. Die Einführung in das Analysewerkzeug findet dabei größtenteils im Rahmen eines Arbeitsblattes statt, bei dem die Schülerinnen und Schüler aufgrund der begrenzten Zeit relativ stark geleitet werden. Durch offene Aufgabenstellungen wird ihnen aber gleichzeitig der nötige kreative Freiraum gegeben, um Spaß an der Thematik zu entwickeln. Die Analyse wird daraufhin hinsichtlich ihrer Qualität und auch hinsichtlich ethischer Einflüsse diskutiert, was dadurch erleichtert wird, dass, wie oben beschrieben, ein Kontext gewählt wurde, der die Schülerinnen und Schüler direkt betrifft.

Doppelstunde 4: Kritische Reflexion der Thematik. Für die vierte Doppelstunde ist eine weitere Vertiefung der Diskussion der rechtlichen, ethischen und moralischen Aspekte solcher Datenanalysen im Rahmen eines Gruppenpuzzles vorgesehen, in dem die Schülerinnen und Schüler verschiedene weitere Datenanalysen kennenlernen, deren Funktionsweise und Auswirkungen diskutieren und für sich selbst mithilfe des erworbenen Wissens bewerten.

12.3 Erprobung

Die Erprobung erfolgte in zwei Gruppen mit 15 bzw. 12 Schülerinnen und Schülern (davon drei bzw. null weiblich) der neunten Klasse einer bayerischen Realschule im Rahmen des regulären IT-Unterrichts. Es standen dafür drei Doppelstunden zur Verfügung, sodass die zuvor skizzierte vierte Doppelstunde in der Erprobung entsprechend entfiel bzw. auf eine kurze Diskussion gekürzt wurde. Der gesamte Unterricht wurde von der regulären Lehrkraft durchgeführt und vom Autor dieser Arbeit, wie in Abschnitt 12.3.2 detaillierter beschrieben, beobachtet.

12.3.1 Ziele der Untersuchung

Im Rahmen der Erprobung sollte insbesondere ein evaluiert werden, ob und in welchem Maß die angestrebten Themen im Unterricht umsetzbar sind und wie stark das Interesse

der Lernenden am Thema ist, um so das Potenzial für den Unterricht abschätzen zu können. Entsprechend wurden folgende Fragen in den Mittelpunkt gestellt:

Q1 Welche Schwierigkeiten treten bei der Umsetzung des Unterrichtskonzepts sowohl aus Lehrer- als auch Schülersicht auf?

Q2 Wie gut eignen sich die Zugänge und Beispiele, um die Schülerinnen und Schüler zu motivieren bzw. deren Interesse zu wecken?

Q3 Wie starkes Interesse zeigen die Schülerinnen und Schüler am Thema Datenanalyse und Vorhersage?

12.3.2 Untersuchungsmethoden: Beobachtung, Interview und Fragebogenstudie

Um diese drei Fragen zu beantworten, wurde eine qualitative Herangehensweise gewählt, da diese Fragen eher explorativer Natur sind. Eine quantitative Untersuchung wäre in diesem Fall eher hinderlich: Beispielsweise würde dabei die Notwendigkeit bestehen, vorher zur ersten Frage bereits Kategorien von Schwierigkeiten festzulegen und zu untersuchen, inwiefern diese auftreten. Dies würde jedoch die Exploration stark einschränken, da unerwartete Schwierigkeiten auf diese Weise möglicherweise unerkannt bleiben. Die qualitative Herangehensweise zeigt daher zur Beantwortung dieser Forschungsfragen ein höheres Potenzial und wurde in anderen Arbeiten zur Evaluation von Unterrichtserprobungen (beispielsweise von *Freischlad (2009)* und *Kohl (2009)*) bereits erfolgreich eingesetzt.

Um die verschiedenen Perspektiven, die im Unterricht eine Rolle spielen, zu erfassen, werden verschiedene Untersuchungsmethoden eingesetzt:

- Eine Unterrichtsbeobachtung durch den Forscher liefert eine Außensicht auf den Unterricht und ist geeignet, um beispielsweise auftretende Probleme neutral zu erkennen, den Erkenntnisgewinn der Schülerinnen und Schüler im Unterrichtsgespräch nachzuverfolgen und deren nach außen gezeigtes Interesse am Unterrichtsgegenstand nachzuvollziehen.
- Eine Schülerbefragung mit Fragebögen ermöglicht Aussagen über das von den Schülerinnen und Schülern geäußerte Interesse am Thema und durch Einbezug offener Fragen auch über deren Wahrnehmung des Themas und dessen Bedeutung. Gleichzeitig kann auch ein Eindruck des durch sie erworbenen Wissens gewonnen werden.
- Durch ein Interview mit der unterrichtenden Lehrkraft kann zusätzlich auch deren Eigenperspektive auf das Unterrichtsgeschehen erfasst und beispielsweise Schwierigkeiten, das Interesse der Schülerinnen und Schüler in Vergleich mit anderen Unterrichtsthemen und der Lernfortschritt beurteilt werden.

Die konkrete Ausgestaltung dieser drei Untersuchungsmethoden wird im Folgenden detaillierter beschrieben.

Unterrichtsbeobachtung. Die Unterrichtsbeobachtung wurde genutzt, um einen externalen Eindruck des Unterrichtsgeschehens zu gewinnen, der nicht aus Schüler- oder Lehrersicht geprägt ist. Diese Beobachtung erfolgte zwangsweise an der Unterrichtssituation teilnehmend, da keine Videoaufnahmen oder Ähnliches möglich waren, jedoch soweit möglich ohne aktive Einbindung des Beobachters in das Unterrichtsgeschehen[66]. Die Beobachtung wurde auf qualitative Weise durchgeführt, um einen möglichst guten Eindruck der gesamten Umsetzung der Unterrichtssequenz zu bekommen: *„Ziel der qualitativen Beobachtung ist es, soziales Geschehen sinnverstehend und möglichst ganzheitlich zu erfassen." (Döring und Bortz, 2016)* Diese Art der Beobachtung ist zwar durch verschiedene potenzielle Fehlerquellen und subjektive Beeinflussung geprägt, wird aber von Experten trotzdem als ausreichend valide angesehen und in verschiedenen Aspekten – insbesondere hinsichtlich der Authentizität der gewonnenen Daten *(vgl. Atteslander, 2010)* – als besonders vorteilhaft erachtet. Um jedoch eine gewisse Systematisierung der Beobachtung vorzunehmen, wurde – in Anlehnung an quantitative Beobachtungsmethoden – ein Beobachtungsbogen erstellt. Dieser legt die regelmäßige Erfassung (mindestens alle 5 Minuten, je nach Unterrichtsphase häufiger) einiger grundsätzlicher Aspekte (Unterrichtsphase, Methode) nahe, gleichzeitig wurde auch versucht, immer festzuhalten, ob die Kontrolle bei Schülerinnen bzw. Schülern oder der Lehrkraft liegt und die Stimmung der Schülerinnen und Schüler zu erfassen. Durch die zusätzliche Möglichkeit weitere Anmerkungen festzuhalten, die umfangreich genutzt wurde, wirkte dieser Beobachtungsbogen aber nur als Richtlinie und schränkte die Beobachtung nicht auf diese Aspekte ein. Insbesondere wurde auch versucht, die Einstellungen der Lernenden zum Thema sowie deren Vorwissen zu erfassen, das sich im Unterricht zum Teil auch durch ihre geäußerten Vorstellungen klar zeigte. Auf diese Weise kann, ergänzend zu der bereits zuvor in einer anderen Schülergruppe ermittelten Schülerperspektive (vgl. Abschnitt 6.2), ein vertiefter Einblick gewonnen werden. Der verwendete Beobachtungsbogen ist in Anhang G angehängt. Die Unterrichtsbeobachtung ist damit eine zentrale Methode für die Beantwortung aller drei Forschungsfragen.

Schülerfragebögen. Die Schülerfragebögen wurden hingegen eingesetzt, um die Perspektive der Schülerinnen und Schüler tiefergehend zu erfassen. Dabei wurde darauf verzichtet, in einem Vorabfragebogen deren Wissen über Begriffe aus dem Themenbereich oder anderes Vorwissen zu erfassen (ähnlich wie in Abschnitt 6.2), da dadurch gleichzeitig auch die nachfolgende Beobachtung mindestens zum Teil beeinflusst worden wäre: Durch Nennung von Begriffen aus dem Fachgebiet sind diese den Schülerinnen und Schülern bereits aus dem Fragebogen bekannt, sodass im Unterricht deren Vorwissen schwerer einschätzbar wäre und insbesondere bei einer möglicherweise auftretenden Verwendung von Fachbegriffen durch die Schülerinnen und Schüler nicht mehr klar wäre, ob sie diesen Begriff eher aus dem Alltag kennengelernt haben oder nur (ggf. unterbewusst) versuchen, diesen aus dem Fragebogen zu übernehmen. Entsprechend wurde auch nicht versucht, den Wissenszuwachs der Schülerinnen und Schüler zu erfassen, da dazu ein Pre-Test nötig

[66] Der Beobachter/Forscher war nicht explizit in das Unterrichtsgeschehen eingebunden. Bei der Verwendung des Softwarewerkzeugs stand er jedoch insbesondere bei technischen Fragen, die nicht durch die Lehrkraft beantwortet werden konnten, als sekundärer Ansprechpartner zur Verfügung.

wäre, diese Untersuchung jedoch gleichzeitig aufgrund der geringen Teilnehmerzahl und ohne Vergleichsgruppen nur wenig aussagekräftig wäre. Stattdessen wird an dieser Stelle der Fokus auf das Interesse, das die Schülerinnen und Schüler dem Thema Datenanalyse im Vergleich zu anderen Themen des IT-Unterrichts an Realschulen entgegenbringen gesetzt, und darauf, ob die Aufgaben für diese umsetzbar/bearbeitbar waren und Ihnen Spaß bereitet haben – d. h. auf subjektive Aspekte, die auch dazu beitragen können, die Beobachtungen im Unterricht zu validieren. Diese beiden Aspekte werden im Schülerfragebogen in Form einer Likert-skalierten Frage zum Interesse an verschiedenen Themen des Informatikunterrichts (Skala: *sehr interessant, etwas interessant, kaum interessant, gar nicht interessant, ist mir unbekannt*) sowie mehreren Aussagen, zu denen die Schülerinnen und Schüler ihre Zustimmung bzw. Ablehnung erklären sollen, erhoben. Zusätzlich wurden die Teilnehmenden aufgefordert, zu benennen, was sie im Unterricht gelernt/bemerkt/entdeckt haben, um einen Eindruck davon zu bekommen, welche die für sie relevantesten Aspekte des Unterrichts waren. Um einen gewissen Einblick in das von den Schülerinnen und Schülern erworbene Wissen zu bekommen, wurden Sie außerdem aufgefordert, die Bedeutung des für den Unterricht zentralen Begriffs „Klassifikation“ und die Schritte des Analyseprozesses bis hin zur Vorhersage zu erläutern und Aspekte zu nennen, mit denen die Qualität von Vorhersagen verbessert werden kann. Zuletzt wurden die Teilnehmenden mit einer für sie neuen und bisher im Unterricht nicht diskutierten Situation, der Datenanalyse im Gesundheitswesen, konfrontiert, zu der sie sich Gedanken um die möglicherweise vorhandenen Daten und jeweiligen Quellen, die Probleme einer Untersuchung von Patienten durch Ärzte primär anhand von Daten und um die Fehlerfreiheit von Datenanalysen Gedanken machen und diese zu Papier bringen sollten. Somit trägt diese Teiluntersuchung insbesondere zur zweiten und dritten Forschungsfrage bei. Der dafür genutzte Fragebogen ist in Anhang H angehängt.

Lehrerinterview. Um als dritten Aspekt auch die Lehrerperspektive auf das Thema detaillierter zu erfassen, wurde diese in einem Interview befragt. Dieses war zwar sehr offen gestaltet, es wurde jedoch Wert auf die Beantwortung folgender Leitfragen gelegt:

- Wie war der Gesamteindruck von der Unterrichtssequenz?
- Wie gut hat der Unterricht funktioniert?
- Was haben die Schülerinnen und Schüler gelernt/mitgenommen?
- Was war am allgemeinen Aufbau der Unterrichtssequenz sinnvoll oder überarbeitungswürdig?
- Wie starkes Interesse haben die Schülerinnen und Schüler am Thema gezeigt?
- Welche Probleme wurden bei der Durchführung aus Perspektive der Lehrkraft erkannt?
- Fazit: Warum ist die geplante Unterrichtssequenz und das Thema allgemein für den Informatikunterricht wichtig und geeignet?

Entsprechend kann anhand dieses Interviews ein Beitrag insbesondere zu den ersten beiden Forschungsfragen geleistet werden, aber in Teilen auch zur dritten.

12.3.3 Durchführung und Auswertung

Die Durchführung der drei geplanten Teiluntersuchungen konnte wie geplant erfolgen. Die Ergebnisse dieser drei Untersuchungen werden im Folgenden dargestellt.

Außenperspektive: Unterrichtsbeobachtung

Während des gesamten Unterrichts konnte in beiden Gruppen eine kontinuierliche Beobachtung durch den Verfasser dieser Arbeit gewährleistet werden. Dabei wurde der vorher entworfene Beobachtungsbogen verwendet, um das Unterrichtsgeschehen möglichst umfassend zu erfassen, was dem subjektiven Eindruck nach gut funktionierte. Es konnten verschiedene Eindrücke vom Umgang der Schülerinnen und Schüler mit dem Thema und dem genutzten Werkzeug, aber auch von deren Interessen und Vorwissen gewonnen werden:

Allgemeine Eindrücke. Der Unterricht verlief in den beiden Klassen trotz derselben Lehrkraft, Jahrgangsstufe und desselben Vorwissens stark unterschiedlich:

- In der ersten Gruppe wurde der Unterricht stark durch einige unaufmerksame und augenscheinlich wenig motivierte Schüler beeinflusst, die sich häufig fremdbeschäftigten und durch Unterrichtsstörungen den Unterrichtsverlauf deutlich ausbremsten. Dies war von Anfang an und alle drei Doppelstunden hindurch klar erkennbar, wobei sich die negativen Auswirkungen auf den Unterricht im Verlauf der Unterrichtssequenz weiter verstärkten.
- In der zweiten Gruppe wurde der Unterricht hingegen eher dadurch geprägt, dass diese Schüler im Allgemeinen sehr vielseitig interessiert wirkten und sehr diskussionsfreudig waren. Dadurch bestand regelmäßig die Gefahr, vom eigentlichen Unterrichtsziel abzugleiten, sodass auch hier die Lehrkraft eine stark regelnde Funktion einnehmen musste, diese aber völlig anders ausgeprägt war.

Allgemein war in beiden Klassen feststellbar, dass nur wenige Schülerinnen bzw. Schüler sich mit dem Unterrichtsthema auch außerhalb des Unterrichts auseinandersetzten: Insbesondere bei den als Lehrer-Schüler-Gespräch durchgeführten Wiederholungen am Anfang der zweiten und dritten Doppelstunde war erkennbar, dass die Lernenden zwar grundsätzlich wissen, was das Thema der letzten Stunde und deren Ziel war und dass sie in eigenen Worten das Vorgehen beispielsweise bei der Datenanalyse beschreiben können, gleichzeitig konnten sie jedoch eher auf Faktenwissen ausgerichtete Fragen kaum ausreichend beantworten. Möglicherweise kann dies jedoch nicht nur durch mangelndes Interesse am Thema

– dies schien im Unterricht nicht der Fall zu sein – sondern auch dadurch erklärt werden, dass die Unterrichtserprobung schon relativ nahe am Schuljahresende stattfand, sodass die Aufmerksamkeit der Lernenden gegebenenfalls auf andere Fächer gerichtet war oder ihre Motivation im Allgemeinen bereits nachließ.

Interesse am Thema und Bezug zur Alltagswelt. Beide Gruppen schienen aus Sicht des Beobachters im Allgemeinen eher interessiert an der Thematik Datenanalyse und -vorhersage. Das zeigte sich insbesondere durch umfangreiche Diskussionen und dadurch, dass viele Schülerinnen und Schüler aus beiden Klassen eigene Beispiele aus dem Alltag eingebracht haben. Aufgrund der stärker ausgeprägten Diskussionskultur in der zweiten Klasse wurden diese dort jedoch deutlich intensiver diskutiert. Gerade in der zweiten Gruppe, aber in Ansätzen auch in der ersten, zeigten sich Versuche der Schülerinnen und Schüler, sich Datenanalysen und darauf basierende Vorhersagen intuitiv zu erklären, wobei zum Teil relativ gute Vorstellungen erkennbar waren, sodass vermutet werden kann, dass dieses Thema durchaus eine Rolle im Alltag der Lernenden spielt. Auch verschiedenste Daten und Datenquellen konnten durch die Schülerinnen und Schüler benannt werden, einige von ihnen stellten Bezug zu im Alltag viel diskutierten Themen, wie der Weitergabe von Facebookdaten an Cambridge Analytica, her.

Vorkenntnisse und Erfahrungen. Die Schülerinnen und Schüler konnten in beiden Klassen den Begriff „Big Data“ auf Nachfrage nicht erklären, auf die Frage ob dieser bekannt ist, meldete sich niemand. Trotzdem konnte in der einführenden Diskussion erkannt werden, dass die Schülerinnen und Schüler zum Großteil bereits Wissen über die unter diesem Begriff zusammengefassten Analysen mitbringen. In beiden Klassen brachten die Lernenden bereits erste Vorstellungen über die Funktionsweise von Datenanalysen und -vorhersagen mit:

- Die Notwendigkeit, Daten erst einmal zu speichern bevor sie analysiert werden können, war für nahezu alle Schülerinnen und Schüler offensichtlich.
- Das Erstellen von Regeln als Grundlage für die Vorhersage, die die Lernenden oft als Muster bezeichneten, schien für die meisten Lernenden intuitiv zu sein.
- Dadurch, dass die meisten Schülerinnen und Schüler regelmäßig versuchten, entdeckte Korrelationen durch Kausalzusammenhänge zu erklären, war erkennbar, dass mindestens implizit die Vorstellung vorherrscht, dass Vorhersagen immer kausal erklärbar sein müssen.
- In einer Diskussion, in der ein Schüler die Vorstellung äußerte, dass beispielsweise Onlinehändler ohne Probleme wissen, nach welchen Suchbegriffen Personen auch auf anderen Webseiten wie beispielsweise Suchmaschinen gesucht haben, zeigte sich, dass diese Vorstellung der Allwissenheit von Unternehmen bei einem nicht näher bezifferbaren Teil der Lernenden vorherrschte, während ein anderer Teil versuchte,

diese Vorstellung durch Argumente wie Datenschutz und technische Einschränkungen auszuräumen.

Insgesamt zeigte sich somit klar, dass die Schülerinnen und Schüler durch den hohen Alltagsbezug des Themas bereits verschieden ausgeprägtes Vorwissen und Vorstellungen über Datenanalysen mitbringen, sodass eine Notwendigkeit besteht, diese im Unterricht aufzugreifen und in Richtung eines fachlich korrekten Verständnisses hin auszubauen.

Verständnis der zentralen Inhalte und Ideen. Während im Allgemeinen nur vereinzelte Schwierigkeiten feststellbar waren, die sich schnell auflösten, zeigte sich in beiden Klassen insbesondere, dass die Thematisierung des Unterschieds zwischen korrelativen und kausalen Zusammenhängen sowie die jeweilige Bedeutung stärker in den Vordergrund rücken müsste, da die Nutzung korrelativer Zusammenhänge für die Schülerinnen und Schüler – obwohl sie eigene Beispiele nannten, in denen solche eine wichtige Rolle spielen – unintuitiv war und das Verständnis zu beeinträchtigen schien. Bis auf diesen Aspekt schien die Unterrichtssequenz zu einem Verständnis der angestrebten Inhalte und Ideen klar beizutragen: Insbesondere griffen die Schülerinnen und Schüler verschiedene Aspekte auch in eigenen Worten selbstständig auf und bauten sie in ihre Argumentationen ein. Es wurde realisiert, dass verschiedenste und auch unerwartete Daten zu einer Vorhersage beitragen können und somit von außen kaum erkennbar ist, welche Daten für eine Analyse potenziell wichtig sein könnten. Auch die zunehmende Vorhersagequalität bei Vergrößerung einer Datenmenge war für die Schülerinnen und Schüler logisch. Die genutzten Analysemethoden, insbesondere die Erkennung von Mustern und Abbildung in Klassifikationsbäumen schien für die Schülerinnen und Schüler die Analyse gut nachvollziehbar zu machen und zeigte bei der Analyse am PC ihr Potenzial: Einige der Lernenden waren sehr überrascht, wie schnell ein solcher Baum automatisch erstellt und auf gefühlt große Datenmengen angewendet werden kann.

Aufbau und Werkzeugauswahl. Obwohl aus Sicht des Beobachters gute Ergebnisse erzielt wurden, war der Aufbau der Unterrichtssequenz an manchen Stellen nicht ideal, insbesondere benötigt die Unterscheidung zwischen Korrelation und Kausalität mehr Zeit und Beispiele. Ansonsten hat sich der Aufbau jedoch als sinnvoll erwiesen, insbesondere die Trennung des Unterrichts in händische Datenanalyse und deren Automatisierung schien hilfreich. In beiden Gruppen war jedoch zu beobachten, dass die Motivation stark zugenommen hat, als von händischer auf automatisierte Datenanalyse umgestiegen wurde. Entsprechend wäre eine weitere mögliche Verbesserung, diese beiden Phasen stärker zu verflechten und gegebenenfalls häufigere Wechsel zwischen diesen stattfinden zu lassen. Eine solche Verflechtung scheint weiterhin auch sinnvoll, da der Umgang der Schülerinnen und Schüler mit dem gewählten Werkzeug weniger schwierig erschien als vermutet, im Gegenteil schien das Werkzeug sehr intuitiv, sodass nur geringe Einarbeitungszeiten notwendig waren. Dies zeigte sich insbesondere beim Erstkontakt mit dem Werkzeug: Als es an den praktischen Teil ging und die Schülerinnen und Schüler begannen ihre Computer

vorzubereiten, sank die Aufmerksamkeit drastisch, sodass beim Lesen des Arbeitsblattes übersehen wurde, dass eine Projektvorlage zur Verfügung stand, die den Einstieg erleichtern sollte. Statt diese Vorlage, auf die auch durch die Lehrkraft explizit hingewiesen wurde, zu nutzen, begannen alle Teilnehmenden direkt, die als Grafik auf dem Arbeitsblatt abgebildete Projektvorlage selbst nachzubauen – ohne dass eine Einführung in das Werkzeug und dessen Bedienung gegeben wurde. Bedienungsprobleme waren dabei nur vereinzelt feststellbar, da manche Fenster sich anders verhalten haben als erwartet (insbesondere mussten Änderungen nicht bestätigt werden, sondern wurden direkt wirksam). Der aus Zeitgründen notwendige Verzicht auf das geplante Gruppenpuzzle hat sich als ungünstig erwiesen: Im Gruppenpuzzle wäre das Thema durch weitere Beispiele abgerundet und die Übertragbarkeit des Wissens demonstriert worden. Ein Verzicht darauf sollte daher vermieden bzw. dieses adäquat ersetzt werden. Alles in allem hat sich das grundlegende Konzept der Unterrichtssequenz und deren Aufbau jedoch aus Sicht des Beobachters bewährt.

Schülerperspektive: Fragebögen

Die Bearbeitung des Fragebogens durch die Schülerinnen und Schüler wurde am Ende der dritten Doppelstunde angesetzt, wobei ihnen mit knapp 20 Minuten ein ausreichender Zeitumfang zur Verfügung stand, sodass alle Teilnehmenden den kompletten Fragebogen bearbeiten konnten. In der letzten Doppelstunde waren 14 bzw. 12 Schülerinnen und Schüler anwesend und standen somit für den Fragebogen zur Verfügung, der eine Rücklaufquote von 100% erzielte. Da kein Fragebogen aus offensichtlichen Gründen auszusortieren war, kann damit auf 26 Fragebögen zurückgegriffen werden. Die Bearbeitung dieser Fragebögen zeigte im Allgemeinen jedoch ein sehr differentes Bild: Während manche Schülerinnen bzw. Schüler sich intensiv mit den Fragestellungen auseinandergesetzt haben, zeigt sich insbesondere bei den Wissens- und Transferfragen, dass viele den Fragebogen nur sehr knapp und oberflächlich beantwortet haben. Dies bestätigt auch die Beobachtung während des Ausfüllens: In beiden Gruppen waren einige bereits nach wenigen Minuten mit der Bearbeitung fertig, während andere die ca. dreifache Zeit benötigten. Entsprechend sind die im Folgenden dargestellten Ergebnisse zum Teil wenig belastbar und weisen oft eine hohe Streuung auf, insbesondere bei der ersten Frage. Dennoch können daraus Indizien für die Weiterentwicklung der Unterrichtssequenz gewonnen werden, die in Abschnitt 12.3.4 zusammen mit der Beobachtung und dem Lehrerinterview ausgewertet werden.

Interesse an verschiedenen Unterrichtsthemen. Die Fragen nach dem Interesse an verschiedenen (potenziellen) Unterrichtsthemen wurden von den einzelnen Schülerinnen und Schülern sehr unterschiedlich beantwortet (vgl. Tabelle 12.1). Während einige von ihnen nahezu nur *kaum interessant* oder *gar nicht interessant* angekreuzt haben, haben andere sich eher auf den positiven Teil der Skala konzentriert. Insbesondere gibt es auch zwischen den beiden Gruppen klare Abweichungen: Die erste, im Unterricht schwierigere, Gruppe zeigte allgemein weniger Interesse an den zur Auswahl stehenden Themen als die zweite Gruppe. Dieser Unterschied zeigt sich deutlich in Abbildung 12.2.

Thema	Gruppe 1				Gruppe 2				Gesamt			
	Mittel-wert	Median	Modus	$\bar{d}_{0.5}$	Mittel-wert	Median	Modus	$\bar{d}_{0.5}$	Mittel-wert	Median	Modus	$\bar{d}_{0.5}$
Programmierung	1,9	2,0	2,0	0,6	2,5	3,0	3,0	0,5	2,2	2,0	2,0	0,6
Entwickeln von Computerspielen	1,9	2,0	3,0	1,0	2,8	3,0	3,0	0,2	2,3	3,0	3,0	0,7
Technisches Zeichnen / CAD	1,5	1,0	1,0	0,8	1,8	2,0	3,0	0,9	1,6	2,0	1,0	0,9
Datenspeicherung	0,8	1,0	0,0	0,7	1,3	1,0	2,0	0,6	1,0	1,0	1,0	0,7
Datenanalyse	0,7	0,0	0,0	0,7	1,4	1,5	2,0	0,6	1,0	1,0	2,0	0,7
Künstliche Intelligenz	2,4	3,0	3,0	0,6	2,4	3,0	3,0	0,6	2,4	3,0	3,0	0,6
Computergrafik	1,7	1,5	1,0	0,8	2,4	3,0	3,0	0,6	2,0	2,0	3,0	0,8
Entwickeln von Smartphone-Apps	1,9	2,0	2,0	0,6	2,3	2,5	3,0	0,8	2,1	2,0	2,0	0,7
Office-Programme	1,2	1,0	1,0	0,5	1,5	2,0	2,0	0,6	1,3	1,0	1,0	0,6

Tabelle 12.1: Interesse der befragten Schülerinnen und Schüler an verschiedenen Themenbereichen. Skala: 0 ≙ *gar nicht interessant* bis 3 ≙ *sehr interessant*

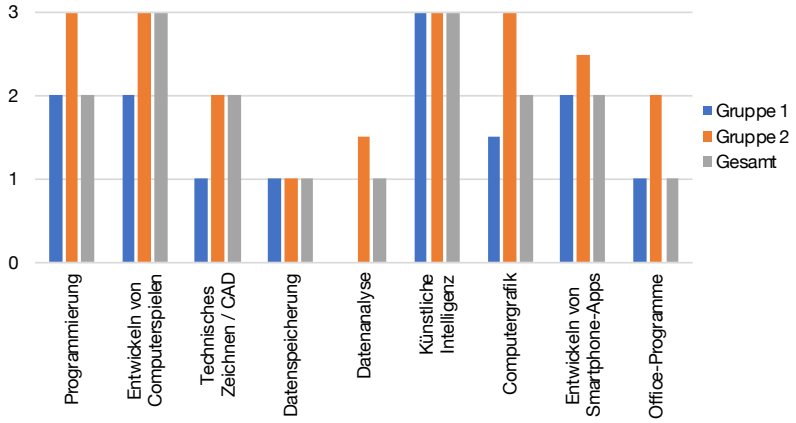


Abbildung 12.2: Median des Interesses der befragten Schülerinnen und Schüler an verschiedenen Themenbereichen. Skala: 0 ≙ *gar nicht interessant* bis 3 ≙ *sehr interessant*

Die Ergebnisse zeigen, dass das Thema *Datenanalyse* im Verhältnis zu den meisten anderen Themen als eher wenig interessant eingeschätzt wird, was der Einschätzung während des Unterrichts klar widerspricht. Da die Streuung der Werte mit einer mittleren Abweichung vom Median im Bereich von 0,5 bis 1,0 relativ hoch ist, lohnt es sich diesem Widerspruch weiter nachzugehen. Dazu wurde in einer relativen Betrachtung der Mittelwert des Interesses aller Teilnehmenden an der Datenanalyse berechnet und mit dem Mittelwert aller anderen Themen verglichen (vgl. Tabelle 12.2): Dieses mittlere Interesse an allen anderen Themen liegt bei 1,9 (1,7 bzw. 2,1 bei getrennter Betrachtung der Gruppen), während das mittlere Interesse an Datenanalysen bei 1,0 (in den Gruppen 0,7 bzw. 1,4) liegt. Damit ist auch bei dieser Betrachtung der deutliche Unterschied erkennbar.

	Mittelwert Gruppe 1	Mittelwert Gruppe 2	Mittelwert Gesamt
Alle Themen außer Datenanalyse	1,7	2,1	1,9
Alle bereits bekannten Themen	1,5	1,8	1,6
Datenanalyse	0,7	1,4	1,0
Standardabweichung der Werte zur Datenanalyse	0,82	0,66	0,70

Tabelle 12.2: Interesse der befragten Schülerinnen und Schüler an verschiedenen Themenbereichen. Skala: 0 ≙ *gar nicht interessant* bis 3 ≙ *sehr interessant*

Nur wenige der gefragten Themen waren bereits Unterrichtsthema in den beiden Klassen, nämlich *Technisches Zeichnen / CAD, Office-Programme* und *Computergrafik*. In einer weiteren Betrachtung wurde daher statt des mittleren Interesses an allen Themen als Referenz zu nehmen, nur das an diesen bereits bekannten Themen geäußerte Interesse betrachtet. Für diese Auswahl liegt das mittlere Interesse bei 1,6 (alle Teilnehmenden), 1,5 (Gruppe 1) bzw. 1,8 (Gruppe 2) und somit deutlich näher am Ergebnis für Datenanalysen, weiterhin schneiden die Datenanalysen aber klar schlechter ab.

Somit zeigt sich, dass das von den Schülerinnen und Schülern selbst eingeschätzte Interesse am Thema zwar tendenziell geringer ist als an anderen Themen. Die Ergebnisse unterliegen jedoch einer starken Streuung: Eine Betrachtung der Varianz der erhobenen Daten zur Datenanalyse zeigt, dass diese nur eine vage Tendenz widerspiegeln und keinesfalls als signifikant gelten können: Die Differenz des mittleren Interesses an bereits bekannten Themen und dem an Datenanalysen ist kleiner als die Standardabweichung der Ergebnisse zur Datenanalyse. Entsprechend ist deren Schwankungsbereich so groß, dass hier kaum valide Aussagen getroffen werden können. Der Widerspruch zur Einschätzung des im Unterricht gezeigten Interesses durch den externen Beobachter zeigt, dass hier eine weitere Untersuchung sinnvoll ist, da die Gründe für diesen Widerspruch offenbleiben und an dieser Stelle nicht erkannt werden können. Eine gemeinsame Vermutung des Autors dieser Arbeit und der unterrichtenden Lehrkraft ist jedoch, dass möglicherweise die Begriffswahl *Datenanalyse* ungünstig war, da entsprechend möglicherweise der Aspekt der Vorhersage – der im Unterricht zentral war – von den Schülerinnen und Schülern nicht miteinbezogen wurde. Eine weitere noch genauer zu prüfende Hypothese ist, dass die für die Realschule in der neunten Klasse zentrale Berufsorientierung zu kurz kam bzw. zu wenig hervorgehoben wurde. Außerdem kann auch vermutet werden, dass das Thema im Unterricht ausreichend detailliert thematisiert wurde, um den Schülerinnen und Schüler den Eindruck zu vermitteln, das Thema bereits ausreichend verstanden zu haben[67]. Entsprechend müsste der Kontext bzw. die Problemstellung des Unterrichts stärker aufgeweitet werden, um umfassendere Betrachtungen zu ermöglichen, auch das ursprünglich geplante Gruppenpuzzle könnte hier Abhilfe schaffen. Zusätzlich deutet sich an, dass die im Unterricht bereits thematisierten Aspekte allgemein als weniger interessant eingeschätzt werden als noch unbekannte Themen, was ein weiterer Einflussfaktor sein könnte.

Fragen zur Unterrichtssequenz. Die Fragen zur Unterrichtssequenz zeigen im Allgemeinen ein eher positives Ergebnis (vgl. Abbildung 12.3 und Tabelle 12.3): Die Schülerinnen und Schüler stimmen den Aussagen, dass die Aufgaben im Allgemeinen gut lösbar und das genutzte Werkzeug gut bedienbar waren teils eher zu (beide Median 3,0[68]). Sehr positiv ist, dass die Schülerinnen und Schüler selbst das Gefühl hatten, nun zu verstehen, was mit Daten gemacht werden kann (Median 4), obwohl sie sich nur zum Teil zutrauen, selbst solche Analysen durchzuführen (Median 2,0).

[67] Darauf deutet die Antwort auf die entsprechende Frage „Ich verstehe jetzt, was mit Daten gemacht werden kann“ hin, die später ausgewertet wird.

[68] Die Antworten wurden auf die Skala $0 \mathrel{\hat{=}} \textit{stimme nicht zu} \ldots 4 \mathrel{\hat{=}} \textit{stimme zu}$ abgebildet.

Frage	Gruppe 1				Gruppe 2				Gesamt			
	Mittelwert	Median	Modus	$\bar{d}_{0.5}$	Mittelwert	Median	Modus	$\bar{d}_{0.5}$	Mittelwert	Median	Modus	$\bar{d}_{0.5}$
Aufgaben gut lösbar	2,93	3,00	3,00	3,00	2,92	3,00	4,00	4,00	2,92	3,00	3,00	3,00
Analysewerkzeug gut bedienbar	3,08	3,00	4,00	4,00	3,00	3,00	4,00	4,00	3,04	3,00	4,00	4,00
Verständnis davon, was mit Daten gemacht werden kann	3,07	3,00	4,00	4,00	3,58	4,00	4,00	4,00	3,31	4,00	4,00	4,00
Kann selbst Daten analysieren	2,36	2,00	2,00	2,00	2,08	2,00	2,00	2,00	2,23	2,00	2,00	2,00
Mehr Spaß als bei anderen Themen im IT-Unterricht	1,57	1,50	1,00	1,00	1,67	1,50	0,00	0,00	1,62	1,50	1,00	1,00
Mehr über Datenanalyse im Unterricht gewünscht	1,38	1,00	1,00	1,00	1,33	1,50	0,00	0,00	1,36	1,00	0,00	0,00
Meine Daten soll nicht jeder bekommen	3,15	4,00	4,00	4,00	2,92	3,00	3,00	3,00	3,04	3,00	4,00	4,00
Thema war kompliziert	1,79	1,50	1,00	1,00	2,08	2,00	4,00	4,00	1,92	2,00	1,00	1,00
Vermeidung einer Weitergabe von Daten	2,86	3,00	3,00	3,00	2,50	2,50	4,00	4,00	2,69	3,00	4,00	4,00
Mit meinen Daten kann niemand etwas anfangen	1,64	1,50	0,00	0,00	1,42	0,50	0,00	0,00	1,54	1,00	0,00	0,00
Überrascht wie einfach Datenanalysen sind	2,43	2,50	3,00	3,00	2,17	3,00	3,00	3,00	2,31	3,00	3,00	3,00

Tabelle 12.3: Antworten der befragten Schülerinnen und Schüler zu den Fragen zur Unterrichtssequenz. Skala: 0 ≙ *gar nicht interessant* bis 3 ≙ *sehr interessant*

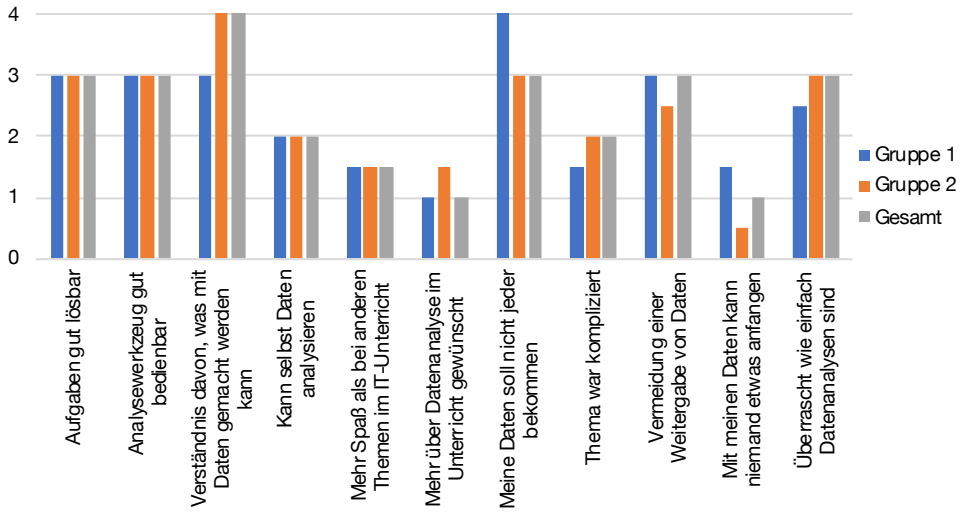


Abbildung 12.3: Median der Antworten der befragten Schülerinnen und Schüler zu den Fragen zur Unterrichtssequenz. Skala: 0 ≙ *stimme nicht zu* bis 4 ≙ *stimme zu*

Die Fragen, ob das Thema mehr Spaß als andere im IT-Unterricht gemacht haben und ob sie mehr über Datenanalysen im Unterricht wünschen, zeigen ein ähnliches Bild wie die Frage nach dem Interesse der Lernenden: Mit einem Median von 1,5 bzw. 1,0 werden diese Aussagen eher abgelehnt. In der Beobachtung schien es, als seien die drei Doppelstunden den Schülerinnen und Schülern stellenweise zu langwierig gewesen, was möglicherweise dadurch bedingt war, dass die Arbeitszeiten bei den meisten länger als geplant waren, da gerade die Arbeitsphasen häufig durch einige Schüler gestört wurden. Die Aussage zeigt daher klar, dass der Versuch unternommen werden muss, das Thema etwas spannen-

der erscheinen zu lassen, beispielsweise indem mehr verschiedene Beispiele eingebracht werden.

Trotz des eher geringen Interesses am Thema äußerten die Schülerinnen und Schüler jedoch in den Fragen zur Privatsphäre eine stärker reflektierte Wahrnehmung als in der anfänglichen Unterrichtsbeobachtung vermutet: Am Anfang der Sequenz wurde von vielen Schülerinnen und Schülern beispielsweise eine Datenweitergabe an Dritte oder die Einsehbarkeit von Daten durch diese als kaum problematisch angesehen. Dies sieht hier deutlich anders aus: Mit einem Median von 3,0 wird der Aussage „Ich möchte nicht, dass jeder Daten bzw. Informationen über mich bekommt“ eher zugestimmt, es wird auch stark angezweifelt (Median 1,0), dass niemand mit den persönlichen Daten der Schülerinnen und Schüler etwas anfangen kann.

Die Antworten der Lernenden rechtfertigen somit die Vermutung, dass trotz des unerwartet geringen geäußerten Interesses, viele von ihnen einen guten Eindruck von Datenanalysen und Vorhersagen bekommen haben. Ein Bezug zu ihrem Alltag und ihrer Privatsphäre konnte anscheinend hergestellt werden, was sich dadurch zeigt, dass die Schülerinnen und Schüler im Fragebogen einen eher reflektierten Umgang mit ihren Daten zeigten. Somit kann angenommen werden, dass der Unterricht einen gewissen Erfolg in diesem Bereich zeigte.

Erworbenes Wissen. Die nächste Frage des Fragebogens forderte die Schülerinnen und Schüler auf, im Rahmen kurzer Aussagen zu reflektieren, was sie im Unterricht gelernt hatten. Von den meisten Teilnehmenden wurden jedoch nur einzelne Begriffe genannt. Diese wurden ausgewertet, indem die genannten Aspekte zu verschiedenen induktiv gebildeteten Kategorien zusammengefasst und die Anzahl der Vorkommnisse gezählt wurden. Dadurch konnten sieben Kategorien identifiziert werden, von denen insbesondere zwei besonders häufig genannt wurden (vgl. Tabelle 12.4): Durch den Unterrichtskontext lässt sich erklären, dass 11 von 27 Schülerinnen und Schülern Aspekte aus dem Bereich „*Datensammlung, Analyse und Verkauf von Daten durch Unternehmen*“ genannt haben. Aber auch der Bereich „*Ablauf bzw. die Funktionsweise von Datenanalysen*“ inklusive verschiedener Aspekte die diesem Prozess, der den zentralen Bestandteil des Unterrichts darstellte, zuzuordnen sind, wurde von elf Schülerinnen bzw. Schülern genannt. Dies zeigt, dass die angestrebten Themen auch im Unterricht klar erkennbar waren und den Lernenden im Gedächtnis blieben. Für immerhin sechs Personen war außerdem die Einfachheit von Datenanalysen ein zentraler Aspekt, was zeigt, dass diese Thematik auch in dieser Altersgruppe durchaus verständlich ist.

Während diese Frage einen Einblick in die von den Lernenden als zentral wahrgenommen Unterrichtsthemen gab, wurde in den drei darauffolgenden Fragen deren Wissen hinsichtlich drei konkreter Themen überprüft. Diese Fragen wurden ausgewertet, indem die Antworten der Schülerinnen und Schüler qualitativ bewertet wurden. Statt sie jedoch klassisch mit Schulnoten zu bewerten, wurden sie den Kategorien *unbeantwortet, falsche Antwort, Ansätze erkennbar, unpräzise ausgedrückt, am Beispiel korrekt* und *allgemeingültig korrekt*

Bereich	Anzahl Nennungen
Datensammlung, Analyse, Verkauf durch Unternehmen	11
Anwendungsbeispiel *Werbung*	1
Ablauf/Funktionsweise von Datenanalysen	11
Einfachkeit von Datenanalysen	6
Möglichkeiten und Umfang von Datenanalysen	3
Schlussfolgerungen: Wert, Vertrauen, Handlungsleitfaden	3
Qualität der Analyse	2

Tabelle 12.4: Von den Schülerinnen und Schülern genannte Bereiche auf die Frage, was sie im Unterricht gelernt haben.

zugeordnet, da auf diese Weise ein klarerer Einblick in das von den Lernenden erworbene Wissen gewonnen werden kann. Den Kategorien wurden dabei beispielsweise folgende Antworten zugeordnet:

- *unbeantwortet:* Leere Antworten, sinnlose Kommentare oder Antworten die in keinerlei Zusammenhang zur Fragestellung stehen.
- *falsche Antwort:* Antworten, die zumindest einen gewissen Bezug zur Frage aufweisen, aber in eine falsche Richtung gehen. Beispielsweise nannte eine Person auf die Frage nach den Schritten des Analyse- und Vorhersageprozesses vier verschiedene Attribute, die vorher im Unterricht in anderem Kontext genannt wurden. Die Antwort beschäftigte sich daher mit der Datenanalyse, aber griff einen völlig falschen Aspekt auf, sodass sie klar als falsch gewertet werden konnte.
- *Ansätze erkennbar:* Die Antwort weist zwar Aspekte auf, die in die richtige Richtung gehen, ist aber im Großen und Ganzen unklar oder weist starke Schwächen auf. Beispielsweise wurde folgende Antwort auf die Frage nach den Schritten des Analyse- und Vorhersageprozesses in diese Kategorie einsortiert: *Datenauswertung, Zusammenhänge erstellen*. Diese weist zwar richtige Ansätze auf, indem Zusammenhänge bzw. besser Muster in den Daten gesucht und diese dazu ausgewertet werden, aber weitere Aspekte, wie die Anwendung der Muster zur Vorhersagegenerierung oder die Erhebung/Sammlung der Daten fehlen.
- *unpräzise ausgedrückt:* Es ist zwar erkennbar, dass die Antwort vermutlich das Richtige meint, sie ist aber eher unpräzise und interpretierbar. Dazu zählt beispielsweise die Antwort *mehr Daten* auf die Frage nach Möglichkeiten zur Verbesserung der Analysequalität. Diese Antwort ist nicht vollständig korrekt und unpräzise, da sie nicht benennt welche Datenmenge erhöht werden muss (die Ausgangsdatenmenge, die Stichprobengröße, die Menge an Daten, auf welche die Vorhersage angewandt wird, ...). Aus dem Unterrichtskontext ist jedoch zu vermuten, dass das richtige gemeint ist.
- *am Beispiel korrekt:* Diese Kategorie wurde gewählt, wenn an einem konkreten Beispiel ein Sachverhalt (weitestgehend) korrekt erläutert, aber nicht verallgemeinert beschrieben wurde. Dies ist beispielsweise der Fall, wenn der Analyseprozess konkret im Hinblick auf die Begriffswahl im Analysetool Orange 3 statt allgemein beschrieben

oder der Begriff Klassifikation auf die Einordnung von Menschen in verschiedene Gruppen hin reduziert wurde.

- *allgemeingültig korrekt:* Diese Kategorie wurde für (weitestgehend) korrekte Antworten vergeben, die ausreichend präzise ausgedrückt und nicht nur am Beispiel, sondern allgemeingültig ausgeführt sind.

Durch diese Einordnung konnte ein Einblick in das Wissen der Lernenden zu den drei abgefragten Unterrichtsthemen gewonnen werden. Der Begriff *Klassifikation* schien dabei in beiden Gruppen weitestgehend ähnlich unbekannt oder nur in Ansätzen bekannt zu sein (vgl. Abbildung 12.4), obwohl er im Unterricht umfangreich thematisiert und mehrfach wiederholt wurde. Dies spiegelt das auch in der Beobachtung gewonnene Bild wider, dass beide Gruppen sich außerhalb des Unterrichts kaum mit dem Fach beschäftigten und die Unterrichtsthemen auch zu Hause nicht wiederholten, sodass nur vages Faktenwissen erworben werden konnte. Laut der unterrichtenden Lehrkraft scheint dies ein allgemeines Problem in beiden Klassen zu sein, sodass diese Ergebnisse kaum verwundern. Die falsche Erklärung des Begriffs durch immerhin sechs Schülerinnen und Schülern kann verschiedene Ursachen haben, da vier dieser Personen aus der ersten Klasse kommen, liegt die Vermutung nahe, dass die deutliche Unterrichtsstörung in dieser Klasse den Lernerfolg an dieser Stelle gemindert hat bzw. es sich bei diesen Teilnehmenden um solche handelt, die keine Beteiligung am Unterricht zeigten.

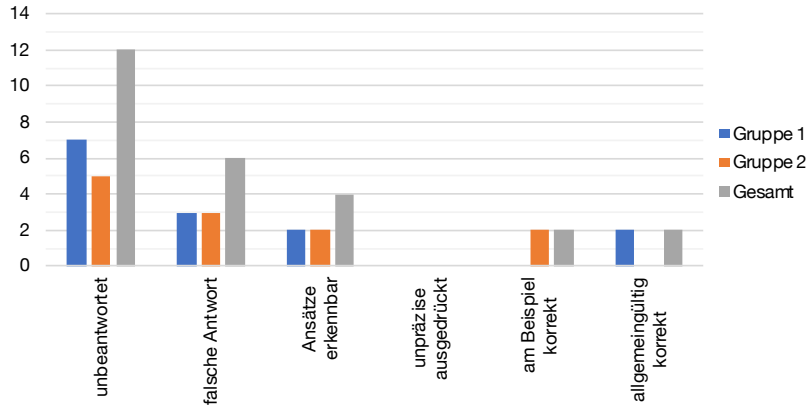


Abbildung 12.4: Auswertung der Antworten der Schülerinnen und Schüler zur Definition des Begriffs Klassifikation.

Auch die Frage zu möglichen Maßnahmen zur Erhöhung der Analysequalität wurde in beiden Kursen ähnlich beantwortet. Dabei ist erfreulich, dass diese nur durch einen Teilnehmer oder eine Teilnehmerin unbeantwortet blieb sowie durch einen bzw. eine falsch beantwortet wurde (vgl. Abbildung 12.5). Bei 18 von 26 Schülerinnen und Schülern und damit einer klaren Mehrheit war eine korrekte Antwort erkennbar, die aber in 15 Fällen nur relativ unpräzise ausgedrückt wurde. Dies kann dadurch erklärt werden, dass aus Zeitgründen die Sicherung nur sehr kurz ausfiel. Das relativ gute Abschneiden beider Klassen in dieser Frage liegt vermutlich daran, dass die Erhöhung der Analysequalität ein wichtiges Thema der dritten Doppelstunde war und damit dem Fragebogen direkt vorherging.

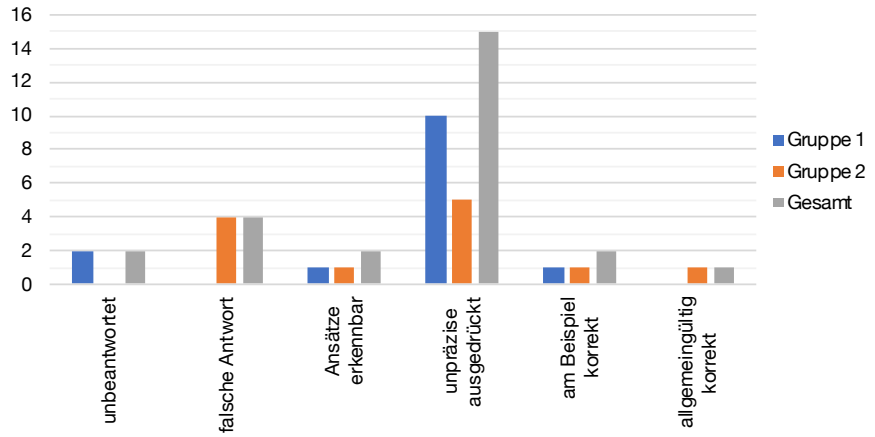


Abbildung 12.5: Auswertung der Antworten der Schülerinnen und Schüler zu Maßnahmen zur Erhöhung der Analysequalität.

Die dritte Frage, in der die Teilnehmenden die Schritte des Analyse- und Vorhersageprozesses nennen/beschreiben sollten, wurde deutlich unterschiedlich in beiden Gruppen bearbeitet: Während in der zweiten Gruppe sieben von zwölf Schülern eine korrekte oder nur unpräzise ausgedrückte Antwort gaben, waren es in der ersten Gruppe nur vier der vierzehn Schülerinnen und Schüler (vgl. Abbildung 12.6). Umgekehrt verhielt es sich bei den falschen Antworten, hier waren in der zweiten Gruppe nur zwei, in der ersten Gruppe hingegen vier zu verzeichnen, während die Enthaltungen gleich hoch waren. Obwohl dieses in der ersten Doppelstunde erworbene Wissen zwar in beiden darauffolgenden Stunden am Anfang wiederholt wurde, kann vermutet werden, dass die Schüler der zweiten Gruppe hier zwar auch nur zum Teil korrektes Faktenwissen erworben haben, vermutlich aber zumindest ein besseres Verständnis erworben haben als die Schülerinnen und Schüler der ersten – und wesentlich unruhigeren – Gruppe.

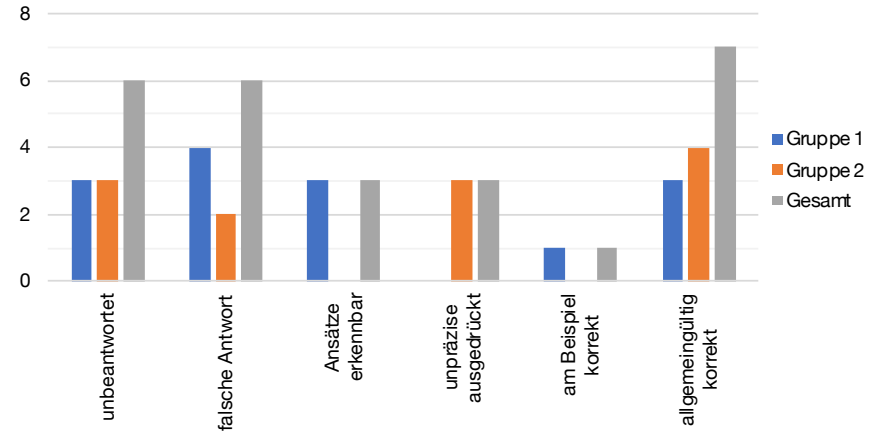


Abbildung 12.6: Auswertung der Antworten der Schülerinnen und Schüler zur Beschreibung des Analyseprozesses.

Anwendung des Wissens auf eine neue Situation. Während die Schülerinnen und Schüler in der vorherigen Frage nach konkretem Faktenwissen zu verschiedenen Begriffen gefragt wurden, sollte im letzten Teil des Fragebogens eine konkrete Situation betrachtet werden, die ihnen bisher im Unterricht so nicht begegnet ist, sodass sie ihr erworbenes Wissen auf diese Situation übertragen sollten. Als Kontext wurden dazu Datenanalysen im Gesundheitswesen gewählt. Auch zu diesen Fragen wurden, ähnlich wie bei der vorherigen, Kategorien gebildet und die Antworten der Schülerinnen und Schüler in diese einsortiert (vgl. Tabelle 12.5).

	Gruppe 1	Gruppe 2	Gesamt
Frage 1: Daten und Datenquellen			
unbeantwortet	3	0	3
naheliegende Daten genannt	9	10	19
mehr als nur naheliegende genannt	2	2	4
Frage 2: Potentielle Probleme			
unbeantwortet	0	1	1
kein nachvollziehbares Problem erkannt	0	2	2
oberflächliches Problem erkannt	11	7	18
tiefergehendes Problem erkannt	3	2	5
Frage 3: Irrtumsfreiheit von Datenanalysen			
unbeantwortet	7	1	8
falsche Antwort	0	0	0
richtiger Ansatz erkennbar	0	4	4
(min.) eine korrekte Idee genannt	7	7	14

Tabelle 12.5: Auswertung der Antworten der Schülerinnen und Schüler zu den Fragen zum Kontext Daten im Gesundheitswesen.

Als erste Frage wurden die Schülerinnen und Schüler aufgefordert, verschiedene Daten und deren Herkunft zu nennen, die Ärzte für potenzielle auf Datenanalysen basierende Diagnosen nutzen könnten. Dabei konnten zwar die meisten Teilnehmenden verschiedene logische Ideen nennen, die jedoch in den meisten Fällen eher naheliegend waren (übliche Krankheits- und Personendaten). Nur vier Personen nannten weitere Daten, wie beispielsweise aus sozialen Medien oder Einkäufe im Supermarkt. Um hier ein tiefergehendes Bewusstsein zu schaffen, wäre vermutlich das geplante aber aus Zeitgründen nicht mehr durchführbare Gruppenpuzzle zu verschiedenen Datenanalysen hilfreich gewesen.

Auch die von den Schülerinnen und Schülern genannten Probleme, die auftreten könnten, wenn Ärzte ihre Patienten zukünftig primär anhand von Daten untersuchen, waren eher oberflächlicher Natur: Beispielsweise wurde dabei häufig genannt, dass als Ergebnis falsche Medikamente verschrieben werden könnten. Obwohl diese Antwort sicherlich an sich korrekt ist, wurde nicht über die Ursachen nachgedacht bzw. diese nicht genannt. Diese und ähnliche Antworten wurden von 18 Teilnehmenden genannt. Nur fünf Schülerinnen bzw. Schüler nannten Probleme, deren Erkennen etwas mehr Verständnis benötigt, insbesondere die Nichterkennung seltener Krankheiten.

Bei der letzten Frage wurden die Schülerinnen und Schüler aufgefordert zu beurteilen, ob eine Analyse sich irren kann oder grundsätzlich fehlerfrei ist und mögliche Probleme

in dieser Hinsicht nennen. Während ein Großteil der Teilnehmenden zwar eine richtige Entscheidung traf und 14 von ihnen auch zumeist ein nachvollziehbares Problem nennen konnten, gingen jedoch nur wenige Schülerinnen bzw. Schüler stärker in die Tiefe als unbedingt gefordert.

Zusammenfassung Zusammenfassend zeigt die Fragebogenstudie im Allgemeinen eher mäßige Ergebnisse: Die Schülerinnen und Schüler nehmen das Thema als weniger interessant wahr als erwartet. Gleichzeitig konnten sie in den Fragebögen zu verschiedenen Themen nur eher geringes Wissen demonstrieren. Es zeigte sich jedoch in vielen Fällen, dass verschiedene Aspekte verinnerlicht wurden, auch konnte das erworbene Wissen zumindest zum Teil auf neue Situationen übertragen werden. Das trotzdem relativ schlechte Abschneiden lässt sich vermutlich auf die im Unterricht eher geringe Motivation zurückführen, die durch die unterrichtende Lehrkraft sowohl für diese Unterrichtsstunde als auch den IT-Unterricht im Allgemeinen bestätigt wurde (vgl. nächster Abschnitt). Entsprechend kann als positiv hervorgehoben werden, dass trotz der – insbesondere in der ersten Gruppe – eher problematischen Unterrichtssituation, verschiedene Schülerinnen und Schüler einen eher guten Eindruck vom Unterrichtsthema bekommen haben und durchaus zeigen konnten, dass sie nicht nur Faktenwissen erworben haben, sondern dieses auch auf neue Situationen übertragen können. Eine vertieftere Betrachtung der Ursachen verschiedener Antworten benötigt jedoch eine weitere Untersuchung mit einer höheren Zahl an Teilnehmenden. Trotz der großen Streuung können die Ergebnisse Indizien für die weitere Entwicklung des Unterrichts liefern, die in Abschnitt 12.3.4 miteinbezogen werden.

Lehrerperspektive: Interview

Ein deutlich anderes Bild als die Schülerperspektive zeigte das Lehrerinterview, das eher die im vorletzten Abschnitt geschilderten Beobachtungen bestätigt. Dieses Interview wurde zwei Tage nach der letzten Doppelstunde retrospektiv durchgeführt und mit Zustimmung der Lehrkraft aufgezeichnet. Es wurde an den bereits zuvor genannten Leitfragen orientiert durchgeführt, nach denen auch die im Folgenden dargestellten Ergebnisse des Interviews gegliedert sind.

Wie war der Gesamteindruck von der Unterrichtssequenz? Die Lehrkraft verdeutlichte im Interview einen positiven Gesamteindruck der Unterrichtssequenz. Insbesondere betonte sie, dass sie persönlich das Thema spannend und wichtig findet, da heute viele Daten weitergegeben werden, ohne darüber nachzudenken, was mit diesen passieren kann bzw. gemacht wird. Sie äußerte die Hoffnung, dass die Schülerinnen und Schüler diese Datenweitergabe und -nutzung nun stärker hinterfragen. Hinsichtlich der Unterrichtsdurchführung wurde betont, dass die beiden Gruppen stark unterschiedlich waren – was die Beobachtung im Unterricht bestätigte. Außerdem wurde angesprochen, dass manche Aspekte zu ausführlich thematisiert wurden und es so für die Schülerinnen und Schüler

zu viel Zeit mit dem Thema war. Trotzdem stimmte die Lehrkraft dem Vorschlag des Interviewers zu, das Thema *Korrelation vs. Kausalität* vertiefter zu betrachten, da auch aus Lehrerperspektive die Lernenden Probleme bei diesem Thema zeigten – nach Aussage der Lehrkraft liegt dies aber vermutlich nicht nur an der knappen Thematisierung, sondern auch daran, dass die meisten Schülerinnen und Schüler sich die Materialien zu Hause nicht noch einmal angesehen haben und gar nicht versuchen, etwas zu verstehen.

Wie gut hat der Unterricht funktioniert? Der Eindruck der Lehrkraft wurde zuerst als *gut* zusammengefasst, dies wurde jedoch noch konkretisiert: *„In der ersten Gruppe war ja alles recht zäh, da muss man alles aus der Nase ziehen“*. Entsprechend ging der Unterricht deutlich schlechter voran. Im Allgemeinen funktioniert in dieser Gruppe der Unterricht, solange die Schülerinnen und Schüler konkrete Aufgaben bekommen, sobald jedoch beispielsweise eine Diskussion stattfinden soll, wird es nach Aussage der Lehrkraft oft zäh. Dies bestätigt auch die Beobachtung im Unterricht. Die andere Gruppe war laut der Lehrkraft hingegen insbesondere in Diskussionsphasen sehr aktiv und beteiligte sich rege, sodass dabei die Gefahr vorherrschte, vom Thema abzuweichen. Die Lehrkraft musste hier daher regelmäßig wieder zum Thema zurückführen. Beide Eindrücke sind jedoch nicht auf diese Unterrichtssequenz beschränkt, sondern zeigen sich bei beiden Kursen allgemein im IT-Unterricht, sodass hier keinerlei Einfluss des konkreten Unterrichtsthemas angenommen werden kann.

Was haben die Schülerinnen und Schüler gelernt/mitgenommen? Nach Einschätzung der Lehrkraft haben die Schülerinnen und Schüler vermutlich das angestrebte Wissen erworben: Sie geht davon aus, dass diese nun besser darauf schauen, was mit Daten passiert und wissen, warum das so ist. Als zentrale Inhalte kennen die Schülerinnen und Schüler nun eine Möglichkeit zur Datenanalyse und Vorhersage und verstehen auch das Prinzip der Kategorisierung.

War der allgemeine Aufbau der Unterrichtssequenz sinnvoll? Grundsätzlich erachtete die Lehrkraft es als sinnvoll, die Analyse erst theoretisch/händisch zu thematisieren, bevor dies dann am Computer ausprobiert wird. Jedoch merkte sie an, dass es am Anfang *„vielleicht ein bisschen zu lang“* war und es gut wäre, wenn dieser Bereich etwas kürzer wäre. Ein Vorschlag des Interviewers, zu versuchen die beiden Phasen stärker miteinander zu verflechten, sodass bereits früher am Computer gearbeitet werden könnte, wurde als sinnvoll erachtet. Es wurde jedoch auch angemerkt, dass auch das manchen der Schülerinnen und Schüler nicht helfen wird, da es Personen gab, die mehrfach äußerten, dass sie *„das gar nicht kapieren“* – die Lehrkraft vermutete hier jedoch, dass es bei diesen vielleicht auch einfach so war, *„dass sie nichts verstehen wollen“*.

Wie starkes Interesse haben die Schülerinnen und Schüler am Thema gezeigt? In der ersten Gruppe wurde laut der Lehrkraft das Interesse insbesondere zu dem Zeitpunkt

geweckt, als es um sie persönlich ging: *„[. . .] mit den Noten, da sind sie ja dann aufgewacht."* Dieses Interesse ist jedoch auch schnell wieder abgeflaut, sie haben dann wieder weniger Motivation gezeigt. Auf Nachfrage des Interviewers wurde geäußert, dass dies nicht ungewöhnlich ist, sondern in dieser Gruppe immer so ist und es nicht am Thema liegt. Hinsichtlich der zweiten Gruppe äußerte die Lehrkraft überzeugt, dass das Thema die Schüler interessiert hat.

Welche Probleme wurden bei der Durchführung aus Perspektive der Lehrkraft erkannt? Bei der Nachfrage nach Problemen äußerte die Lehrkraft nochmals, dass es sinnvoll wäre, den ersten (händischen/theoretischen) Teil etwas kürzer zu fassen, beziehungsweise wie zuvor vorgeschlagen die beiden Phasen umzustrukturieren, sodass die Schülerinnen und Schüler schneller selbst etwas machen können. Ansonsten wurde keine Probleme erkannt, stattdessen wurde betont, dass die Nutzung des Programms den Schülerinnen und Schülern gefallen hat. Es wurde darauf hingewiesen, dass beispielsweise die Klassifikationsbäume intensiv besprochen werden sollten, indem ein Weg von der Wurzel zu einer Entscheidung / einem Blatt gemeinsam nachvollzogen wird, damit die Grafik für alle verständlich wird. Zusätzlich betonte die Lehrkraft, dass es schade war, dass die Zeit nicht mehr ausreichte, um den letzten Teil (Gruppenpuzzle) am Ende noch durchzuführen, da dies laut ihrer Einschätzung noch interessant gewesen wäre.

Fazit: Wie geeignet ist das Thema für den Informatikunterricht? Abschließend äußerte die Lehrkraft, dass sie das Thema als wichtig für den Informatikunterricht einschätzt, insbesondere die Sensibilisierung. Als Fazit äußerte sie, dass sie das Thema auch zukünftig erneut thematisieren würde, *„vielleicht nicht ganz so ausführlich, weil einfach die Zeit normalerweise nicht so da ist, aber prinzipiell ist es – finde ich – schon wichtig."*

12.3.4 Synthese und Interpretation: Leitlinien für den Unterricht

Obwohl die Erprobung nur mit zwei Klassen derselben Schule und mit derselben Lehrkraft stattfand, kann sie die Praxistauglichkeit der konzipierten Unterrichtssequenz bestätigen und zeigt, dass für die Schülerinnen und Schüler spannende und relevante Themen ausgewählt wurden, die diese auch zur Diskussion anregen konnten. Es muss jedoch berücksichtigt werden, dass beide Klassen sehr unterschiedlich waren, was sowohl die Lehrkraft als auch die Unterrichtsbeobachtung bestätigten, und sich auch am Fragebogen durch verschiedene Differenzen zwischen beiden Gruppen zeigt. Gleichzeitig muss berücksichtigt werden, dass die Schülerbefragung deutlich unterschiedliche Ergebnisse im Vergleich mit der Beobachtung sowie der Lehrerbefragung zeigt: Während Einschätzung der Schülerinnen und Schüler insbesondere eher geringes Interesse am Thema zeigt und nur einen eingeschränkten Kompetenzzuwachs vermuten lässt, war die Einschätzung insbesondere in diesen Bereichen sowohl aus Sicht des externen Beobachters als auch der Lehrkraft konträr dazu. Durch die kleine Stichprobe und die sich zum Teil klar widersprechenden

Ergebnisse der drei Untersuchungsmethoden, können die Ergebnisse dieser Erprobung nicht als stichhaltig und zwingend auf andere Unterrichtssituationen und auch Schularten übertragbar angenommen werden. Die genauere Betrachtung insbesondere der Differenzen der verschiedenen Perspektiven kann aber wichtige Indizien für die Weiterentwicklung der Unterrichtssequenz und für die zukünftige Konzeption weiterer Unterrichtseinheiten liefern. Aus diesem Grund werden, nach einer kurzen Zusammenfassung der zentralen Aspekte der Erprobung, auf dieser Basis Leitlinien entwickelt, die sich in der Erprobung als wichtig für eine erfolgreiche Durchführung erwiesen haben.

Zusammenfassend kann aus den drei zuvor beschriebenen Untersuchungen gefolgert werden, dass die Unterrichtsziele in Weiten teilen erreicht werden konnten: Obwohl die Schülerinnen und Schüler selbst angaben, wenig Interesse am Thema zu haben und in der Befragung nur wenig konkret fassbares Fachwissen zeigten, schienen sie einen klaren Eindruck von der Datenanalyse und Vorhersage bekommen zu haben und zeigten im Unterricht, dass sie deutlich mehr Wissen erworben haben, als ihnen bewusst zu sein schien. Insbesondere den Eindrücken der Lehrkraft und des Beobachters nach schien ein Großteil beider Klassen die angesprochenen Themen größtenteils gut nachvollziehen zu können und einen klaren Eindruck der Möglichkeiten von Datenanalysen bekommen zu haben. Entsprechend konnte die Unterrichtssequenz als Sensibilisierung für die Schülerinnen und Schüler wirken, was sich auch in den Rückmeldungen im Fragebogen zum Teil gezeigt hat. Ein spannender Aspekt zur Vertiefung wäre jedoch, zu untersuchen ob die Intervention das Verhalten der Schülerinnen und Schüler in Situationen, die eine Datenweitergabe fordern, beeinflusst hat. Auch muss eine weitere Erprobung mit mehreren zusätzlichen Schulklassen und verschiedenen Lehrkräften erfolgen, um allgemeingültigere und von subjektiven Einflüssen unabhängigere Ergebnisse zu erhalten.

Hinsichtlich der Machbarkeit im Unterricht und des Schwierigkeitsgrads waren sich jedoch alle drei Untersuchungen einig, dass die Aufgaben für die Schülerinnen und Schüler nachvollziehbar und lösbar waren. Entsprechend treffen die nicht nur von Lehrkräften, sondern auch von verschiedenen Wissenschaftlerinnen und Wissenschaftlern geäußerten Befürchtungen, dass das Thema für die Schule zu komplex ist und die Schülerinnen und Schüler überfordern würde, nicht zu. Dies ist insbesondere spannend, da durch die verschiedenen Klassenstufen und Schularten verschiedene intellektuelle Niveaus angesprochen wurden, die aber durch die Fokussierung auf die informatischen Prinzipien und Ideen hinter dem Thema beide weder über- noch unterfordert wurden. Es schien sogar so, dass auch an Realschulen der Schwierigkeitsgrad noch etwas angepasst werden kann, um die besseren und interessierteren Teilnehmenden noch stärker zu fordern und möglicherweise – durch tiefergehende Aspekte – auch deren Interesse am Thema zu steigern. Dieses Interesse am Thema wurde durch den Beobachter und die unterrichtende Lehrkraft als größtenteils relativ hoch eingeschätzt, obwohl die Schülerinnen und Schüler es selbst als eher niedrig eingestuft haben. Eine mögliche Vermutung ist, dass die Schülerinnen und Schüler stellenweise unterfordert waren (insbesondere gute und interessierte Schülerinnen und Schüler im sehr unruhigen Kurs) und somit die Aufmerksamkeit im Unterricht fehlte. Eine andere, aus der subjektiven Einschätzung des Beobachters nach wahrscheinlichere Hypothese ist

jedoch, dass gerade das Problem, dass Schülerinnen und Schüler regelmäßig Kausalzusammenhänge suchten, obwohl das Ziel die Diskussion von Korrelationen war, zu deutlichen Verständnisschwierigkeiten führte und die Methodik somit von vielen Lernenden als sinnlos, realitätsfern oder einfach unverständlich eingeschätzt wurde. Zusätzlich könnten eine ausführlichere Synthese und der Transfer des Wissens auf weitere Beispiele – die aus Zeitgründen wesentlich knapper ausfallen mussten als geplant – die Anwendbarkeit des Wissens stärker verdeutlichen und somit auch zu einer positiveren Einschätzung beitragen. Trotz dieses eher geringen von den Lernenden geäußerten Interesses, zeigte sich insbesondere im Unterricht, dass die meisten Schülerinnen und Schüler – obwohl ihnen dies dem Fragebogen nach nicht bekannt war – einerseits begannen ein gewisses Verständnis dafür zu entwickeln, was mit Daten gemacht werden kann, andererseits aber auch eine realistischere Einschätzung der Möglichkeiten zu bekommen, die sie heute selbst im Kontext der Datenanalyse haben.

Aus diesen Untersuchungen können daher anhand der positiven und negativen Erfahrungen folgende Leitlinien für die weitere Unterrichtsgestaltung im Bereich der Datenanalyse und -vorhersage abgeleitet werden, die möglicherweise aber auch auf das gesamte Themengebiet übertragbar sind:

- **Nutzen von Alltagsbeispielen als Ankerpunkte:** Die Nutzung von Alltagsbeispielen war für die Schülerinnen und Schüler sichtlich motivierend und hat dafür gesorgt, dass diese eine Verbindung zu ihrer Lebenswelt herstellen konnten. Dies zeigte sich insbesondere durch von ihnen eingebrachte eigene Beispiele. Die Wahl der Beispiele – und insbesondere die Analyse und Vorhersage der Schulnoten – war unter diesem Gesichtspunkt eindeutig förderlich. Eine kontinuierliche Ausrichtung der Unterrichtssequenz und das regelmäßige Heranziehen verschiedener Beispiele, wie Werbeschaltungen auf Webseiten, Produktempfehlungen bei Onlineshops oder Freundesvorschläge in sozialen Medien, scheint daher sinnvoll und hilfreich.

- **Schaffen einer Möglichkeit zur Diskussion:** Die Möglichkeit zur Diskussion wurde zwar nur von einer der beiden Gruppen intensiv genutzt, bei dieser hat diese jedoch klar zu einer intensiven Betrachtung des Themas beigetragen. Selbst in der weniger diskussionsfreudigen Gruppe wurden jedoch mehrfach eigene Beispiele und Ideen eingebracht – wenn auch nur von einem kleineren Teil der Schülerinnen und Schüler. Diese Möglichkeit sollte daher auf jeden Fall geboten und ein entsprechender idealerweise etwas variabler Zeitumfang eingeplant werden.

- **Abwechseln von händischen Datenanalysen und Automatisierung:** Die sowohl händische als auch automatisierte Durchführung der Datenanalysen im Unterricht hat sich als hilfreich erwiesen, um einerseits das Verständnis der Schülerinnen und Schüler für die zugrundeliegende Methodik zu fördern, ihnen andererseits aber auch einen Einblick in die Effizienz und das Potenzial realer Datenanalysen zu geben. Wie die Lehrkraft im Interview betonte und wie auch im Unterricht erkennbar war, war die anfängliche Arbeitsphase, die rein händisch stattfand, jedoch zu lang, was auf einige der Schülerinnen und Schüler augenscheinlich demotivierend wirkte. Ein

mehrfacher Wechsel zwischen händischer und computergestützter Analyse erscheint hier nach übereinstimmender Meinung des Autors und der durchführenden Lehrkraft sinnvoll, um den Unterricht abwechslungsreicher zu gestalten.

- **Nutzen echter Datensätze:** Während im Unterricht ursprünglich nur nebenbei erwähnt wurde, dass der für die computergestützte Analyse genutzte Datensatz auf echten Schülerdaten einer portugiesischen Schule basiert, schien dies für die Schülerinnen und Schüler ein wichtiger Aspekt zu sein: Im Laufe der Analyse wurde dies in beiden Gruppen nochmals nachgefragt und in der zweiten auch kurz die Weitergabe solcher Daten durch die portugiesischen Schulen hinterfragt. Dass die Daten – und damit deren Zusammenhänge – nicht ausgedacht sind, schien für Lernenden ein zentraler Aspekt zu sein, was sich insbesondere dadurch zeigte, dass hier deutlich weniger versucht wurde, die Sinnhaftigkeit der Daten zu hinterfragen als bei den Einführungsbeispielen, da die Korrektheit aufgrund der Echtheit der Daten anscheinend eher akzeptierbar war.

- **Ausführliches Thematisieren des Unterschieds zwischen kausalitäts- und korrelationsbasierten Analysen:** Wie bereits erwähnt, war es für die Schülerinnen und Schüler beider Gruppen naheliegend, jegliche entdeckten Zusammenhänge zu hinterfragen und zu versuchen, diese kausal zu begründen. Falls Zusammenhänge nicht kausal erklärbar waren, wurden diese eher angezweifelt, insbesondere bei den anfangs verwendeten künstlichen Datensätzen. Dies führte offensichtlich zu Verständnisproblemen insbesondere bei solchen Schülerinnen und Schülern, die die Thematik im Detail verstehen wollten und entsprechend motiviert waren. Die Thematisierung des Unterschieds zwischen kausalitäts- und korrelationsbasierten Analysen sollte daher im Unterricht ausgeweitet und stärker an Beispielen orientiert werden, um das Verständnis für den Unterschied zu verdeutlichen.

- **Reflektieren des Gelernten anhand mehrerer Anwendungsbeispiele:** Die geplante Reflexion des Gelernten, die die Unterrichtssequenz in Form eines Gruppenpuzzles mit verschiedenen weiteren alltagsnahen und realen Anwendungsbeispielen abschließen sollte, wäre eine wichtige Festigung gewesen. Sie hätte vermutlich auch motivierend beigetragen, indem die Übertragbarkeit des Wissens auf weitere Beispiele demonstriert worden wäre. Somit sollte auch bei Zeitproblemen vermieden werden, diese Phase komplett zu streichen. Eine alternative Anpassung der Unterrichtssequenz wird im nächsten Kapitel dargestellt.

- **Nutzen geeigneter und didaktisch reduzierter Werkzeuge:** Das gewählte Werkzeug hat sich im Unterricht klar bewährt. Ausschlaggebend dürften laut der Unterrichtsbeobachtung insbesondere die einfache Bedienoberfläche und die nachvollziehbare Struktur des Analyseaufbaus in Form einer (nicht explizit thematisierten) Datenflussmodellierung sein, aber auch die durch den Autor durchgeführte Reduktion des Tools um für die konkrete Sequenz nicht notwendige Module, die das Werkzeug deutlich komplexer gemacht hätten. Aufgrund des einfachen Aufbaus des Werkzeugs, kann jedoch bei Wahl dieses oder eines ähnlichen Werkzeugs darauf verzichtet werden,

eine Projektvorlage zur Verfügung zu stellen, da es den Schülerinnen und Schülern so im Unterricht möglich war, das Werkzeug freier zu explorieren.

Zusammenfassend konnten anhand dieser Unterrichtserprobung somit, trotz der zum Teil eher schwierigen Situation im Unterricht und der deutlich differenten Ergebnisse aus den verschiedenen Perspektiven auf den Unterricht, wichtige Ideen und Leitlinien für den Unterricht abgeleitet werden. Um weitere Erfahrungen miteinbeziehen und die Validität dieser Leitlinien zu überprüfen, müssen hier jedoch noch weitere Untersuchungen erfolgen, damit die Anzahl der Teilnehmenden Schülerinnen und Schüler erhöht, insbesondere aber durch Einbezug verschiedener Lehrkräfte die Unabhängigkeit der Ergebnisse von der Lehrkraft sichergestellt werden kann. Ein erster Ansatz dazu wurde im Rahmen einer weiteren Unterrichtserprobung gemacht, die insbesondere die vertiefte Reflexion des gelernten miteinbezieht und den Unterschied zwischen kausalen und korrelativen Zusammenhängen etwas stärker in den Fokus nahm, dabei aber die Möglichkeiten für die Lernenden, eigene Erfahrungen mit einem Analysewerkzeug zu gewinnen, einschränkte.

12.4 Weitere Unterrichtserfahrungen

Diese weitere Erprobung wurde in der neunten Jahrgangsstufe eines bayerischen Gymnasiums durchgeführt, konnte jedoch vom Autor dieser Arbeit nicht näher begleitet werden. Die dabei gewonnenen Erfahrungen basieren daher auf einer schriftlich durchgeführten Befragung der Lehrkraft im Nachgang des Unterrichts, die an denselben Leitfragen wie sie in der vorherigen Erprobung im Lehrerinterview gestellt wurden, ausgerichtet war. Auch dieser Unterricht wurde am Schuljahresende durchgeführt. Zuvor wurden alle durch den zum Untersuchungszeitpunkt aktuellen Lehrplan vorgeschriebenen Themen im Unterricht bereits thematisiert, sodass den Schülerinnen und Schülern insbesondere relationale Datenbanken, relationale Datenmodellierung und die Arbeit mit diesen (insbesondere mit SQL) bekannt waren. Der Unterricht wurde am selben Konzept orientiert wie die vorher beschriebene Unterrichtserprobung, jedoch wurden von der Lehrkraft, die langjährige Unterrichtserfahrungen mitbringt und auch in der zweiten Phase der Informatiklehrerausbildung tätig ist, verschiedene Anpassungen vorgenommen, um die eigenen Gegebenheiten zu berücksichtigen:

- Während im eigentlichen Konzept und auch in der zuvor beschriebenen Durchführung an der Realschule für die Arbeitsblätter 1 und 2 sowie 3 und 4 jeweils zusammen eine Doppelstunde vorgesehen war, wurden diese vier Arbeitsblätter – laut Aussage der Lehrkraft problemlos – in einer Doppelstunde bearbeitet und somit die Unterrichtssequenz auf diese Weise zeitlich komprimiert.
- Da im stark restriktierten Schulnetzwerk weder die Installation des Softwarepakets Orange 3 noch die Nutzung der zur Verfügung gestellten portablen Version ohne großen Zeitaufwand möglich war, wurde die Unterrichtssequenz an dieser Stelle angepasst: Statt die Schülerinnen und Schüler den Arbeitsauftrag mit Orange 3 be-

arbeiten zu lassen, wurde dieser in der Vorbereitung durch die Lehrkraft bearbeitet und in Form eines Videos festgehalten. Laut Rückmeldung der Lehrkraft konnte das Prinzip auf diese Weise trotzdem überzeugend demonstriert und eine Diskussion darüber angestoßen werden.

- Im Vergleich mit der Realschulerprobung war es hier jedoch möglich, auch das sechste Arbeitsblatt zu bearbeiten und somit die in den beiden Realschulklassen zu kurz gekommene Reflexion miteinzubeziehen. Hierzu wurde jedoch aus Zeitgründen auf die Nutzung des vorgeschlagenen Gruppenpuzzles verzichtet, stattdessen wurden in einer Gruppenphase die verschiedenen Beispiele für die Nutzung von Datenanalysen betrachtet und dann in einer Plenumsphase insbesondere über die Chancen und Risiken der Vorhersagen diskutiert.

Trotz der eingeschränkten Zeit und der Durchführung der beiden Doppelstunden kurz vor Schuljahresende zog die Lehrkraft ein sehr positives Fazit: Aus ihrer Sicht war das Thema sehr interessant und motivierend für die Schülerinnen und Schüler, die auch allgemein sehr interessiert am (Informatik-)Unterricht sind. Insbesondere betonte die Lehrkraft, dass bei *allen* Schülerinnen und Schülern ein wichtiger Lernerfolg erzielt wurde: Ihnen wurden die Augen hinsichtlich Datenanalysen geöffnet und Hintergründe aufgezeigt, die ihnen ansonsten verborgen geblieben wären. Dabei haben sie laut Lehrkraft aber auch gelernt, dass Zusammenhänge im Kontext von Datenanalysen nicht immer kausal begründbar sind, wie Analysen automatisiert durchgeführt werden können, wie daraus Vorhersagen getroffen werden und dass dies Vor- und Nachteile haben kann. Insgesamt wurde das Fazit gezogen, dass die kurze Unterrichtssequenz trotz der notwendigen Komprimierung problemlos funktioniert hat, zur Erreichung der Unterrichtsziele sehr geeignet ist und aus allgemeinbildender Sicht dringend notwendig ist.

Diese Rückmeldungen der Lehrkraft bestätigen die Eindrücke des Autors bei der Erprobung an der Realschule und zeigen, dass die entsprechenden zuvor gewonnenen Ergebnisse auch durch diese Erprobung größtenteils – durch die weniger nahe Begleitung aber nicht im Detail – bestätigt werden. Die aus der vorherigen Unterrichtserprobung abgeleiteten Leitlinien scheinen also auch durch diese zweite Erprobung bestätigt zu werden, wobei deutlich wurde, dass es insbesondere wichtig ist, den Schülerinnen und Schülern Einblicke in Datenanalysen zu geben, auch wenn sie diese im Größeren nicht unbedingt im Unterricht durchführen können.

Alles in allem hat sich der Aufbau und die Vorbereitung der Unterrichtssequenz daher im Unterricht in beiden Schulformen grundsätzlich bewährt, obwohl die Ergebnisse der Befragung der Schülerinnen und Schüler zeigen, dass weiteres Verbesserungspotential besteht. Zusätzlich konnten erste Ideen für eine Verbesserung der Materialien sowohl für Lehrerinnen und Lehrer als auch für Schülerinnen und Schüler gewonnen werden, die in eine neue Version dieser Materialien einfließen werden.

Teil V:

Abschluss

13 Zusammenfassung der Arbeit

13.1 Zusammenfassung

Das Ziel dieser Arbeit war die informatikdidaktische Aufarbeitung des Fachgebietes *Datenmanagement*, dessen neuere Aspekte trotz ihrer hohen Relevanz in Informatik, Gesellschaft und Alltag zugleich bisher allenfalls als Randthema der informatikdidaktischen Forschung thematisiert wurden.

Basierend auf dem Modell der Didaktischen Rekonstruktion für den Informatikunterricht bzw. aus dessen verschiedenen Perspektiven wurde der gesamte Themenkomplex in dieser Arbeit umfassend betrachtet:

- *Fachliche Klärung (Teil I):*
  Um einen Überblick über das Datenmanagement aus fachwissenschaftlicher Perspektive zu geben, wurde dieses anhand fünf zentraler Themen und Forschungsfelder detaillierter beschrieben: *Big Data*, *verteilte Datenspeicher* (inkl. *Cloud-Speicherung*), *Metadaten*, *Data Mining* und *Datenstromsysteme*. Diese Themen charakterisieren die Breite des Fachgebiets. Zusätzlich wurden Bezüge zu den verwandten Themenbereichen *Data Science* und *Data Literacy* hergestellt.

- *Klärung gesellschaftlicher Ansprüche:*
  Zur Begründung der Notwendigkeit, das Fachgebiet aus informatikdidaktischer Perspektive verstärkt zu betrachten, wurde die Bedeutung verschiedener Themen des Datenmanagements in Hinblick auf Gesellschaft, Alltag und Beruf expliziert. Dazu wurden insbesondere verschiedene Anforderungen, die die digitale Gesellschaft an ihre Bürgerinnen und Bürger stellt, beschrieben. Zusätzlich wurde auch die Bedeutung von Datenmanagement im beruflichen Umfeld und die in diesem stattfindenden Veränderungen, wie beispielsweise die Entstehung der neuen Berufsgruppe *Data Scientist*, herausgearbeitet.

- *Auswahl informatischer Phänomene und Kontexte:*
  Anhand der vorher bereits charakterisierten zentralen Themen des Fachgebiets Datenmanagement wurden exemplarische Alltagskontexte, in denen der Umgang mit Daten eine zentrale Rolle spielt, herausgearbeitet, wie beispielsweise das Leben im *Smart Home* oder die Nutzung des *öffenlichen Nahverkehrs*. Durch Phänomene, denen Schülerinnen und Schüler heute beim Umgang mit Daten begegnen, wurde eine Verknüpfung zum Alltag hergestellt.

- *Erfassung von Lehrerperspektiven:*
  Durch eine Befragung von Lehrkräften wurde gezeigt, dass bei diesen im Allgemeinen eine große Offenheit für und ein Interesse an diesem Themengebiet vorhanden ist. Obwohl die befragten Lehrkräfte das Themenfeld im Gesamten als eher span-

nend für den Unterricht einschätzen, fehlt ihnen jedoch das notwendige Wissen, um Datenmanagementthemen verstärkt im Unterricht zu thematisieren.

- *Erfassung von Schülerperspektiven:*
  Auch die Befragung von Schülerinnen und Schülern bestätigte die Relevanz des Themas. Dabei zeigte sich, dass die meisten Befragten bereits außerunterrichtliche Erfahrungen mit verschiedenen Themen des Datenmanagements gemacht und dabei gewisse Vorstellungen darüber entwickelt haben. Dies bestätigte sich auch später im Rahmen der Beobachtung einer Unterrichtssequenz. Trotz der augenscheinlich relativ hohen Bedeutung verschiedener der angesprochenen Themen im Leben der Schülerinnen und Schüler, zeigen sie jedoch relativ wenig fachlich fundiertes Wissen, wodurch die Notwendigkeit eines adäquaten Unterrichts bestätigt wird.

- *Fachliche Klärung (Teil II):*
  Um der Notwendigkeit nachzukommen, dieses Fachgebiet verstärkt zu betrachten, wurden in Anlehnung an bewährte Ansätze, wie die *Fundamentalen Ideen* der Informatik, in einem empirischen Ansatz *Schlüsselkonzepte des Datenmanagements* ermittelt. Das entstandene, an die *Great Principles of Computing* angelehnte, Modell charakterisiert Datenmanagement aus den vier Perspektiven *Praktiken* (z. B. *Datenerfassung*, *Modellierung*), *Kerntechnologien* (wie *Datenbanken* und *Datenstromsysteme*), *Entwurfsprinzipien* (u. a. *Integrität*, *Verfügbarkeit*) und *Mechanismen* (bspw. *Repräsentation*, *Replikation*). Diese Charakterisierung des Fachgebiets dient als Grundlage für die Vermittlung eines Verständnisses der zentralen Aspekte dieses Fachgebiets, für die Aufarbeitung verschiedener Themen für den Informatikunterricht, aber auch für die weitere didaktische Forschung in diesem Bereich. Zusätzlich wurde durch die fundierte Betrachtung von *Data Literacy* als Aspekt des Informatikunterrichts auch dieses Thema erstmalig aufgegriffen und der Weg für eine vertieftere Betrachtung geebnet, indem basierend auf den theoretischen Grundlagen dieser Arbeit ein für den Schulunterricht geeignetes und zu den GI-Empfehlungen für Bildungsstandards in Informatik kompatibles Data-Literacy-Kompetenzmodell entwickelt wurde.

- *Didaktische Strukturierung von Informatikunterricht:*
  Im Rahmen der didaktischen Strukturierung von Informatikunterricht wurde durch zwei Implementierungen von Werkzeugen die Umsetzbarkeit von zwei exemplarisch ausgewählten Datenmanagementthemen im Informatikunterricht demonstriert. Dabei konnten im Rahmen einer Unterrichtserprobung und diverser Lehrerfortbildungen erste Erfahrungen in diesem Bereich erzielt werden. Diese gingen in die Planung einer Unterrichtssequenz zum Thema Data Mining ein, die auf das zuvor entwickelte Data-Literacy-Kompetenzmodell aufsetzte und im zeitlichen Rahmen von drei Doppelstunden einen ersten Einblick in die Thematik der Datenanalyse und darauf aufbauenden Prognose geben konnte. Im Rahmen von zwei Erprobungen konnten im Gesamten positive Ergebnisse erzielt und insbesondere demonstriert werden, dass diese durchaus komplex erscheinenden Themen der Informatik auch geeignet sind, um im Schulunterricht wichtige Kompetenzen zu vermitteln.

Basierend auf den dieser Arbeit zugrundeliegenden Forschungsfragen können die zentralen Forschungsergebnisse damit wie folgt zusammengefasst werden:

RQ1) Welche Einflüsse haben die Entwicklungen der letzten Jahre im Bereich des Datenmanagements auf den Umgang mit und die Bedeutung von Daten in Informatik, Alltag und Beruf?
Datenmanagement beinhaltet eine Reihe von Themen, wie Metadaten, die eine hohe Bedeutung nicht nur in der Informatik, sondern auch in Alltag und Beruf haben. Diese Bezüge werden in Kapitel 5 dargestellt und anhand von Alltagsphänomenen, z. B. Konflikten bei der Synchronisation zwischen mehreren Geräten, expliziert. Um diese und ähnliche Phänomene verstehen und mit diesen umgehen zu können, erweisen sich grundlegende Kompetenzen im Bereich des Datenmanagements als unerlässlich. Aktuelle Entwicklungen zeigen, dass die Bedeutung von Themen aus dem Umfeld des Datenmanagements in allen Bereichen unserer Gesellschaft vermutlich weiter zunehmen wird und ein fundierter Umgang mit diesen essenziell ist.

RQ2) Welche Bedeutung haben das Fachgebiet Datenmanagement bzw. mit diesem in Zusammenhang stehende Themen bereits im Informatikunterricht?
Das Fachgebiet Datenmanagement ist im Informatikunterricht heute weitestgehend nur mit dem Thema *Datenbanken* vertreten, was sich einerseits bei einem explorativen Blick auf den aktuellen Unterricht anhand von Schulbüchern und Unterrichtsmaterialien (vgl. Abschnitt 4.3) zeigt, andererseits aber auch in einer Analyse internationaler Curricula und Bildungsstandards bestätigt werden konnte. Obwohl vermutet werden kann, dass der Informatikunterricht in verschiedenen Einzelfällen auch über diese Mindeststandards hinaus geht, ist es daher gerechtfertigt anzunehmen, dass die meisten Themen des Datenmanagements, mit Ausnahme tradierter Themen wie beispielsweise Datenbanken oder dem zum Teil noch oberflächlich thematisierten Datenschutz, im Allgemeinen allenfalls eine sehr geringe Bedeutung im Informatikunterricht haben. Eine abschließende Überprüfung dieser Ergebnisse, mehrere Jahre nach der eigentlichen Analyse, bestätigte diese, obwohl in den letzten Jahren erkannt werden kann, dass dem Thema *Daten* langsam ein höheres Gewicht im Informatikunterricht beigemessen zu werden scheint.

RQ3) Wie sind Vorwissen und Erfahrungen von Schülerinnen und Schülern zu dem Datenmanagement zugehörigen Themen ausgeprägt?
Obwohl verschiedene Themen des Datenmanagements durchaus eine Rolle im Leben der Schülerinnen und Schüler spielen und sie bereits mit diesen Erfahrungen gemacht und Vorstellungen dazu aufgebaut haben, deuten die Ergebnisse einer durchgeführten Fragebogenstudie auf eher gering ausgeprägtes und wenig fachlich fundiertes Vorwissen hin. Trotzdem offenbaren die Ergebnisse dieser Studie genau wie Beobachtungen im Unterricht, dass die Lernenden zu alltagsnahen oder gesellschaftlich diskutierten Themen bereits erste Vorstellungen aufgebaut haben. Die im Rahmen der beschriebenen Studie gewonnenen Ergebnisse müssen jedoch aufgrund der geringen und nicht repräsentativen Stichprobe reflektiert verwendet werden und können nur einen ersten Eindruck geben. Für eine vertiefte Untersuchung unter Einbeziehung von Lehrkräften, beispielsweise im Rahmen eines Design-Based-Research-Ansatzes,

wurde durch die fachliche Aufbereitung dieses Themengebiets im Rahmen der vorliegenden Arbeit wichtige Grundlagen gelegt.

RQ4) Welche Unterstützung benötigen Lehrkräfte bei der Integration von Aspekten des Datenmanagements bzw. der Data Literacy in ihren Unterricht?
Grundsätzlich messen die befragten Lehrkräfte zentralen Themen des Datenmanagements eine hohe Relevanz bei und vermuten, dass diese im Unterricht motivierend thematisiert werden können. Sie sehen dabei jedoch Schwierigkeiten, insbesondere aufgrund ihrer eigenen fehlenden Fachkenntnisse, die sie als kaum ausreichend einschätzen, um die Themen in ihren Unterricht zu integrieren. Um dieses Problem anzugehen, muss einerseits eine umfangreiche Schulung entsprechender Kenntnisse und Kompetenzen im Rahmen von Lehrkräftefortbildungsmaßnahmen stattfinden, aber auch die Unterstützung der Lehrkräfte durch gut ausgearbeitetes Unterrichtsmaterial, auf das sie ihren Unterricht aufbauen können. Ein weitere wichtige Herausforderung sind die fehlenden Werkzeuge: Während zwar verschiedenste Tools im Umfeld des Datenmanagements existieren, sind diese häufig zu komplex für den Einsatz im Unterricht. Entsprechend muss eine Neuentwicklung oder Adaption von Werkzeugen in Hinblick auf die angestrebte Zielgruppe stattfinden, wie in dieser Arbeit bereits exemplarisch geschehen.

RQ5) Welche sind die zentralen Konzepte und Praktiken des Fachgebiets Datenmanagement, insbesondere in Hinblick auf den Informatikunterricht an Sekundarschulen?
Basierend auf dem Modell der *Great Principles of Computing* wurde das Modell der Schlüsselkonzepte des Datenmanagements entwickelt, das einen umfassenden Blick auf das Fachgebiet erlaubt. Neben den Praktiken des Datenmanagements, die den kompletten Lebenszyklus von Daten von der *Erfassung* und *Bereinigung* über die *Modellierung*, *Implementierung*, *Analyse*, *Visualisierung* und *Evaluation* bis hin zu *Austausch*, *Archivierung* und *Löschung* beinhalten, stellt dieses die zentralen informatischen Konzepte, die dem Fachgebiet zugrundeliegen und für dessen Verständnis zentral sind, dar. Diese werden durch die Entwurfsprinzipien und Mechanismen des Datenmanagements repräsentiert. Die Entwurfsprinzipien (z. B. *Datenunabhängigkeit*, *Isolierung*, *Redundanz*) haben je nach zu erreichendem Ziel eine unterschiedliche Relevanz und sind unterschiedlich umzusetzen: Beispielsweise stellt *Redundanz* in relationalen Datenbanken üblicherweise eine zu vermeidende Eigenschaft dar, da redudante Datenspeicherung zu Inkonsistenzen führen kann, während verteilte Datenbanken *Redundanz* gezielt nutzen, um eine höhere Ausfallsicherheit zu erreichen. Die Entwurfsprinzipien erlauben es somit, verschiedene Anwendungen des Datenmanagements hinsichtlich der Erfüllung bzw. Relevanz der verschiedenen Entwurfsprinzipien zu charakterisieren. Die Mechanismen stellen hingegen keine Entwurfsentscheidungen dar, sondern charakterisieren die zentralen Aspekte der Funktionsweise von Datenmanagementsystemen, wie beispielsweise die *Synchronisation*, *Replikation* oder *Repräsentation* von Daten.

RQ6) Welche Struktur liegt den für einen kritischen und verantwortungsbewussten Umgang mit Daten im Sinne einer Data Literacy allgemein notwendigen Kompetenzen zugrunde?

Um die Konsequenzen der andauernd stattfindenden und immer umfassender werdenden Datenerfassung und -analyse einschätzen und von den sich dabei eröffnenden Möglichkeiten selbst profitieren zu können, muss heute Jeder grundlegende Kompetenzen im Umgang mit Daten erwerben. Diese Kompetenzen werden durch den Begriff der *Data Literacy* charakterisiert und im in dieser Arbeit entwickelten Data-Literacy-Kompetenzmodell charakterisiert, dass sie in eher fachlich orientierte Inhaltsbereiche und eher auf den Umgang mit Daten und die praktischen Tätigkeiten fokussierte Prozessbereiche aufgliedert. Eine zukünftige Einführung verschiedener Kompetenzstufen wird die Einsetzbarkeit des Modells weiter verbessern, gleichzeitig muss eine tiefergehende Evaluation mit fachlichen Experten und Lehrkräften noch erfolgen.

RQ7) Inwiefern ist es möglich, im Informatikunterricht grundlegende Data-Literacy-Kompetenzen herauszubilden?
Anhand von drei Beispielen wurde in Teil IV die Aufbereitung von Datenmanagementthemen für den Informatikunterricht an Sekundarschulen demonstriert. Dabei wurde gezeigt, dass es möglich ist, verschiedene Aspekte der Data Literacy und des Datenmanagements erfolgreich im Unterricht zu thematisieren. Die Schülerinnen und Schüler bekamen im Rahmen der Erprobungen erste Eindrücke der Echtzeitverarbeitung mithilfe von Datenstromsystemen und von der effizienten Prognose von Attributen auf Basis eines Data-Mining-Ansatzes. Sie konnten dabei einen Einblick in die Aussagekraft von Daten und Metadaten in den dabei durchgeführten Datenanalysen bekommen, aber auch die hohe Qualität und schnelle und einfache Durchführbarkeit solcher Analysen selbst erleben. Durch die Verknüpfung der gewählten Beispiele mit ihrem Alltag und durch die Diskussion gesellschaftlich relevanter Themen konnten sie die gewonnen Erkenntnisse direkt auf weitere Beispiele übertragen und somit die Bedeutung dieser Themen für ihr eigenes Leben erkennen. Im durchgeführten Unterricht konnten somit verschiedene Kompetenzen, beispielsweise hinsichtlich der Struktur und Durchführung solcher Analysen, aber auch in Hinblick auf die Beurteilung der Auswirkungen auf Mensch und Gesellschaft erworben werden. Basierend auf den im Rahmen der Erprobungen gewonnenen Erkenntnisse wurden Leitlinien für die Gestaltung von Unterricht im Bereich Datenmanagement entwickelt.

13.2 Ausblick und Fazit

In den einzelnen Kapiteln dieser Arbeit wurden bereits verschiedene Anknüpfungsmöglichkeiten für die zukünftige informatikdidaktische Forschung im Bereich des Datenmanagements und der Data Literacy benannt. Dabei erscheint insbesondere die praktische Anwendung und die damit einhergehende weitergehende Evaluation der Ergebnisse dieser Arbeit zentral: Während in Kapitel 11 bis 12 bereits erste praktische Erfahrungen beschrieben wurden, müssen diese durch eine größer angelegte Untersuchung validiert werden. Dabei kann insbesondere eine vertiefte Betrachtung der bisher im Rahmen der Schüler-

und Lehrerbefragungen erhaltenen Ergebnisse erfolgen, sodass im Sinne des Modells der didaktischen Rekonstruktion die gewonnenen Ergebnisse in einer weiteren Iteration über alle Bereiche des Modells in eine Weiterentwicklung einfließen können.

Auch der Bereich der *Data Literacy* ist ein vielversprechendes und zukünftig höchst relevantes Thema: Wie im Rahmen dieser Arbeit argumentiert, stellen Daten ein für den Alltag heute bereits essenzielles Thema dar, dessen Bedeutung zukünftig weiterhin anwachsen wird. Während im Kontext der tertiären Bildung insbesondere im letzten Jahr verstärkte Bemühungen erkennbar sind, entsprechende Kompetenzen im Grundlagenbereich aller Fächer zu verankern[69], stellen *Daten* im Bereich der Allgemeinbildung weiterhin ein eher nur nebenbei angesprochenes Thema des Informatikunterrichts dar. Das im Rahmen dieser Arbeit entwickelte Data-Literacy-Kompetenzmodell kann hier zu einer ersten Fundierung entsprechender Unterrichtsversuche beitragen und charakterisiert zusammen mit dem Modell der Schlüsselkonzepte des Datenmanagements gleichzeitig diesen wichtigen Themenbereich. Insbesondere zeichnet sich somit Data Literacy als vielversprechendes Forschungsgebiet ab, dass auf die in der vorliegenden Arbeit gewonnen Ergebnisse auf vielfältige Weise zurückgreifen kann. Dieses Thema wird daher unter anderem im GI-Präsidiumsarbeitskreis *Data Science und Data Literacy Education* aufgegriffen.

Nicht nur an der Data Literacy, sondern auch am Datenmanagement selbst, konnte in vielfältigen Diskussionen mit Kolleginnen und Kollegen der nationalen und internationalen informatikdidaktischen Fachcommunity aber auch mit Lehrerinnen und Lehrern ein zunehmendes Interesse festgestellt werden. Durch verschiedene parallel zur Forschung durchgeführte Lehrerfortbildungen konnte auch in die Unterrichtspraxis hineingewirkt werden: Immer häufiger zeigen Anfragen von Lehrerinnen und Lehrern ein umfassendes Interesse daran, basierend auf den im Rahmen dieser Arbeit entwickelten Unterrichtsideen, verschiedene Aspekte des Datenmanagements in ihren Informatikunterricht zu integrieren. Gleichzeitig wurde die Relevanz der Themen auch von anderen Akteuren im Bereich der Schulbildung erkannt, sodass diese beispielsweise durch die Medienpädagogik *(vgl. Tulodziecki, 2016)* aufgegriffen werden. Auch bei der Neugestaltung von Lehrplänen werden insbesondere Big Data und Data Mining meist diskutiert. Dies schafft ideale Bedingungen, um zukünftig eine tiefergehende Erforschung der Unterrichtspraxis in diesem Bereich zu ermöglichen, die in dieser Arbeit aufgrund der Notwendigkeit, das Thema zuerst von Grund auf aufzuarbeiten, nur eingeschränkt stattfinden konnte.

Ein weiteres spannendes Forschungsgebiet, dass einen gewissen Bezug zum Datenmanagement und noch stärker zur Data Literacy aufweist, stellen moderne Aspekte der Künstlichen Intelligenz bzw. insbesondere des Maschinenlernens dar. Maschinelles Lernen wird dabei heute zu einem Grundprinzip von immer mehr Anwendungen, die bereits seit langem Einzug in das Alltagsleben genommen haben und von den meisten Personen bereits selbstverständlich und unbemerkt verwendet werden. Dieses Fachgebiet steht vor einer ähnlichen Situation wie das Datenmanagement: Während Künstliche Intelligenz (nicht

[69] Beispielsweise wurden durch den Stifterverband Fördermittel zur *Vermittlung von Datenkompetenzen an Studierende aller Fächer* ausgeschrieben, auf die sich 47 Hochschulen mit entsprechenden Umsetzungskonzepten beworben haben.

nur) von Fachfremden häufig noch auf Anwendungen wie die bereits 1966 von Weizenbaum entwickelte ELIZA-KI reduziert betrachtet wird, konnten in den letzten Jahrzehnten enorme Fortschritte erzielt werden, die unseren Umgang mit Informatiksystemen und unser Leben in der heutigen digitalen Gesellschaft deutlich verändert haben. Entsprechend scheint es sinnvoll, dieses Fachgebiet auch, analog zum Datenmanagement, aus informatikdidaktischer Perspektive umfassend aufzubereiten. Hierzu kann die in dieser Arbeit genutzte Methodik eine mögliche Basis liefern.

Zusammenfassend stellt diese Arbeit einen Beitrag zur informatikdidaktischen Forschung in einem Bereich dar, der über mehr als zwei Jahrzehnte in der Forschung trotz zunehmender Bedeutung und umfassender Veränderungen stark vernachlässigt wurde, liefert aber gleichzeitig wichtige methodische Ansätze für die Aufbereitung anderer Themenbereiche nicht nur der Informatik. Durch diese Arbeit wurde Datenmanagement zu einem geeigneten Zeitpunkt erneut in den Mittelpunkt gerufen: Im Kontext der fortwährenden Digitalisierung jeglicher Lebensbereiche und der Herausforderungen, die mit der digitalen Gesellschaft einhergehen, stellt der fundierte und kritische Umgang mit Daten heute eine wichtige Schlüsselkompetenz dar. Um diese zu fördern, muss sich allgemeinbildender Schulunterricht verschiedener Themen des Datenmanagements annehmen und diese in aktuellen Kontexten aufgreifen, um Informatik und Informatiksysteme zu entmystifizieren.

Verzeichnisse

Literaturverzeichnis

Alle gegebenenfalls angegebenen Weblinks wurden zuletzt überprüft am 01.11.2018.

Acker, Amelia und Bowler, Leanne (2017). „What is Your Data Silhouette?: Raising Teen Awareness of Their Data Traces in Social Media“. In: *Proceedings of the 8th International Conference on Social Media & Society*. #SMSociety17. Toronto, ON, Canada: ACM, 26:1–26:5 (siehe S. 42).

ACRL Board (2016). *Framework for Information Literacy for Higher Education* (siehe S. 21, 22).

Ali-ud-din Khan, M., Fahim Uddin, Muhammad und Gupta, Navarun (Apr. 2014). „Seven V's of Big Data understanding Big Data to extract value“. In: *Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education*. IEEE Computer Society, S. 1–5 (siehe S. 27).

Antonitsch, Peter K. (2007). „Datenbanken – (etwas) anders gesehen“. 12. GI-Fachtagung „Informatik und Schule – INFOS 2007“. 19.–21. September 2007 an der Universität Siegen. In: *Didaktik der Informatik in Theorie und Praxis*. Hrsg. von S. Schubert. Bd. P. Lecture Notes in Informatics (LNI) 112. Bonn: Gesellschaft für Informatik e.V., S. 229–240 (siehe S. 5, 41).

Arbeitskreis Bildungsstandards (2008). „Grundsätze und Standards für die Informatik in der Schule: Bildungsstandards Informatik für die Sekundarstufe I“. In: *Beilage zu LOG IN* 28 (150/151) (siehe S. 4, 43, 47, 145).

Arbeitskreis Bildungsstandards SII (2016). „Bildungsstandards Informatik für die Sekundarstufe II“. In: *Beilage zu LOG IN* 36 (183/184) (siehe S. 43, 145).

Arbeitskreis der Technologietransferstellen niedersächsischer Hochschulen (2017). *Die Automatisierung der Gesellschaft*. `https://www.uni-hannover.de/fileadmin/Dezernat4/SG43/Publikationen/Technologie-Informationen/MWK_ti_1_2_2017_1703_Web.pdf` (siehe S. 60).

Armstrong, Deborah J. (Feb. 2006). „The Quarks of Object-oriented Development“. In: *Commun. ACM* 49 (2), S. 123–128 (siehe S. 100–103, 106).

Arnold, Norbert und Thomas Köhler, Hrsg. (2018). *Digitale Gesellschaft: Gestaltungsräume*. `http://www.kas.de/wf/doc/kas_51277-544-1-30.pdf?180112082252`. Konrad-Adenauer-Stiftung (siehe S. 60).

Atteslander, Peter (2010). *Methoden der empirischen Sozialforschung*. Berlin: Erich Schmidt Verlag (siehe S. 197).

Baumann, Rüdeger (1998). „Fundamentale Ideen der Informatik – gibt es das?“ In: *Informatische Bildung in Deutschland. Perspektiven für das 21. Jahrhundert*. Hrsg. von Bernhard Koerber und Ingo-Rüdiger Peters. LOG IN Verlag, S. 89–107 (siehe S. 138).

Bayerisches Staatsministerium für Unterricht und Kultus (2012). *Medienbildung. Medienerziehung und informationstechnische Bildung in der Schule*. `http://www.gesetze-bayern.de/Content/Document/BayVwV270223` (siehe S. 182).

Beauchamp, Nicholas (2017). „Predicting and Interpolating State-Level Polls Using Twitter Textual Data“. In: *American Journal of Political Science* 61 (2), S. 490–503 (siehe S. 170).

Beckel, Christian et al. (2014). „Revealing household characteristics from smart meter data“. In: *Energy* 78, S. 397–410 (siehe S. 62).

Behörde für Schule und Berufsbildung, Hamburg (2009). *Bildungsplan gymnasiale Oberstufe: Informatik* (siehe S. 47).

– (2011). *Bildungsplan Gymnasium Sekundarstufe I: Informatik Wahlpflichtfach* (siehe S. 47).

Bell, Tim, Tymann, Paul und Yehudai, Amiram (2011). *The Big Ideas of K-12 Computer Science Education*. `http://www.cosc.canterbury.ac.nz/research/RG/CSE/big-ideas/BigIdeas-webdocument-7-May-2011.pdf` (siehe S. 99, 103).

Berendt, Bettina et al. (2014). „Kostenlos ist nicht kostenfrei“. In: *LOG IN* 34 (178/179), S. 41–56 (siehe S. 169).

Bierschneider-Jakobs, Andrea (2004). „Datenmodellierung und Datenbanksysteme“. In: *LOG IN* 24 (127), S. 28–34 (siehe S. 5).

Borg, Bernhard (1987). „Didaktisch-methodische Aspekte des Einsatzes von Datenbanksystemen“. In: *LOG IN* 7 (5/6), S. 30–35 (siehe S. 4).

Borowski, Christian, Diethelm, Ira und Wilken, Henning (2016). „What Children Ask About Computers, the Internet, Robots, Mobiles, Games etc.“ In: *Proceedings of the 11th Workshop in Primary and Secondary Computing Education*. WiPSCE '16. New York: ACM, S. 72–75 (siehe S. 12).

Brewer, Eric (2012). „CAP Twelve Years Later: How the "Rules"Have Changed“. In: *Computer* 45 (2), S. 23–29 (siehe S. 28, 29).

Brichzin, Peter et al. (11. Sep. 2007). *Informatik I*. München: Oldenbourg Schulbuchverlag (siehe S. 4).

Brinda, Torsten (2004). „Didaktisches System für objektorientiertes Modellieren im Informatikunterricht der Sekundarstufe II“. Diss. Universität Siegen (siehe S. 43).

– (2016). *Stellungnahme zum KMK-Strategiepapier "Bildung in der digitalen Welt"*. `https://fb-iad.gi.de/fileadmin/stellungnahmen/gi-fbiad-stellungnahme-kmk-strategie-digitale-bildung.pdf`. Gesellschaft für Informatik e.V. (siehe S. 60, 61).

Budgen, David et al. (2008). „Using Mapping Studies in Software Engineering“. In: *PPIG 2008: Proceedings of the 20th Annual Meeting of the Pschology of Programming Interest Group*. Hrsg. von Jim Buckley, John Rooksby und Roman Bednarik. Lancaster: University of Lancaster (siehe S. 44).

Buffum, Philip Sheridan et al. (2014). „CS principles goes to middle school: learning how to teach "Big Data"“. In: *Proceedings of the 45th ACM technical symposium on Computer science education – SIGCSE'14*. New York: ACM, S. 151–156 (siehe S. 5, 42).

Bundesamt für Sicherheit in der Informationstechnik (2012). *Leitfaden Informationssicherheit: IT-Grundschutz kompakt*. `https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Grundschutz/Leitfaden/GS-Leitfaden_pdf.pdf` (siehe S. 134).

Bundesministerium für Bildung (2012). „Verordnung des Bundesministers für Unterricht und Kunst vom 14. November 1984 über die Lehrpläne der allgemeinbildenden höheren Schulen; Bekanntmachung der Lehrpläne für den Religionsunterricht an diesen Schulen (in der Fassung vom 24.10.2012)“. In: *Bundesgesetzblatt für die Republik Österreich* (siehe S. 47).

Bundesministerium für Familie, Senioren, Frauen und Jugend (2017). *Gutes Leben in der Digitalen Gesellschaft*. https://www.bmfsfj.de/blob/108988/aeec36ee21b4c6ac7fdea86c976e4128/gutes-familienleben-in-der-digitalen-gesellschaft-data.pdf (siehe S. 59).

Bussmann, Hans und Heymann, Hans Werner (1987). „Computer und Allgemeinbildung". In: *Neue Sammlung* 27 (1), S. 2–39 (siehe S. 141).

Buttke, Robby und Engelmann, Lutz (11. Mai 2007). *Informatik Bayern 9 Gymnasium*. Hrsg. von Lutz Engelmann. Berlin: DUDEN PAETEC. 136 S. (siehe S. 4).

Cao, Longbing (Juni 2017). „Data Science: A Comprehensive Overview". In: *ACM Comput. Surv.* 50 (3), 43:1–43:42 (siehe S. 20).

Carlson, Jake und Lisa R. Johnston, Hrsg. (2015). *Data-Information Literacy*. Purdue University Press (siehe S. 22).

Chang, Fay et al. (Juni 2008). „Bigtable: A Distributed Storage System for Structured Data". In: *ACM Trans. Comput. Syst.* 26 (2), 4:1–4:26 (siehe S. 29).

Chisholm, Malcolm (2015). *7 phases of a data life cycle*. https://www.bloomberg.com/professional/blog/7-phases-of-a-data-life-cycle/ (siehe S. 125, 127).

Codd, Edgar F. (1970). „A Relational Model of Data for Large Shared Data Banks". In: *Commun. ACM* 13 (6), S. 377–387 (siehe S. 25, 39).

Computing At School (2012). *Computer Science: A curriculum for schools*. https://www.computingatschool.org.uk/data/uploads/ComputingCurric.pdf (siehe S. 47).

Cortez, Paulo und Silva, Alice (2008). „Using Data Mining to Predict Secondary School Student Performance". In: *Proceedings of 5th Annual Future Business Technology Conference, Porto, 2008*. Hrsg. von Antonio Carvalho Brito und J. Manuel Feliz Teixeira. EUROSIS-ETI, S. 5–12 (siehe S. 192).

Costa, Arthur L. und Liebmann, Rosemarie M. (1996). *Envisioning Process as Content: Toward a Renaissance Curriculum*. Thousand Oaks, California: Corwin (siehe S. 98).

CSTA (2017). *K-12 Computer Science Standards, Revised 2017*. https://drive.google.com/file/d/0B0TlX1G3mywqbXpydGdIVk0OY1U/view (siehe S. 142).

CSTA Standards Taskforce (2016). *[Interim] CSTA K–12 Computer Science Standards*. https://cdn.ymaws.com/www.csteachers.org/resource/resmgr/Docs/Standards/2016StandardsRevision/INTERIM_StandardsFINAL_07222.pdf (siehe S. 43).

CSTA und ISTE (2011). *Computational Thinking Teacher Resources*. https://www.csteachers.org/resource/resmgr/472.11CTTeacherResources_2ed.pdf (siehe S. 142).

Cukier, Kenneth Neil und Mayer-Schönberger, Viktor (2017). „The rise of big data: How it's changing the way we think about the world's". In: *Foreign Affairs* 92, S. 28–40 (siehe S. 3).

DAMA International (2010). *The DAMA Guide to the Data Management Body of Knowledge: (DAMA-DMBOK Guide)*. Basking Ridge: Technics Publications (siehe S. 19, 48, 105, 107).

– (2017). *DAMA-DMBOK. Data Management Body of Knowledge*. Hrsg. von Deborah Henderson, Susan Earley und Laura Sebastian-Coleman. 2. Aufl. Basking Ridge: Technics Publications (siehe S. 35, 37, 125, 126).

Deahl, Erica (2016). „Better the Data You Know: Developing Youth Data Literacy in Schools and Informal Learning Environments“. Magisterarb. Massachusetts Institute of Technology (siehe S. 154).

Dean, Jeffrey und Ghemawat, Sanjay (Jan. 2008). „MapReduce: Simplified Data Processing on Large Clusters“. In: *Commun. ACM* 51 (1), S. 107–113 (siehe S. 27, 30).

Demchenko, Yuri, Belloum, Adam und Wiktorski, Tomasz (2017). *EDISON Data Science Framework: Part 2. Data Science Body of Knowledge (DS-BoK) Release 2*. EDISON (siehe S. 287).

Demchenko, Yuri, Manieri, Andrea und Belloum, Adam (2017). *EDISON Data Science Framework: Part 1. Data Science Competence Framework (CF-DS) Release 2*. EDISON (siehe S. 287).

Denning, Peter J. (2003a). „Computer Science“. In: *Encyclopedia of Computer Science*. Chichester, UK: John Wiley und Sons, S. 405–419 (siehe S. 133).

– (2003b). „Great Principles of Computing“. In: *Commun. ACM* 46 (11), S. 15–20 (siehe S. 7, 92, 93, 97, 98, 103, 113).

– (2004). „Great Principles in Computing Curricula“. In: *Proceedings of the 35th SIGCSE Technical Symposium on Computer Science Education*. SIGCSE '04. New York: ACM, S. 336–341 (siehe S. 102, 113, 139).

– (Apr. 2005). „Is Computer Science Science?“ In: *Commun. ACM* 48 (4), S. 27–31 (siehe S. 97).

Denning, Peter J. und Martell, Craig H. (2015). *Great Principles of Computing*. Cambridge, London: The MIT Press (siehe S. 113).

Diethelm, Ira (2007). „‚Strictly models and objects first' – Unterrichtskonzept und -methodik für objektorientierte Modellierung im Informatikunterricht“. Diss. Universität Kassel (siehe S. 43).

Diethelm, Ira, Borowski, Christian und Weber, Thomas (2010). „Identifying relevant CS contexts using the miracle question“. In: *Proceedings of the 10th Koli Calling International Conference on Computing Education Research*. New York: ACM, S. 74–75 (siehe S. 12).

Diethelm, Ira und Dörge, Christina (2011). „Zur Diskussion von Kontexten und Phänomenen in der Informatikdidaktik“. In: *Informatik in Bildung und Beruf – INFOS 2011 – 14. GI-Fachtagung Informatik und Schule*. Hrsg. von Marco Thomas. Bonn: Gesellschaft für Informatik e.V., S. 67–76 (siehe S. 63).

Diethelm, Ira, Dörge, Christina et al. (2011). „Die Didaktische Rekonstruktion für den Informatikunterricht“. In: *Informatik in Bildung und Beruf – 14. GI-Fachtagung Ïnformatik und Schule – INFOS 2011, 12.-15. September 2011 an der Westfälischen Wilhelms-Universität Münster*. Hrsg. von Marco Thomas. Bd. P. Lecture Notes in Informatics (LNI) 189. Bonn: Gesellschaft für Informatik e.V., S. 77–86 (siehe S. 10, 11, 13, 14, 63).

Diethelm, Ira, Hubwieser, Peter und Klaus, Robert (2012). „Students, Teachers and Phenomena: Educational Reconstruction for Computer Science Education“. In: *12th Koli Calling conference on computing education research*. Hrsg. von Robert McCartney und Mikko-Jussi Laakso. New York: ACM (siehe S. 12).

Diethelm, Ira, Wilken, Henning und Zumbrägel, Stefan (2012). „An investigation of secondary school students' conceptions on how the internet works“. In: *Proceedings of the 12th*

Koli Calling International Conference on Computing Education Research. New York: ACM, S. 67–73 (siehe S. 12).

Dietz, Alexander und Oppermann, Frank (2011). „Planspiel "Datenschutz 2.0"“. In: *Beilage zu LOG IN* 31.171 (siehe S. 41).

DIN 2342:2011-08 (2011). *Begriffe der Terminologielehre* (siehe S. 92).

Döring, Nicola und Bortz, Jürgen (2016). *Forschungsmethoden und Evaluation*. Berlin Heidelberg: Springer (siehe S. 106, 197).

Dorn, Julian (2017). *InstaHub - durch Datenbanken Möglichkeiten und Risiken in sozialen Netzwerken verstehen*. `https://blog.wi-wissen.de/post/instahub` (siehe S. 41).

Dorschel, Joachim (2015). *Praxishandbuch Big Data*. Wiesbaden: Springer Gabler (siehe S. 67, 189).

Dryer, Amber, Walia, Nicole und Chattopadhyay, Ankur (2018). „A Middle-School Module for Introducing Data-Mining, Big-Data, Ethics and Privacy Using RapidMiner and a Hollywood Theme“. In: *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*. SIGCSE '18. Baltimore, Maryland, USA: ACM, S. 753–758 (siehe S. 5).

Dwoskin, Elizabeth (2014). *In a Single Tweet, as Many Pieces of Metadata as There Are Characters*. `http://blogs.wsj.com/digits/2014/06/06/in-a-single-tweet-as-many-pieces-of-metadata-as-there-are-characters` (siehe S. 171).

Eckert, Claudia (2014). *IT-Sicherheit: Konzepte - Verfahren - Protokolle*. München: De Gruyter Oldenbourg (siehe S. 259).

Edlich, Stefan et al. (2011). *NoSQL*. 2. Aufl. München: Carl Hanser Verlag (siehe S. 27–29, 107, 266).

Elmasri, Ramez A. und Navathe, Shamkant B. (2009). *Grundlagen von Datenbanksystemen*. 3. aktualisierte Auflage. München: Pearson Deutschland GmbH (siehe S. 107, 257).

Engelmann, Lutz (2006). *Informatische Grundbildung 3*. DUDEN PAETEC (siehe S. 40).

Ester, Martin und Sander, Jörg (2000). *Knowledge Discovery in Databases*. Berlin Heidelberg: Springer-Verlag (siehe S. 168).

Europäische Union (2016). „Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung)“. In: *Amtsblatt der Europäischen Union* 56 (L 119) (siehe S. 23).

Exner, Martin (2009). „Die Entdeckung der Cholera-Ätiologie durch Robert Koch 1883/84.“ In: *HygMed* 34 (4), 144ff (siehe S. 25).

FDP (2016). *Chancen der digitalen Gesellschaft. Beschluss des 67. Ord. Bundesparteitags der FDP, Berlin, 3.-24. April 2016*. `https://www.fdp.de/sites/default/files/filefield_paths/2016_04_24_bpt_chancen_der_digitalen_gesellschaft.pdf` (siehe S. 60).

Fischer, Helmar, Knapp, Thomas und Neupert, Heiko (2006). *Grundlagen der Informatik II*. Oldenbourg (siehe S. 41).

Fischer, Stephan (2014). „Big Data: Herausforderungen und Potenziale für deutsche Softwareunternehmen“. In: *Informatik-Spektrum* 37 (2), S. 112–119 (siehe S. 26).

Freischlad, Stefan (2009). „Entwicklung und Erprobung des didaktischen Systems Internetworking im Informatikunterricht“. ger. Diss. Universität Siegen (siehe S. 196).

Gal-Ezer, Judith und Harel, David (1999). *Curriculum and Course Syllabi for a High-School Program in Computer Science*. https://www.openu.ac.il/personal_sites/download/galezer/curr_and_syll.pdf. Jerusalem, Israel, Israel (siehe S. 47).

Gantz, John und Reinsel, David (2012). *THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East* (siehe S. 36).

Gesellschaft für Fachdidaktik (2015). *Formate Fachdidaktischer Forschung. Definition und Reflexion des Begriffs*. Diskussionspapier der GFD 2015. http://www.fachdidaktik.org/wp-content/uploads/2015/09/GFD-Positionspapier-18-Formate-Fachdidaktischer-Forschung.pdf (siehe S. 9).

Gillmor, D. (2014). *As we sweat government surveillance, companies like Google collect our Data*. https://www.theguardian.com/commentisfree/2014/apr/18/corporations-google-should-not-sell-customer-data (siehe S. 4).

Ginsberg, Jeremy et al. (Feb. 2009). „Detecting influenza epidemics using search engine query data". In: *Nature* 457 (7232), S. 1012–1014 (siehe S. 32).

Golab, Lukasz und Özsu, M. Tamer (Juni 2003). „Issues in Data Stream Management". In: *SIGMOD Rec.* 32 (2), S. 5–14 (siehe S. 34).

Gräsel, Cornelia und Parchmann, Ilka (2004). „Implementationsforschung – oder: der steinige Weg, Unterricht zu verändern". In: *Unterrichtswissenschaft* 32 (3), S. 196–214 (siehe S. 9).

Grillenberger, Andreas, Przybylla, Mareen und Romeike, Ralf (2016). „Bringing CS Innovations to the Classroom: a Process Model of Educational Reconstruction". In: *International Conference on Informatics in Schools. ISSEP 2016. Proceedings*. Hrsg. von Andrej Brodnik und Françoise Tort, S. 31–39 (siehe S. vii).

Grillenberger, Andreas und Romeike, Ralf (2014a). „A Comparison of the Field Data Management and its Representation in Secondary CS Curricula". In: *Proceedings of the 9th Workshop in Primary and Secondary Computing Education*. Hrsg. von Carsten Schulte, Michael E. Caspersen und Judith Gal-Ezer. New York: ACM, S. 29–36 (siehe S. vii).

– (2014b). „Teaching Data Management: Key Competencies and Opportunities". In: *KEYCIT 2014 – Key Competencies in Informatics and ICT*. Hrsg. von Torsten Brinda, Nicholas Reynolds und Ralf Romeike. Potsdam: Universitätsverlag Potsdam, S. 133–150 (siehe S. vii).

– (2015a). „Analyzing the Twitter Data Stream Using the Snap! Learning Environment". In: *International Conference on Informatics in Schools: Situation, Evolution, and Perspectives*. Hrsg. von Andrej Brodnik und Jan Vahrenhold. Springer, Cham. Cham: Springer, S. 155–164 (siehe S. vii).

– (2015b). „Big Data im Informatikunterricht: Motivation und Umsetzung". In: *Informatik allgemeinbildend begreifen*. Hrsg. von Jens Gallenbacher. Bd. P. Lecture Notes in Informatics (LNI) 249. Bonn: Gesellschaft für Informatik e.V., S. 125–134 (siehe S. vii).

– (2015c). „Big-Data-Analyse im Informatikunterricht mit Datenstromsystemen: Ein Unterrichtsbeispiel". In: *Informatik allgemeinbildend begreifen*. Hrsg. von Jens Gallenbacher. Bd. P. Lecture Notes in Informatics (LNI) 249. Bonn: Gesellschaft für Informatik e.V., S. 135–144 (siehe S. vii).

– (2015d). „Bringing the Innovations in Data Management to CS Education: An Educational Reconstruction Approach“. In: *Proceedings of the Workshop in Primary and Secondary Computing Education*. WiPSCE '15. New York: ACM, S. 88–91 (siehe S. vii).
– (2017a). „Empirische Ermittlung der Schlüsselkonzepte des Fachgebiets Datenmanagement“. In: *Informatische Bildung zum Verstehen und Gestalten der digitalen Welt*. Hrsg. von Ira Diethelm. Bd. P. Lecture Notes in Informatics (LNI) 274. Bonn: Gesellschaft für Informatik e.V., S. 157–166 (siehe S. vii).
– (2017b). „Key Concepts of Data Management: An Empirical Approach“. In: *Proceedings of the 17th Koli Calling International Conference on Computing Education Research*. Koli Calling '17. New York: ACM, S. 30–39 (siehe S. vii).
– (2017c). „Real-Time Data Analyses in Secondary Schools Using a Block-Based Programming Language“. In: *International Conference on Informatics in Schools: Situation, Evolution, and Perspectives*. Hrsg. von Valentina Dagienė und Arto Hellas. Springer, Cham. Cham: Springer, S. 207–218 (siehe S. vii).
– (2017d). „What Teachers and Students Know about Data Management“. In: *Tomorrow's learning: Involving everyone – Learning with and about technologies and computing* (Dublin). Hrsg. von Arthur Tatnall und Mary Webb. Bd. 515. IFIP AICT. Heidelberg: Springer, S. 557–566 (siehe S. vii).
– (2018a). „Datenmanagement als Thema für den Informatikunterricht“. In: *LOG IN* 37 (187/188), S. 44–52 (siehe S. vii).
– (2018b). „Developing a Theoretically Founded Data Literacy Competency Model“. In: *Proceedings of the 13th Workshop in Primary and Secondary Computing Education*. ACM, NY, USA (siehe S. vii).
– (2018c). „Was ist Data Science? Ermittlung der informatischen Inhalte durch Analyse von Studienangeboten“. In: *Hochschuldidaktik der Informatik HDI 2018*. Hrsg. von Nadine Bergner et al. Commentarii informaticae didacticae (CID) 12. Universitätsverlag Potsdam, S. 119–134 (siehe S. vii, 145, 287).

Gropengießer, Harald (1997). *Didaktische Rekonstruktion des SSehens". Wissenschaftliche Theorien und die Sicht er Schüler in der Perspektive der Vermittlung*. Beiträge zur Didaktischen Rekonstruktion. Oldenburg: Carl-von-Ossietzky-Universität Oldenburg, Didaktisches Zentrum (siehe S. 10).

Hammer, Volker und Prodesch, Ulrich (1987). *Planspiel Datenschutz in vernetzten Informationssystemen*. Die Schulpraxis (siehe S. 41).

Hartmann, Werner, Näf, Michael und Reichert, Raimond (2006). *Informatikunterricht planen und durchführen*. Berlin Heidelberg: Springer-Verlag (siehe S. 91).

Heidrich, Jens, Bauer, Pascal und Krupka, Daniel (2018). *Studie zu übergreifenden Kompetenzen und Studieninhalten in der digitalen Welt am Beispiel von Data Literacy*. IESE-Report 014.17/D. Fraunhofer IESE (siehe S. 21).

Heinemann, Birte et al. (2018). „Drafting a Data Science Curriculum for Secondary Schools“. In: *Proceedings of the 18th Koli Calling International Conference on Computing Education Research*. Koli Calling '18. Koli, Finland: ACM, 17:1–17:5 (siehe S. 42).

Hessisches Kultusministerium (2010). *Lehrplan Informatik, Gymnasialer Bildungsgang, Gymnasiale Oberstufe* (siehe S. 47).

Hey, Tony, Tansley, Stewart und Tolle, Kristin (Okt. 2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, Washington: Microsoft Research (siehe S. 21, 142).
Hoadley, Christopher und Favaro, Sharon (2015). „Digital Literacy in Higher Education". In: Hrsg. von J. Michael Spector. SAGE Publications (siehe S. 21, 22).
Hochschulforum Digitalisierung (2017). *Ausschreibung: „Übergreifende Kompetenzen und Studieninhalte in der digitalen Welt am Beispiel von Data Literacy"*. `https://hochschulforumdigitalisierung.de/de/news/ausschreibung-data-literacy` (siehe S. 21, 22, 144).
Hörsch, C. (2007). *Biologie verstehen: Mikroorganismen und mikrobielle Prozesse im Menschen*. Beiträge zur didaktischen Rekonstruktion. Oldenburg: Didaktisches Zentrum, Carl-von-Ossietzky-Univ. (siehe S. 10).
Hubwieser, Peter (2007). *Didaktik der Informatik*. Berlin Heidelberg: Springer (siehe S. 4).
Käberich, Günther und Steigerwald, Friedhelm (1986). *Schüler arbeiten mit einer Datenbank*. Metzler/Teubner (siehe S. 39).
Kantardzic, Mehmed (2011). *Data Mining*. 2. Aufl. Hoboken, New Jersey: John Wiley & Sons (siehe S. 32).
Kastl, Petra und Romeike, Ralf (2015). „"Now They Just Start Working, and Organize Themselves"First Results of Introducing Agile Practices in Lessons". In: *Proceedings of the Workshop in Primary and Secondary Computing Education*. WiPSCE '15. New York: ACM, S. 25–28 (siehe S. 43).
Kattmann, Ulrich (2007). *Theorien in der biologiedidaktischen Forschung. Didaktische Rekonstruktion – eine praktische Theorie*. Hrsg. von Dirk Krüger und Helmut Vogt. Berlin Heidelberg: Springer (siehe S. 10).
Kattmann, Ulrich et al. (1997). „Das Modell der didaktischen Rekonstruktion – Ein Rahmen für naturwissenschaftsdidaktische Forschung und Entwicklung". In: *Zeitschrift für Didaktik der Naturwissenschaften* 3 (3), S. 3–18 (siehe S. 9, 10, 12, 13).
Kemper, Alfons und Eickler, André (2013). *Datenbanksysteme: Eine Einführung*. München: Oldenbourg (siehe S. 107).
– (2015). *Datenbanksysteme: Eine Einführung*. Berlin/Boston: De Gruyter Oldenbourg. 870 S. (siehe S. 25, 26, 28, 262, 264).
Klein, Dominik, Tran-Gia, Phuoc und Hartmann, Matthias (1. Juni 2013). „Big Data". In: *Informatik-Spektrum* 36 (3), S. 319–323 (siehe S. 26).
Koch, Robert (1893). „Wasserfiltration und Cholera". In: *Zeitschrift für Hygiene und Infektionskrankheiten* 14 (1), S. 393–426 (siehe S. 25).
Kohl, Lutz (2009). „Kompetenzorientierter Informatikunterricht in der Sekundarstufe I unter Verwendung der visuellen Programmiersprache Puck". Diss. Friedrich-Schiller-Universität Jena (siehe S. 196).
Komorek, Michael und Prediger, Susanne (2013). *Der lange Weg zum Unterrichtsdesign*. Fachdidaktische Forschungen. Münster: Waxmann Verlag (siehe S. 12).
Kraynova, Aleksandra (2012). *Didaktische Rekonstruktion der Nanophysik : analytische und empirische Untersuchungen in einem interdisziplinären Forschungsfeld*. Beiträge zur Didaktischen Rekonstruktion. Oldenburg: Carl-von-Ossietzky-Universität Oldenburg, Didaktisches Zentrum (siehe S. 10).

Krikorian, Raffi (2013). *New Tweets per second record, and how!* https://blog.twitter.com/node/2845 (siehe S. 170).
Kudraß, Thomas, Hrsg. (2015). *Taschenbuch Datenbanken*. München: Carl Hanser Verlag (siehe S. 107).
Kultusministerkonferenz (2004). *Einheitliche Prüfungsanforderungen Informatik*. Kultusministerkonferenz (siehe S. 47).
– (2016). *Bildung in der digitalen Welt. Strategie der Kultusministerkonferenz.* https://www.kmk.org/fileadmin/Dateien/pdf/PresseUndAktuelles/2016/Bildung_digitale_Welt_Webversion.pdf (siehe S. 60).
Laney, Douglas (Feb. 2001). *3D Data Management: Controlling Data Volume, Velocity, and Variety*. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (siehe S. 27, 128).
Lischka, Konrad (2011). *Ermittler saugen 40.000 persönliche Datensätze*. http://www.spiegel.de/netzwelt/netzpolitik/handy-ueberwachung-in-dresden-ermittler-saugen-40-000-persoenliche-datensaetze-a-776465.html (siehe S. 61).
Lockemann, Peter C. (1986). „Konsistenz, Konkurrenz, Persistenz – Grundbegriffe der Informatik? – Zur Diskussion gestellt". In: *Informatik Spektrum* 9 (5), S. 300–305 (siehe S. 136, 269).
Lück, Willi van (1990). „Datenbanken in Schule und Unterricht". In: *LOG IN* 10 (6), S. 61–66 (siehe S. 4).
Maury, Matthew Fontaine (1855). *The Physical Geography of the Sea*. New York: Harper & Brothers (siehe S. 24).
Mayer-Schönberger, Viktor und Cukier, Kenneth Neil (2013). *Big Data – Die Revolution, die unser Leben verändern wird*. München: FinanzBuch Verlag (siehe S. 24, 32, 161).
Mayring, Philipp (2010). *Qualitative Inhaltsanalyse: Grundlagen und Techniken*. Weinheim: Beltz (siehe S. 45, 106, 288).
McBurney, Vincent (2013). *The Origin and Growth of Big Data Buzz*. http://it.toolbox.com/blogs/infosphere/the-origin-and-growth-of-big-data-buzz-51509 (siehe S. 26).
Michaeli, Tilman und Romeike, Ralf (2017). „Addressing Teaching Practices Regarding Software Quality: Testing and Debugging in the Classroom". In: *Proceedings of the 12th Workshop on Primary and Secondary Computing Education*. WiPSCE '17. New York: ACM, S. 105–106 (siehe S. 43).
Ministerium für Bildung, Wissenschaft, Jugend und Kultur RLP (2011a). *Lehrplan Informatik: Grund- und Leistungsfach. Einführungsphase und Qualifikationsphase der gymnasialen Oberstufe (Mainzer Studienstufe)* (siehe S. 47).
– (2011b). *Lehrplan Informatik: Wahlfach und Wahlpflichtfach an Gymnasien und Integrierten Gesamtschulen (Sekundarstufe I)* (siehe S. 47).
Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen (2013). *Kernlehrplan für die Sekundarstufe II Gymnasium / Gesamtschule in Nordrhein-Westfalen. Informatik.* (Siehe S. 47).

Mishra, Punya und Koehler, Matthew J. (2006). „Technological Pedagogical Content Knowledge: A new framework for teacher knowledge“. In: *Teachers College Record* 108 (6), S. 1017–1054 (siehe S. 69, 70).

Modrow, Eckart (1996). *Dateien, Datenbanken, Datenschutz*. Dümmler (siehe S. 40).

– (2003). „Pragmatischer Konstruktivismus und fundamentale Ideen als Leitlinien der Curriculumentwicklung am Beispiel der theoretischen und technischen Informatik“. Diss. Mathematisch-Naturwissenschaftlich-Technischen Fakultät der Martin-Luther-Universität Halle-Wittenberg (siehe S. 95, 97, 103, 138).

Modrow, Eckart und Strecker, Kerstin (2016). *Didaktik der Informatik*. Berlin/Boston: de Gruyter (siehe S. 97, 99, 103, 138).

NIST Big Data Public Working Group (2015). *NIST Big Data Interoperability Framework: Volume 1, Definitions*. National Institute of Standards and Technology (siehe S. 20).

Ogden, Charles Kay und Richards, Ivor Armstrong (1923). *The meaning of meaning: A study of the influence of thought and of the science of symbolism*. New York: Harcourt, Brace & World, Inc. (siehe S. 92, 93).

Omnicore Agency (2018). *Twitter by the Numbers: Stats, Demographics & Fun Facts*. `https://www.omnicoreagency.com/twitter-statistics` (siehe S. 170).

Ontario Ministry of Education (2007). *The Ontario Curriculum Grades 1-8: Science and Technology* (siehe S. 47, 55).

Open Knowledge International (2017). *The Open Definition*. `http://opendefinition.org` (siehe S. 4).

Österreichisches Institut für angewandte Telekommunikation (2015). *Studie: Dynamic Pricing – Die Individualisierung von Preisen im e-Commerce*. `https://www.guetezeichen.at/fileadmin/daten/Blog/DynamicPricing_OIAT.pdf` (siehe S. 175).

Palmer, Shelly (2015). *Data Science in the C-Suite*. New York, NY, USA: Digital Living Press (siehe S. 20).

Penon, Johann (2013). *VideoCenter*. `http://dokumentation.videocenter.schule.de` (siehe S. 41, 167).

– (2017). *FitnessCenter*. `http://fitnesscenter.schule.de` (siehe S. 41, 167).

Pentzold, Christian und Fischer, Charlotte (2017). „Framing Big Data: The discursive construction of a radio cell query in Germany“. In: *Big Data & Society* 4 (2), S. 2053951717745897. eprint: `https://doi.org/10.1177/2053951717745897` (siehe S. 61).

Piepmeyer, Lothar (2011). *Grundkurs Datenbanksysteme*. München: Carl Hanser Verlag (siehe S. 107).

Principles and Standards for School Mathematics (2000). National Council of Teachers of Mathematics (siehe S. 146).

Przybylla, Mareen (2018). „From Embedded Systems to Physical Computing: Challenges of the "Digital Worldïn Secondary Computer Science Education“. Diss. Universität Potsdam (siehe S. 43).

Qin, Jian und D'ignazio, John (2010). „The Central Role of Metadata in a Science Data Literacy Course“. In: *Journal of Library Metadata* 10.2-3, S. 188–204. eprint: `https://doi.org/10.1080/19386389.2010.506379` (siehe S. 22).

RatSWD (2018). „Forschungsdatenmanagement in den Sozial-, Verhaltens- und Wirtschaftswissenschaften – Orientierungshilfen für die Beantragungund Begutachtung datengenerierender und datennutzender Forschungsprojekte“. In: *RatSWD Output* 3.5 (siehe S. 19).

Rechenberg, Peter und Gustav Pomberger, Hrsg. (2002). *Informatik-Handbuch*. München: Hanser Fachbuch (siehe S. 61).

Reinsel, David, Gantz, John und Rydning, John (2017). *Data Age 2025: The Evolution of Data to Life-Critical* (siehe S. 24, 36, 37).

Resnick, Mitchel und Silverman, Brian (2005). „Some Reflections on Designing Construction Kits for Kids“. In: *Proceedings of the 2005 Conference on Interaction Design and Children*. IDC '05. New York: ACM, S. 117–122 (siehe S. 167).

Ridgway, Rosie, Nicholson, James und Gal, Iddo (2018). „Understanding statistics about society: A framework of knowledge and skills needed to engage with Civic Statistics“. In: *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018)*. Hrsg. von M. A. Sorto, A. White und L. Guyot. Voorburg, The Netherlands: International Statistical Institute (siehe S. 42).

Ridsdale, Chantel et al. (2015). *Strategies and Best Practices for Data Literacy Education: Knowledge Synthesis Report*. Dalhousie University (siehe S. 21, 22, 142–144, 157).

Riley, Jenn (2017). *Understanding Metadata: What is Metadata, and What is it For?: A Primer*. Baltimore: NISO (siehe S. 31).

Rubin, Andee (2005). „Math that matters: The case for replacing the algebra/calculus track with data literacy—a critical skill for modern life“. In: *Threshold Magazine: Exploring the Threshold of Education* (Spring 2005 issue), S. 22–25 (siehe S. 3).

Runkler, Thomas A. (2015). *Data Mining: Modelle und Algorithmen intelligenter Datenanalyse*. Wiesbaden: Springer Vieweg (siehe S. 125, 126).

Rutke, Ulrike (2007). „Schülervorstellungen und wissenschaftliche Vorstellungen zur Entstehung und Entwicklung des menschlichen Lebens“. Diss. Ludwig-Maximilians-Universität München (siehe S. 10).

Schield, Milo (2018). „Information Literacy, Statistical Literacy and Data Literacy“. In: *IASSIST QUARTERLY (IQ)* (siehe S. 21).

Schubert, Sigrid und Schwill, Andreas (2011). *Didaktik der Informatik*. Heidelberg: Spektrum Akademischer Verlag (siehe S. 92).

Schuh, Bernhard et al. (2002). *Grundzüge der Informatik III*. Manz (siehe S. 40).

Schulte, Carsten (2003). *Lehr-, Lernprozesse im Informatik-Anfangsunterricht : theoriegeleitete Entwicklung und Evaluation eines Unterrichtskonzepts zur Objektorientierung in der Sekundarstufe II [Elektronische Ressource]*. Paderborn, Univ., Diss., 2004 (siehe S. 43).

Schwanewedel, Julia (2010). „Biologie verstehen: Gene und Gesundheit“. Diss. Carl-von-Ossietzky-Universität Oldenburg (siehe S. 10).

Schwill, Andreas (1993). „Fundamentale Ideen der Informatik“. In: *Zentralblatt für Didaktik der Mathematik* 25 (1), S. 20–31 (siehe S. 7, 92, 94, 96, 103, 138, 139, 191).

– (1995). „Programmierstile im Anfangsunterricht“. In: *Innovative Konzepte für die Ausbildung: 6. GI-Fachtagung Informatik und Schule, INFOS '95, Chemnitz, 25.–28. September 1995*. Hrsg. von Sigrid Schubert. Berlin Heidelberg: Springer, S. 178–187 (siehe S. 43).

Schwill, Andreas (1998). *Fundamentale Ideen der Informatik und Modellierung im Informatikunterricht*. http://www.informatikdidaktik.de/didaktik/Forschung/Vortragsfolien FundIdeenMNU.pdf (siehe S. 94–96).
– (2004). „Philosophical aspects of fundamental ideas: Ideas and concepts. Concepts of Empirical Research and Standardisation of Measurement in the Area of Didactics of Informatics“. In: *Informatics and Student Assessment*. Bd. S. Lecture Notes in Informatics (LNI) 1. Bonn: Gesellschaft für Gesellschaft für Informatik e.V., S. 145–157 (siehe S. 92).
Seehorn, Deborah et al. (2011). *K–12 Computer Science Standards* (siehe S. 47).
Shannon, Claude E. (1948). „A mathematical theory of communication“. In: *The Bell System Technical Journal* 27.3, S. 379–423 (siehe S. 149).
Sharma, Sashi (2017). „Definitions and models of statistical literacy: a literature review“. In: *Open Review of Educational Research* 4.1, S. 118–133. eprint: https://doi.org/10.1080/23265507.2017.1354313 (siehe S. 42).
Shaw, Mary (1992). „We can teach software better“. In: *Computing Research News* 4 (4), S. 2–12 (siehe S. 91).
Sommer, Sarah (2013). *Warum Amazon weiß, was Ihre Frau mag*. http://www.manager-magazin.de/unternehmen/handel/big-data-analyse-im-online-handel-a-935555.html (siehe S. 26).
Staatsinstitut für Schulqualität und Bildungsforschung (2008). *Lehrplan für die Realschule in Bayern, Fach Informationstechnologie*. Staatsinstitut für Schulqualität und Bildungsforschung (siehe S. 83).
– (2009). *Lehrplan des achtjährigen Gymnasiums in Bayern*. Staatsinstitut für Schulqualität und Bildungsforschung (siehe S. 47, 57, 82).
Stoffers, Ana-Maria (2016). „Subjektive Theorien von Informatiklehrkräften zur fachdidaktischen Strukturierung ihres Unterrichts“. Diss. Carl-von-Ossietzky-Universität Oldenburg (siehe S. 12).
Strava (2017). *Strava Global Heatmap*. https://www.strava.com/heatmap (siehe S. 31).
Strozzi, Carlo (o.D.). *NoSQL: a non-SQL RDBMS*. http://www.strozzi.it (siehe S. 27).
The Economist (2017). *The world's most valuable resource is no longer oil, but data*. https://www.economist.com/news/leaders/21721656-data-economy-demands-new-approach-antitrust-rules-worlds-most-valuable-resource (siehe S. 3).
The Verge (2018). *Strava's fitness tracker heat map reveals the location of military bases*. https://www.theverge.com/2018/1/28/16942626/strava-fitness-tracker-heat-map-military-base-internet-of-things-geolocation (siehe S. 32).
Thirani, Vasudha und Gupta, Arvind (2017). *The Value of Data*. https://www.weforum.org/agenda/2017/09/the-value-of-data/ (siehe S. 66).
Tulodziecki, Gerhard (2016). „Konkurrenz oder Kooperation? Zur Entwicklung des Verhältnisses von Medienbildung und informatischer Bildung“. In: *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung* 25 (0), S. 7–25 (siehe S. 226).
Unland, Rainer und Pernul, Günther (2015). *Datenbanken im Einsatz: Analyse, Modellbildung und Umsetzung*. De Gruyter Studium. Berlin/München/Boston: De Gruyter Oldenbourg (siehe S. 107).

Valdivia, A. et al. (2010). „Monitoring influenza activity in Europe with Google Flu Trends: comparison with the findings of sentinel physician networks-results for 2009-10“. In: *Euro Surveill* 15 (29), S. 19621 (siehe S. 32).

Weinert, Franz Emanuel (2001). „Vergleichende Leistungsmessung in Schulen-eine umstrittene Selbstverständlichkeit“. In: *Leistungsmessungen in schulen*. Beltz, S. 17–32 (siehe S. 142).

Weintrop, David et al. (2016). „Defining computational thinking for mathematics and science classrooms“. In: *Journal of Science Education and Technology* 25.1 (1), S. 127–147 (siehe S. 142).

Wing, Jeannette M (2006). „Computational thinking“. In: *Communications of the ACM* 49 (3), S. 33–35 (siehe S. 142).

Wissenschaftliche Dienste des Deutschen Bundestags (2017). *Digitalisierung und Entwicklungspolitik*. `https://www.bundestag.de/blob/525938/488ea79620fb0b4c452b42519f2afb37/wd-2-051-17-pdf-data.pdf` (siehe S. 59).

Witten, Helmut (1994). „Datenbanken – (k)ein Thema im Informatikunterricht?“ In: *LOG IN* 14 (2), S. 14–19 (siehe S. 4, 39).

Wolff, Annika et al. (2017). „Creating an understanding of data literacy for a data-driven society“. In: *Journal of Community Informatics* 12.3 (3), S. 9–26 (siehe S. 156, 157).

Zendler, Andreas und Spannagel, Christian (2006). „Zentrale Konzepte im Informatikunterricht: eine empirische Grundlegung“. In: *Notes on Educational Informatics—Section A: Concepts and Techniques* 2 (1), S. 1–21 (siehe S. 92, 98, 99, 101, 103, 106).

Zendler, Andreas, Spannagel, Christian und Klaudt, Dieter (2007). „Zentrale Prozesse im Informatikunterricht: eine empirische Grundlegung“. In: *Notes on Educational Informatics – Section A: Concepts and Techniques* 3 (1), S. 1–19 (siehe S. 98, 99, 103, 106).

Abbildungsverzeichnis

Tabellenverzeichnis

Anhang

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

DEPARTMENT
INFORMATIK
DIDAKTIK DER INFORMATIK

Fragebogen zur Lehrerperspektive auf das Thema „Datenmanagement“

Im Rahmen des Forschungsprojekts „Datenmanagement im Informatikunterricht“ an der FAU Erlangen-Nürnberg, erfassen wir u.a. die Perspektive von Lehrerinnen und Lehrern auf dieses Thema. Dazu würden wir Sie bitten, uns durch Beantwortung des folgenden Fragebogens zu unterstützen. Falls Ihnen eine Einschätzung nicht möglich ist, lassen Sie die entsprechende Frage bitte unbeantwortet.

	Wie schätzen Sie ihr eigenes Wissen zu den Themen ein?				Wie interessant finden Sie diese für den Informatikunterricht?				Wo sehen Sie Schwierigkeiten bei der Umsetzung im Unterricht?		
	unbekannt	kaum Wissen	grundlegendes Wissen	detailliertes Wissen	nicht interessant	kaum interessant	eher interessant	sehr interessant	fehlendes eigenes Fachwissen	fehlende Werkzeuge	Thema ist zu komplex
(klassische) Datenbanken	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
NoSQL / non-relationale Datenbanken	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Verteilte Datenbanken	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Cloud-Speicher	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Cloud-Computing	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Datenanalyse (klassisch)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Data Mining	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Big Data	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Open Data	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Verschlüsselung von Daten	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Datenmodellierung	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Funktionsweise von Suchmaschinen	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
CAP-Theorem	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
ACID-Prinzip	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
BASE-Prinzip	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Metadaten	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Datensicherheit (z. B. Backup)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Datenschutz	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Gefahren bei der bzw. durch die maschinelle Verarbeitung von Daten	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Anhang B: Schülerfragebogen

Friedrich-Alexander-Universität Erlangen-Nürnberg
Professur für Didaktik der Informatik

Was wisst ihr zum Thema „Daten“?

Du erstellst mit deinem Smartphone ein Foto. Welche Informationen werden dem Bild automatisch mitgeliefert (d.h. sind in der Bilddatei zusätzlich zum eigentlichen Bild als sog. Metadaten gespeichert)?

- ☐ Datum/Uhrzeit
- ☐ Name aller Personen auf dem Bild
- ☐ Name des Fotografen
- ☐ Genauer Ort der Aufnahme (GPS-Daten)
- ☐ Beschreibung was auf dem Bild zu sehen ist
- ☐ Informationen zur Kamera mit der das Bild erstellt wurde

Du besuchst eine Webseite im Internet. Welche Informationen kann diese über dich herausfinden? (Ohne dass du spezielle Einstellungen machen oder spezielle Programme installieren musstest)

- ☐ Von welcher Webseite ich komme
- ☐ Welches Betriebssystem ich benutze
- ☐ Meinen Namen
- ☐ Meine E-Mail-Adresse
- ☐ Mich eindeutig identifizieren
- ☐ Meine Monitorauflösung
- ☐ In welchem Land ich mich befinde
- ☐ Welchen Browser ich benutze
- ☐ Meinen genauen Standort (GPS-Daten)
- ☐ Name einiger Programme die ich installiert habe
- ☐ Meine Interessen
- ☐ Ob ich ein mobiles Gerät oder einen PC nutze
- ☐ Meine Sprache
- ☐ Mein Alter

Welche der Folgenden Aussagen über Datenbanken sind deiner Meinung nach korrekt?

- ☐ Alle Daten müssen konsistent gespeichert sein
- ☐ Bei kleineren Datenmengen lohnt sich eine Datenbank nicht
- ☐ Große Datenmengen schaffen die meisten Datenbanken nicht
- ☐ Nur bis zu 5 Personen können gleichzeitig mit einer Datenbank arbeiten
- ☐ Jede Datenbank liegt auf einem eigenen Server
- ☐ Um Fotos, Videos und so weiter zu speichern, sind Datenbanken kaum geeignet
- ☐ Cloud-Dienste basieren typischerweise auf Datenbanken

Wie schützt du deine Daten vor Verlust?

- ☐ Ich erstelle regelmäßig ein Backup auf USB-Stick oder externer Festplatte
- ☐ Ich synchronisiere sie in die Cloud (z.B. Dropbox)
- ☐ Meine Daten sind nicht so wertvoll, dass ich sie schützen muss
- ☐ Habe ich mir noch keine Gedanken gemacht

Typischerweise sichere ich folgende Daten: _____

_____

Welche der folgenden Aussagen über Datenanalysen sind deiner Meinung nach korrekt?

- ☐ Datenanalysen dauern sehr lange
- ☐ Kleine Datenmengen sind besser, da die Analyse schneller geht
- ☐ Aus großen Datenmengen kann man wenig herauslesen
- ☐ Oft ist es möglich Informationen über Personen herauszufinden, die gar nicht in den Daten stehen
- ☐ Es ist kaum möglich solch große Datenmengen wie sie beispielsweise die NSA vorhält zu analysieren
- ☐ Die Metadaten sind oft wesentlich interessanter als die eigentlichen Daten
- ☐ Meine Daten dürfen ruhig analysiert werden, die finden sowieso nichts neues heraus
- ☐ Ich habe nichts zu verbergen

Anhang C: Detaillierte Beschreibung der Schlüsselkonzepte des Datenmanagements

C.1 Datenunabhängigkeit

Entwurfsprinzip

„Der Definition zufolge beschreibt Datenunabhängigkeit die Möglichkeit, das Schema auf einer Ebene ändern, ohne das Schema der nächsthöheren Ebene ändern zu müssen.“ (Elmasri und Navathe, 2009)

Es wird zwischen logischer und physischer Datenunabhängigkeit unterschieden: Die Erste befasst sich mit Änderungen des konzeptionellen Schemas und thematisiert somit das Hinzufügen, Verändern oder Entfernen von Datenfeldern oder Datensatztypen. Die Zweite berücksichtigt hingegen die Möglichkeit, die physischen Daten neu zu organisieren, ohne das konzeptionelle Schema ändern zu müssen.

Datenunabhängigkeit ist daher eine Folge der klaren Einteilung eines Systems in mehrere konzeptionelle Schichten.

Kernaussagen

- Die Benutzungsoberfläche von Datenmanagementsystemen abstrahiert von der internen Speicherung.
- Die Sichtbarkeit von internen Aspekten wird je nach Benutzerrolle eingeschränkt.
- Zur Trennung des Systems in Schichten werden klare Schnittstellen definiert.
- Datenmanagementsysteme erlauben es, einzelne Schichten auszutauschen, ohne andere anzupassen.
- Insbesondere kann die konzeptionelle Schicht (das Datenmodell) flexibel an sich ändernde Anforderungen angepasst werden.
- Informatiksysteme abstrahieren interne Details, indem diese geeignet versteckt werden, mit dem Ziel Systeme besser nutzbar zu machen.
- Die entstehenden Schichtenarchitekturen werden in der Informatik verbreitet eingesetzt.
- Jede Benutzergruppe hat unterschiedliche Anforderungen, denen ein System genügen muss.

- Für eine Benutzergruppe irrelevante Informationen und Möglichkeiten sollten erkannt und verborgen werden (Geheimnisprinzip).
- Geeignete Schnittstellen müssen (nicht nur bei der Teilung eines Systems in mehrere Schichten) klar definiert werden.

Verwandte Konzepte des Datenmanagements

- Strukturierung
- Repräsentation
- Dauerhaftigkeit
- Verfügbarkeit
- Partitionierung

Relevanz in Verbindung zu Praktiken des Datenmanagements

Datenunabhängigkeit ist insbesondere im Bereich der Modellierung, Implementierung, Optimierung und Analyse relevant, da sich diese Praktiken konkret mit der Strukturierung und Speicherung von sowie dem Zugriff auf Daten beschäftigen.

Anknüpfungspunkte in der Informatik

- **Schichtenarchitekturen:** Die Trennung eines Systems in verschiedene architekturelle Schichten mit verschiedenen Aufgaben kommt in verschiedensten Bereichen der Informatik vor, beispielsweise in der Softwareentwicklung, der Netzwerkkommunikation oder bei Betriebssystemen.
- **Schnittstellendefinitionen:** Die Definition von Schnittstellen ist insbesondere auch bei der Modularisierung von Anwendungen, aber auch für die Netzwerkkommunikation oder den Zugriff auf externe Datenquellen essenziell.

Begründung der Einordnung

Nicht jedes Datenmanagementsystem stellt Datenunabhängigkeit gleichermaßen sicher. Der Grad an Datenunabhängigkeit ist daher eine wichtige Entscheidung, die beim Entwurf eines Systems getroffen werden muss und die zukünftige Nutzung wesentlich beeinflusst. Aus diesem Grund muss der Grad an Datenunabhängigkeit, die ein System bietet, auch bei der Auswahl von Datenmanagementsystemen berücksichtigt werden. Datenunabhängigkeit ist daher klar als Entwurfsprinzip einzustufen.

C.2 Integrität

Entwurfsprinzip

Die Datenintegrität ist ein wichtiges Schutzziel im Rahmen der IT-Sicherheit: „*Wir sagen, dass das System die Datenintegrität (engl. Integrity) gewährleistet, wenn es Subjekten nicht möglich ist, die zu schützenden Daten unautorisiert und unbemerkt zu manipulieren*“ *(Eckert, 2014)*. Im Datenmanagement spielt die Integrität insbesondere in Zusammenhang mit (oft semantischen) Integritätsbedingungen eine Rolle: Durch diese können einem Datenmanagementsystem deklarativ Bedingungen mitgegeben werden, die der Datenbestand zu jedem Zeitpunkt erfüllen soll. Diese Maßnahme ist dabei eine wichtige Möglichkeit zur Wahrung der Konsistenz eines Datenbestands.

Kernaussagen

- Integrität versucht zu gewährleisten, dass Daten nicht unbemerkt/versehentlich verfälscht werden.
- Die Wahrung der Integrität der gespeicherten Daten ist in den meisten Datenmanagementsystemen zentral.
- Zur Sicherstellung der Integrität können sog. Integritätsbedingungen definiert werden.
- Integritätsbedingungen beschreiben Anforderungen, die der Datenbestand erfüllen soll.
- Ein Datenmanagementsystem das Integrität sicherstellt, ist langsamer als eines, dass dies nicht tut.
- Um Daten vor unberechtigter Änderung zu schützen, sind Integritätsbedingungen nicht ausreichend.
- Unbefugte Änderungen können nur durch ausreichenden Zugriffsschutz und Berechtigungsvergabe verhindert werden.
- Daten und Informatiksysteme im Allgemeinen sind angreif- und manipulierbar. Die Reduzierung solcher Möglichkeiten ist in der Informatik zentral.
- Sicherheit besteht nicht nur aus Zugriffsschutz, sondern beinhaltet weitere Aspekte wie Integrität.
- Für die Sicherstellung von Integrität und Sicherheit im Allgemeinen sind weitere Informationen nötig, wie z. B. Integritätsbedingungen, die ein Informatiksystem auswerten kann.

- Nahezu alle Möglichkeiten zur Erhöhung der Sicherheit verlangsamen ein System und/oder beeinträchtigen das Nutzungserlebnis. Der angestrebte Grad an Sicherheit ist daher je nach Anwendungsfall abzuwägen.

Verwandte Konzepte des Datenmanagements

- Konsistenz
- Konkurrenz
- Strukturierung
- Repräsentation

Relevanz in Verbindung zu Praktiken des Datenmanagements

Die Wahrung der Integrität ist immer dann relevant, wenn Daten nicht nur gelesen, sondern auch geschrieben werden können. Dies ist bei den Praktiken Modellierung bis Optimierung der Fall.

Anknüpfungspunkte in der Informatik

- **IT-Sicherheit:** Integrität stellt eines der zentralen Schutzziele der IT-Sicherheit dar.
- **Vergabe von Zugriffsrechten:** Nicht nur in Datenmanagementsystemen ist die Vergabe von Zugriffsrechten zentral, sondern allgemein in Bezug auf Informatiksysteme.
- **Erkennung von Manipulationen:** Die Erkennung von Manipulationen an Datenbeständen, aber auch Datenübertragungen, Software oder Informatiksystemen im Allgemeinen ist in vielen Bereichen der Informatik wichtig.

Begründung der Einordnung

Integritätsbedingungen sind nicht in allen Datenmanagementsystemen gleichermaßen zentral und können – je nach Konzeption eines Systems – auch nicht ohne weiteres eingeführt werden. So ist beispielsweise bei Multimediadatenbanken oder dokumentenorientierten Datenbanken das Thema „Integrität“ wesentlich komplexer als beispielsweise bei relationalen Datenbanken, bei denen Integritätsbedingungen relativ einfach definiert werden können. Integrität ist daher als Entwurfsprinzip einzustufen.

C.3 Konsistenz

Entwurfsprinzip

In Datenbanken ist Konsistenz seit Jahrzehnten ein zentrales Ziel, das lange Zeit als unumstößlich galt. Konsistenz bezeichnet dabei die Widerspruchsfreiheit des Datenbestands. Diese kann durchgehend oder zu definierten Zeitpunkten erzwungen werden. Im Zusammenhang mit Anfragen an Datenbanken stellt Konsistenz eines der vier im ACID-Prinzip beschriebenen Ziele von Transaktionen dar. Trotz der langjährigen Bedeutung verzichten moderne Datenmanagementsysteme heute teils auf Konsistenz bzw. weichen diese auf (beispielsweise moderne NoSQL-Datenbanken).

Kernaussagen

- Konsistenz liegt vor, wenn ein Datenbestand in sich widerspruchsfrei ist.
- Konsistenz befasst sich mit logischen Widersprüchen in Daten.
- Auftretende Anomalien beim Verändern von Daten müssen verhindert werden, um Konsistenz zu ermöglichen.
- Durch voneinander isolierte Ausführung von Abfragen in Transaktionen kann sichergestellt werden, dass sich nebenläufige Abfragen nicht überschneiden und so Inkonsistenzen herbeiführen.
- Ein durchdachter Aufbau des Datenmodells kann dazu beitragen, Konsistenz zu erreichen, indem Redundanzen vermieden werden.
- Um die Konsistenz von Daten sicherzustellen, müssen über Metadaten oder geeignete Strukturierung zusätzliche Informationen über Daten bereitgestellt werden.
- Das Erkennen von Fehlern in den Daten, zu denen auch Inkonsistenzen zählen, kann nur basierend auf diesen zusätzlichen Informationen stattfinden.
- In vielen Anwendungsfällen ist eine dauerhafte Konsistenz weniger wichtig, sodass eine Aufweichung dieser Anforderung zugunsten der Performanz stattfinden kann.

Verwandte Konzepte des Datenmanagements

- Transaktion
- Integrität
- Konkurrenz
- Redundanz
- Strukturierung
- Synchronisation

Relevanz in Verbindung zu Praktiken des Datenmanagements

Die Wahrung der Konsistenz des Datenbestandes ist bei allen schreibenden Zugriffen, aber auch bei der Entwicklung des Datenmodells, relevant. Dies betrifft daher die Praktiken von Modellierung bis Optimierung.

Anknüpfungspunkte in der Informatik

- **Backup:** Während zur Wahrung der Konsistenz im Allgemeinen dazu tendiert wird, Redundanzen zu vermeiden, stellt ein Backup einen Anwendungsfall dar, in dem eine Inkonsistenz einer redundanten Kopie zu deren Original gezielt angestrebt und genutzt wird.
- **Synchronisation:** Bei der Synchronisation von Daten besteht ein hohes Risiko der Entstehung von Inkonsistenzen. Es müssen Maßnahmen zu deren Vermeidung ergriffen werden.
- **Fehlererkennung:** Die Erkennung von Fehlern funktioniert häufig über die Erkennung von Inkonsistenzen in den Daten bzw. zwischen diesen und über sie gespeicherten Metadaten.

Begründung der Einordnung

Konsistenz war jahrelang eine zentrale Anforderung an Datenmanagementsysteme, stellt heute jedoch eindeutig eine Designentscheidung dar. Je nach Anwendungszweck muss dabei mehr oder weniger Wert auf die konsistente Datenspeicherung gelegt, dafür aber ggf. an anderer Stelle Nachteile in Kauf genommen werden müssen (vgl. CAP-Theorem). Konsistenz ist daher ein Entwurfsprinzip des Datenmanagements.

C.4 Isolierung

Entwurfsprinzip

Die Isolierung zählt zu den zentralen Eigenschaften von transaktionalen Systemen und ist im ACID-Paradigma berücksichtigt. „*Diese Eigenschaft verlangt, dass nebenläufig (parallel, gleichzeitig) ausgeführte Transaktionen sich nicht gegenseitig beeinflussen*“ *(Kemper und Eickler, 2015)*. Das Prinzip der Isolierung nebenläufiger Transaktionen ist für einen störungsfreien Mehrbenutzerbetrieb in Datenmanagementsystemen unverzichtbar.

Kernaussagen

- Nebenläufige Abfragen können isoliert voneinander durchgeführt werden, obwohl sie sich real überschneiden.
- Durch eine (quasi-)parallele Ausführung von Transaktionen besteht eine hohe Gefahr der gegenseitigen Beeinflussung.
- Gegenseitige Abhängigkeiten der Abfragen können zu Verklemmungen im Datenmanagementsystem führen.
- Lesende Zugriffe sind immer voneinander unabhängig und können daher beliebig parallelisiert werden.
- Eine Isolierung von Abfragen ist nur nötig, wenn mehrere Benutzer gleichzeitig arbeiten.
- Die Isolierung verhindert das Auftreten von Anomalien, die die Konsistenz des Datenbestands gefährden.

Verwandte Konzepte des Datenmanagements

- Transaktion
- Konsistenz
- Konkurrenz

Relevanz in Verbindung zu Praktiken des Datenmanagements

Die Isolierung ist relevant, wenn (mindestens) zwei schreibende Transaktionen auf denselben Datenbestand zugreifen. Sie tritt daher insbesondere in den Praktiken Implementierung und Optimierung zutage.

Anknüpfungspunkte in der Informatik

- **Parallelisierung:** Eine Isolierung von nebenläufigen Aktionen voneinander ist nicht nur im Datenmanagement, sondern beispielsweise auch der (parallelen) Programmierung oder in Zusammenhang mit Betriebssystemen nötig.
- **Mehrbenutzerarchitekturen:** Die Vermeidung von Seiteneffekten, die durch die Nutzung eines Informatiksystems mit mehreren gleichzeitigen Benutzern entstehen, ist in allen Mehrbenutzersystemen zentral.
- **Ablaufsteuerung:** Zur Vermeidung von unerwünschten Seiteneffekten und zur Isolierung von Aktionen voneinander gibt es verschiedene Möglichkeiten der Ablaufsteuerung, wie beispielsweise Sperren.

Begründung der Einordnung

Die Isolierung von Anfragen an ein Datenmanagementsystem muss im Rahmen des Entwurfs und der Implementierung des Systems realisiert werden. Dabei steht auch die Entscheidung an, welche Maßnahmen zur Isolierung umgesetzt werden. Isolierung kann daher als Entwurfsprinzip des Datenmanagements betrachtet werden.

C.5 Dauerhaftigkeit

Entwurfsprinzip

Die Dauerhaftigkeit stellt sicher, dass die Wirkung einer erfolgreich abgeschlossenen Transaktion erhalten bleibt, bis sie durch weitere Änderungen obsolet wird (nach *Kemper und Eickler (2015)*). Dies stellt einen zu jederzeit nachvollziehbaren Datenbestand sicher und erlaubt es Benutzern, sich nach Bestätigung der Transaktionsausführung darauf zu verlassen, dass Änderungen auch korrekt durchgeführt worden sind und nicht, beispielsweise durch andere nicht korrekt isolierte Transaktionen, versehentlich überschrieben werden.

Kernaussagen

- Bestätigte Änderungen des Datenbestands in einem Datenmanagementsystem sollen eine dauerhafte Wirkung zeigen, sodass...
  - bestätigte Änderungen nicht versehentlich verloren gehen.
  - einmal gemachte Änderungen nur durch gezieltes Überschreiben (bzw. durch eine gegenteilige Änderung) rückgängig gemacht werden können.
- Auch im Systemausfall müssen bereits bestätigte Änderungen nachvollzogen und wiederhergestellt werden können.
- Im Falle eines Fehlers innerhalb einer Transaktion dürfen keine Rückstände eines Teils dieser Transaktion im System erkennbar sein.
- Eine übliche Implementierung der Dauerhaftigkeit basiert auf dem Zwei-Phasen-Sperrprotokoll.
- Die Dauerhaftigkeit wird nicht nur durch Nutzung zuverlässiger Speicher, sondern auch durch Replikation und/oder Synchronisation von Daten und durch Protokollierung von Veränderungen sichergestellt.

Verwandte Konzepte des Datenmanagements

- Integrität

- Konsistenz
- Transaktion
- Synchronisation
- Replikation

Relevanz in Verbindung zu Praktiken des Datenmanagements

Da das Prinzip der Dauerhaftigkeit bei schreibendem Zugriff auf das Datenmanagementsystem relevant ist, hat es insbesondere bei Implementierung und Optimierung Bedeutung.

Anknüpfungspunkte in der Informatik

- **Sicherung/Wiederherstellung:** Im Falle eines Systemausfalls müssen Maßnahmen ergriffen werden, um Änderungen beispielsweise aus Protokollen wiederherstellen zu können. Dies ist nicht nur in Datenmanagementsystemen relevant, sondern es stellt sich in der Informatik allgemein die Frage, wie Daten vor Verlust geschützt werden können.
- **Datenspeicherung im Allgemeinen:** Im Zusammenhang mit der Datenspeicherung stellt sich im Allgemeinen die Frage, wie verhindert werden kann, dass durchgeführte Änderungen den Datenbestand gefährden. Um beispielsweise beim Verschieben von Dateien im Dateisystem zu vermeiden, dass bei einem „Verschiebefehler“ die ursprünglichen Dateien verloren sind, kopieren viele Nutzer diese erst und löschen das Original danach (was aufgrund der Implementierung im Betriebssystem eigentlich unnötig ist).

Begründung der Einordnung

Der angestrebte Grad an Dauerhaftigkeit ist eine Entscheidung, die bei der Entwicklung eines Datenmanagementsystems getroffen werden muss. Auch entsprechende Maßnahmen zur Umsetzung müssen in dieser Phase bereits getroffen werden. Somit stellt Dauerhaftigkeit ein Entwurfsprinzip dar.

C.6 Verfügbarkeit

Entwurfsprinzip

Die Verfügbarkeit beschreibt, dass ein Datenmanagementsystem zu einem bestimmten bzw. idealerweise jedem beliebigen Zeitpunkt mit einer angemessenen Reaktionszeit er-

reichbar und nutzbar sein muss (vgl. *Edlich et al. (2011)*). Die Verfügbarkeit von Daten wird in Datenmanagementsystemen im Allgemeinen schon seit langem als zentral angesehen. Im Zusammenhang mit immer häufiger in Echtzeit und von vielen Quellen gleichzeitig erfolgenden Zugriffen wird sie aber immer relevanter.

Kernaussagen

- Bei den meisten Datenmanagementsystemen ist Verfügbarkeit eine zentrale Eigenschaft.
- Verfügbarkeit schließt eine hohe Performanz des Systems ein, aber auch möglichst kurze Ausfallzeiten.
- Verfügbarkeit steht damit in direkter Konkurrenz zu Maßnahmen, die die Sicherheit und Integrität des Datenbestands erhöhen, da diese das System im Allgemeinen verlangsamen und gegebenenfalls zeitweise blockieren.
- Ein besonders hoher Grad an Verfügbarkeit wird in Echtzeitsystemen benötigt, da diese bestimmte Zeitschranken einhalten müssen.
- Verfügbarkeit betrifft nicht nur Datenmanagementsysteme, sondern zeigt sich klar in allen Informatiksystemen und ist besonders auch in sicherheitskritischen Anwendungen zentral.
- Zur Erhöhung der Verfügbarkeit werden Daten oft auf mehrere Knoten repliziert.

Verwandte Konzepte des Datenmanagements

- Integrität
- Konkurrenz
- Replikation
- Transport
- Redundanz

Relevanz in Verbindung zu Praktiken des Datenmanagements

Die Verfügbarkeit ist bei allen Tätigkeiten, bei denen direkt mit dem System bzw. den Daten interagiert wird, zentral. Dies trifft insbesondere auf die Praktiken Implementierung bis Analyse und ggf. auch Visualisierung zu.

Anknüpfungspunkte in der Informatik

- **Informatiksysteme allgemein:** Verfügbarkeit ist bei allen Informatiksystemen ein relevantes Thema. Maßnahmen zur Erhöhung der Verfügbarkeit von Datenmanagementsystemen können daher beispielsweise auch übertragen auf die Verfügbarkeit von Software wiedergefunden werden, indem web- bzw. cloudbasierte Anwendungen oft auf vielen Rechnerknoten repliziert zur Verfügung stehen.
- **Echtzeitsysteme:** Die Verarbeitung von Daten in Echtzeit ist ein immer häufiger anzutreffendes Ziel der Informatik. In allen Fällen spielt Verfügbarkeit eine zentrale Rolle.

Begründung der Einordnung

Verfügbarkeit ist eine Eigenschaft von Datenmanagementsystemen, die in klarem Widerspruch zu anderen Eigenschaften wie Sicherheit und Dauerhaftigkeit steht: Jegliche Maßnahmen zur Erhöhung dieser beiden Eigenschaften senken gleichzeitig die Verfügbarkeit des Systems. Es ist daher je nach Anwendungsfall zu entscheiden, welcher Grad an Verfügbarkeit erreicht werden soll, sodass Verfügbarkeit als Entwurfsprinzip einzuordnen ist.

C.7 Partitionstoleranz

Entwurfsprinzip

Die Partitionstoleranz ist ein Entwurfsprinzip, das in verteilten Datenbanken zum Tragen kommt. Ein Datenmanagementsystem kann, je nach Grad der Partitionstoleranz, den Ausfall der Kommunikation mit einem oder mehreren seiner Knoten verkraften, ohne dass die Funktion eingeschränkt wird oder Fehler bzw. Inkonsistenzen auftreten.

Kernaussagen

- Verteilte Systeme müssen mit dem Ausfall eines oder mehrerer Knoten zurechtkommen.
- Neben dem Ausfall der Knoten ist auch ein Ausfall der Kommunikationsverbindung zu berücksichtigen.
- Bei Ausfällen der Kommunikation kann ein System in mehrere Teile getrennt werden („partitioniert“), die unabhängig voneinander arbeiten.
- Bei Wiedervereinigung des Systems müssen die Änderungen an den Datenbeständen so synchronisiert und nachvollzogen werden, als wäre keine Partitionierung erfolgt.

- Die Partitionstoleranz steht damit aber in klarer Konkurrenz zur Konsistenz und Verfügbarkeit: Hohe Partitionstoleranz bei gleichzeitiger Konsistenz sorgt für Performanceverluste.
- Die Schaffung einer Partitionstoleranz ist nicht nur bei verteilten Datenmanagementsystemen, sondern allgemein bei verteilten Systemen eine zentrale Herausforderung.

Verwandte Konzepte des Datenmanagements

- Konsistenz
- Verfügbarkeit
- Synchronisation
- Transport
- Dauerhaftigkeit
- Transaktion

Relevanz in Verbindung zu Praktiken des Datenmanagements

Partitionstoleranz ist insbesondere im Hintergrund relevant und kommt im Fehlerfall zum Tragen, sodass diese idealerweise nach außen kaum sichtbar wird. Falls doch, dann tritt sie jedoch während allen Interaktionen mit Datenmanagementsystemen, also insbesondere während den Tätigkeiten von Implementierung bis Analyse zutage.

Anknüpfungspunkte in der Informatik

- **Verteilte Systeme:** Der Ausfall von Knoten oder der Kommunikation zwischen diesen ist in allen verteilten Systemen ein relevantes Thema.
- **Rechnerkommunikation:** Im Rahmen der Rechnerkommunikation werden möglichst zuverlässige und sichere Kommunikationsmöglichkeiten entwickelt.

Begründung der Einordnung

Je nach Einsatz und Gestaltung eines Systems kann Partitionstoleranz unverzichtbar aber auch relativ überflüssig sein. Ein gewisser Grad an Partitionstoleranz ist dabei in den meisten verteilten Datenmanagementsystemen zentral. Beim Entwurf und der Implementierung des Systems muss jedoch die Entscheidung getroffen werden, wie stark diese Eigenschaft

priorisiert werden soll, da sie in Konkurrenz zu insbesondere Konsistenz und Verfügbarkeit steht. Damit kann Partitionstoleranz als Entwurfsprinzip angesehen werden.

C.8 Nebenläufigkeit

Entwurfsprinzip

Nebenläufigkeit kann in Zusammenhang mit Datenmanagementsystemen als synonym zur Konkurrenz betrachtet werden, diese betont jedoch nach *Lockemann (1986)* stärker den Wettbewerbsgedanken um Ressourcen. Der Begriff betont damit, dass immer wenn mehrere Nutzer gleichzeitig mit einem Datenmanagementsystem arbeiten, eine Konkurrenzsituation entsteht, mit der das System gewissermaßen umgehen muss.

Der Grad an Nebenläufigkeit, den ein Datenmanagementsystem zulässt, steht dabei in direktem Zusammenhang mit der Menge an Maßnahmen, die dieses ergreifen muss, um andere Prinzipien wie Konsistenz, Integrität, Verfügbarkeit u. Ä. sicherzustellen.

Kernaussagen

- Die verschiedenen Nutzer bzw. Aktionen eines Datenmanagementsystems stehen miteinander in Konkurrenz um die Ressourcen des Systems (insbesondere die Daten).
- Mehrbenutzerbetrieb erhöht zwar i. A. die Verfügbarkeit des Systems, sorgt aber gleichzeitig für Einschränkungen bezüglich anderer Prinzipien: Beispielsweise müssen weitere Maßnahmen ergriffen werden, um die Integrität des Datenbestands aufrechtzuerhalten.
- Ohne explizite Behandlung von Wettbewerbssituationen besteht die Gefahr von Inkonsistenzen und der Entstehung von Fehlern im Datenbestand.
- Konkurrenz kann durch gezielte Ablaufsteuerung in den Griff bekommen werden.
- Zur Auflösung von Konkurrenzsituationen werden Maßnahmen ergriffen wie beispielsweise das Sperren von (Lese- und Schreib-)Zugriffen auf Daten für die Dauer anderer (schreibender) Zugriffe.

Verwandte Konzepte des Datenmanagements

- Transaktion
- Verfügbarkeit
- Integrität
- Konsistenz

Relevanz in Verbindung zu Praktiken des Datenmanagements

Die Konkurrenz bzw. Nebenläufigkeit erlaubt eine gleichzeitige Nutzung der Datenmanagementsysteme durch mehrere Nutzer. Daher spielt dieses Entwurfsprinzip bei allen Tätigkeiten, in denen direkt Datenmanagementsysteme verwendet werden, d. h. insbesondere in den Phasen Implementierung bis Analyse eine große Rolle.

Anknüpfungspunkte in der Informatik

- **Mehrbenutzersysteme im Allgemeinen:** Nebenläufigkeit spielt in allen Mehrbenutzersystemen eine zentrale Rolle und muss jeweils geeignet beherrscht werden.
- **Parallelisierung:** Auch im Rahmen von Parallelisierung spielen dieselben Konzepte und Maßnahmen wie bei Nebenläufigkeit bzw. Konkurrenz eine wichtige Rolle.
- **Betriebssysteme:** In Betriebssystemen sind die Konzepte der Konkurrenz bzw. Nebenläufigkeit besonders zentral, da diese sich hier nicht nur bei Mehrbenutzerbetrieb, sondern auch im Bereich des Multi-Tasking auswirken.

Begründung der Einordnung

Der Grad an Nebenläufigkeit, den ein System zulässt, ist eine Entscheidung, die im Rahmen der Entwicklung eines Datenmanagementsystems getroffen und die gegenüber anderen Entwurfsprinzipien abgewogen werden muss. Es handelt sich daher auch bei Nebenläufigkeit um ein Entwurfsprinzip.

C.9 Redundanz

Entwurfsprinzip

Redundanz tritt auf, wenn identische Daten als Kopien an verschiedenen Orten gespeichert werden. Da Redundanzen die Gefahr bergen, zu Anomalien und damit Inkonsistenzen im Datenbestand zu führen, gilt es häufig, diese zu vermeiden. Gleichzeitig wird Redundanz jedoch auch an verschiedenen Stellen gezielt eingesetzt, beispielsweise um Effizienzsteigerungen zu bewirken, aber auch zur Erhöhung der Ausfallsicherheit und Verfügbarkeit von Systemen.

Kernaussagen

- Redundante Datenspeicherung birgt das Risiko der Entstehung von Inkonsistenzen im Datenbestand.

- Zur Vermeidung von Redundanz können Datenbestände beispielsweise normalisiert werden.
- Redundante Daten können zur Erhöhung der Verfügbarkeit eines Systems im Sinne der Replikation der Daten eingesetzt werden.
- Durch Redundanz kann die Sicherheit von Daten erhöht werden, da ggf. verlorene Daten wiederhergestellt werden können.
- Redundanz kann auf verschiedener konzeptioneller Ebene eingesetzt werden bzw. auftreten: Beispielsweise können Informationen in einer Kopie eines Datensatzes redundant sein, aber auch die physischen Daten redundant gespeichert werden.

Verwandte Konzepte des Datenmanagements

- Integrität
- Konsistenz
- Verfügbarkeit
- Replikation
- Synchronisation
- Strukturierung

Relevanz in Verbindung zu Praktiken des Datenmanagements

Die Vermeidung oder Zulassung von Redundanz ist insbesondere in den Planungsphasen des Datenmanagements zentral. Dies betrifft daher die Praktiken Modellierung bis Optimierung.

Anknüpfungspunkte in der Informatik

- **Dateispeicherung allgemein:** Selbst bei Verwendung normaler Dateisysteme kann die Vermeidung von Redundanz zielführend sein, indem beispielsweise Links genutzt werden, anstatt Dateien in Kopie abzuspeichern, da auch dabei Änderungsanomalien auftreten können.
- **Backup:** Backups stellen eines der zentralen Beispiele für die Nutzung von Redundanz dar.
- **Programmierung:** Die hinter dem Prinzip der Redundanznutzung und -vermeidung liegende Idee kann auch in der Programmierung angetroffen werden, die im Allgemeinen eher versucht, Dopplungen im Programmcode aus demselben Grund zu

vermeiden wie bei Datenmanagementsystemen. Auch hier kann es jedoch stellenweise sinnvoll sein, Redundanz auszunutzen um Code verständlicher zu machen, ein Nachladen von Modulen zu verhindern o. Ä.

- **Fehlertoleranz:** Auch abseits des Datenmanagements kann Redundanz zur Erhöhung der Fehlertoleranz eines Gesamtsystems eingesetzt werden, indem beispielsweise redundante Server vorgehalten werden, die bei Ausfall eines Servers automatisiert einspringen können.
- **Fehlererkennung:** Auch zur Erkennung von Fehlern eines Computersystems kann Redundanz hilfreich sein, indem beispielsweise Berechnungen in sicherheitsrelevanten Bereichen (bspw. Fahrstraßensteuerung bei Bahnstellwerken) durch drei unabhängig arbeitende redundante Computer durchgeführt werden und nur als korrekt angenommen werden, falls mindestens zwei dieser Rechner zum selben Ergebnis kommen.

Begründung der Einordnung

Die Nutzung und/oder Vermeidung von Redundanz auf unterschiedlichen Ebenen zeigt, dass es sich bei diesem Prinzip um eine Entscheidung handelt, die im Rahmen des Entwicklungsprozesses eines Datenmanagementsystems, aber auch bei der Strukturierung von Daten getroffen werden muss. Redundanz stellt daher klar ein Entwurfsprinzip des Datenmanagements dar.

C.10 Strukturierung

Mechanismus

Strukturierung ist einer der zentralen Aspekte bei der Speicherung und Verwaltung von, aber auch beim Zugriff auf Daten: Diese müssen immer strukturiert abgelegt werden, damit ein gezielter Zugriff auf diese möglich ist; die Struktur von gespeicherten Daten wird dabei stark durch die Art der Daten und den Einsatzzweck beeinflusst und kann drastische Auswirkungen auf Performanz und Nutzbarkeit von Systemen haben. Daher können geeignete Primär- und Sekundärstrukturen auch eingesetzt werden, um ein System in diesem Zusammenhang zu optimieren.

Kernaussagen

- Um einen Zugriff auf Daten zu ermöglichen, muss immer eine gewisse Strukturierung vorgenommen werden.
- Für die Vermeidung von Inkonsistenzen bietet sich eine möglichst starke Strukturierung der Daten an.

- Eine starke Strukturierung sorgt meist für eine starke Partitionierung des Datenbestands und senkt damit die Performanz des Zugriffs.
- Zur Strukturierung werden in vielen Fällen Metadaten eingesetzt.
- Neben der Strukturierung der eigentlichen Daten durch Festlegung von Primärstrukturen, können zusätzliche Sekundärstrukturen festgelegt werden, die alternative Wege zum Zugriff auf Daten eröffnen.
- Die Strukturierung von Daten kann auf verschiedenen Ebenen des Datenmanagementsystems erfolgen, der physikalischen, logischen oder konzeptuellen Ebene.
- Das im Rahmen der Datenmodellierung meist genutzte und auch am nähesten am Nutzer gelegene Modell ist das konzeptionelle Modell.

Verwandte Konzepte des Datenmanagements

- Datenunabhängigkeit
- Integrität
- Konsistenz
- Redundanz
- Repräsentation

Relevanz in Verbindung zu Praktiken des Datenmanagements

Die Strukturierung ist, je nach Ebene, auf der sie stattfindet, insbesondere in Zusammenhang mit der Modellierung, Implementierung und Optimierung von Bedeutung.

Anknüpfungspunkte in der Informatik

- **Rechnerkommunikation:** Strukturierende Ansätze finden sich in der Rechnerkommunikation beispielsweise beim Aufbau von Rechnernetzen und der Strukturierung von Kommunikation zwischen verschiedenen Systemen.
- **Softwareengineering:** Im Softwareengineering ist die Strukturierung sowohl für den Projektverlauf als auch die Planung des konkreten Produkts zentral.
- **Theoretische Informatik:** In der theoretischen Informatik werden strukturierende Aspekte genutzt, um Probleme zu modellieren, aber auch um Zusammenhänge verschiedener Konzepte der theoretischen Informatik zu betonen (z. B. Chomsky-Hierarchie).

Begründung der Einordnung

Obwohl der Grad an Strukturierung je nach System, Verwendungszweck und Daten entschieden werden muss, handelt es sich bei der Strukturierung nicht um ein Entwurfsprinzip: Strukturierung wird im Datenmanagement insbesondere eingesetzt, um verschiedene Entwurfsprinzipien zu erreichen und nicht zum Selbstzweck. Es handelt sich bei der Strukturierung daher eher um ein Prinzip, das die korrekte Funktion eines Datenmanagementsystems ermöglicht, d. h. um einen Mechanismus.

C.11 Repräsentation

Mechanismus

Unter Repräsentation werden alle Aspekte der internen Speicherung von Informationen in einem Datenmanagementsystem verstanden. Dies beinhaltet Themen wie Datenstrukturen (z. B. Suchbäume), Optimierungen in diesem Bereich (z. B. Kompression, Einsatz von Pufferspeichern) und grundsätzliche Möglichkeiten zur Datenspeicherung (bspw. die Nutzung von Vorder- und Hintergrundspeichern).

Kernaussagen

- Die Repräsentation erfolgt auf unterschiedlichen Ebenen:
  - Informationen müssen zur Speicherung im Datenmanagementsystem als Daten repräsentiert werden.
  - Die Daten werden wiederum für die physikalische Speicherung geeignet repräsentiert.
  - Zur Organisation der Daten werden diese geeigneten Datenstrukturen gespeichert, die sich je nach Art der Daten und Ziel des Datenmanagementsystems unterscheiden.
- Bei der Speicherung von Daten muss zwischen Vorder- und Hintergrundspeichern unterschieden werden.
- Zur Optimierung der Performanz eines Systems können Daten zusätzlich in einem Pufferspeicher vorgehalten werden.
- Das Datenvolumen kann durch Kompression in gewissen Grenzen verringert werden.
- Ohne geeignete Interpretationsvorschriften sind die gespeicherten Daten oft wertlos, da die Information nicht zurückgewonnen werden kann.

Verwandte Konzepte des Datenmanagements

- Strukturierung
- Dauerhaftigkeit
- Transport

Relevanz in Verbindung zu Praktiken des Datenmanagements

Die geeignete Repräsentation von Daten findet üblicherweise im Hintergrund statt und tritt kaum bei den Praktiken des Datenmanagements zutage, außer beispielsweise durch Auswahl von Datentypen, die über die Repräsentation mitentscheiden. Sie ist damit höchstens im Bereich der Modellierung und Implementierung erkennbar. Die zweite Ebene der Repräsentation, nämlich die Darstellung der Informationen als Daten, ist hingegen klar bei der Datenerfassung/-gewinnung relevant, aber auch der Datenbereinigung.

Anknüpfungspunkte in der Informatik

- **Theoretische Informatik:** Die Repräsentation von Informationen wird in der theoretischen Informatik beispielsweise durch die Betrachtung des Informationsgehalts thematisiert.
- **Datenstrukturen:** Die im Bereich der Datenstrukturen thematisierten Möglichkeiten zur Speicherung von Daten finden im Datenmanagement starke Anwendung zur Repräsentation der Daten.
- **Rechnerkommunikation:** In der Rechnerkommunikation müssen Daten so geeignet repräsentiert werden, dass sie zuverlässig und fehlerfrei übertragen werden können.

Begründung der Einordnung

Die Repräsentation findet eher in den ersten Schritten des Datenmanagements, bei der Entwicklung der Systeme und der Gewinnung der Daten, statt. Durch sie werden alle weiteren Möglichkeiten des Datenmanagements erst umsetzbar. Es handelt sich daher bei der Repräsentation um einen dem Datenmanagement zugrundeliegenden Mechanismus.

C.12 Replikation

Mechanismus

Replikation bezeichnet die gezielte redundante Speicherung von Daten an mehreren Orten / auf mehreren Datenspeichern. In verteilten Systemen werden Daten möglichst sofort repliziert, sodass zu jedem Zeitpunkt eine Kopie der aktuellen Daten vorliegt, die beispielsweise im Falle eines Ausfalls eines Dateispeichers als Ersatz genutzt werden kann. Je nach Ziel kann die Replikation synchron ablaufen, d. h. die eigentliche Abfrage solange blockieren, bis die Replikation abgeschlossen ist, oder asynchron und ohne Blockierung im Hintergrund stattfinden. Auch im Heimanwenderbereich bekommt Replikation – beispielsweise in Zusammenhang mit Netzwerkspeichersystemen (NAS) – zunehmende Bedeutung: Hier wird Replikation eingesetzt, indem Festplatten in RAID-Arrays organisiert werden, die Daten u. A. auf zwei Festplatten repliziert speichern können.

Kernaussagen

- Daten können – automatisiert oder manuell – auf verschiedene Datenspeicher repliziert werden.
- Durch die Replikation kann sowohl die Ausfallsicherheit eines Informatiksystems im Gesamten, aber auch die Verlustsicherheit der Daten selbst erhöht werden.
- Gleichzeitig sorgt Replikation durch die dadurch erzeugte Redundanz auch für potenzielle Einschränkungen der Konsistenz der Daten.
- Je nach angestrebter Form der Replikation – d. h. synchron oder asynchron – kann diese die Performanz des Datenmanagementsystems negativ beeinflussen.
- Während eine asynchrone Replikation zwar positive Einflüsse auf die Performanz des Systems hat, wird dadurch die Dauerhaftigkeit und Konsistenz eingeschränkt, da nicht garantiert werden kann, dass im Fehlerfall bereits alle bestätigten Änderungen auch repliziert wurden.

Verwandte Konzepte des Datenmanagements

- Konsistenz
- Integrität
- Verfügbarkeit
- Dauerhaftigkeit
- Partitionstoleranz
- Redundanz

- Synchronisation
- Transport
- Transaktion

Relevanz in Verbindung zu Praktiken des Datenmanagements

Replikation findet im Datenmanagementsystem im Hintergrund statt und sollte durch den Nutzer kaum bemerkbar sein. Daher muss sie insbesondere bei der Implementierung und Optimierung beachtet werden, aber auch bei der Löschung von Daten, die redundante Kopien miteinbeziehen muss.

Anknüpfungspunkte in der Informatik

- **Verteilte Systeme:** Die Replikation spielt in verteilten Systemen allgemein eine Rolle, da, selbst wenn das System kein Datenmanagementsystem darstellt, gewisse Daten repliziert werden müssen. Gleichzeitig wird das Grundprinzip der Replikation, die redundante Vorhaltung zueinander äquivalenter Ressourcen, auch ansonsten vielfältig bei solchen Systemen eingesetzt.
- **Rechnerkommunikation:** Jegliche Verfahren zur Replikation basieren auf solchen aus dem Fachgebiet Rechnerkommunikation. Bei der Replikation spielen verschiedene Konzepte aus der Rechnerkommunikation, wie die synchrone oder asynchrone Datenübertragung, die Grenzen dieser Kommunikation, etc. eine deutliche Rolle.
- **Cloud-Computing:** Im immer wichtiger werdenden Bereich des Cloud-Computing ist Replikation heute zentral, da nur durch Replikation ein Cloud-System ausfallsicher und performant entworfen werden kann.

Begründung der Einordnung

Als eher technischer Aspekt, der in Datenmanagementsystemen im Hintergrund relevant ist, wurde die Replikation als Mechanismus eingeordnet, der im offensichtlicheren Bereich zur Erreichung verschiedener Entwurfsprinzipien beiträgt.

C.13 Synchronisation

Mechanismus

Im Bereich des Datenmanagements können zwei Arten der Synchronisation erkannt werden: Einerseits die Synchronisation von Daten, die über reine Replikation in der Hinsicht

hinausgeht, dass Entscheidungen darüber getroffen werden, wie beispielsweise mit konkurrierenden Änderungen umgegangen wird. Andererseits bezeichnet Synchronisation auch die Koordination gleichzeitiger bzw. konkurrierender Zugriffe auf Datenbestände, die immer dann erfolgen muss, wenn konkurrierende Zugriffe auf Daten erlaubt werden sollen, die sich jedoch nicht beeinflussen dürfen und voneinander unbemerkt ablaufen müssen.

Kernaussagen

- Bei der Synchronisation von Daten entstehen Konflikte, die aufgelöst werden müssen.
- Eine automatische Lösung von Synchronisationskonflikten ist oft nicht möglich.
- Die Synchronisation von Daten kann (theoretisch) auf beliebig viele Geräte/Speicher stattfinden.
- Zur Vermeidung einer gegenseitigen Beeinflussung von konkurrierenden Abfragen bzw. Zugriffen müssen diese geeignet synchronisiert werden.
- Zur Synchronisation werden verschiedene Techniken eingesetzt, wie beispielsweise Sperren oder die Synchronisierung anhand von Zeitstempeln.
- Synchronisierungsmaßnahmen werden insbesondere eingesetzt, um dafür zu sorgen, dass eine serialisierbare Abfolge an Abfragen entsteht.
- Ein verbreitetes Sperrprotokoll stellt das Zwei-Phasen-Sperrprotokoll dar.

Verwandte Konzepte des Datenmanagements

- Integrität
- Isolierung
- Transaktion
- Konkurrenz
- Replikation
- Transport

Relevanz in Verbindung zu Praktiken des Datenmanagements

Die Synchronisation wird in denselben Bereichen relevant, in denen auch Konkurrenz auftreten kann. Es handelt sich daher um alle Praktiken, bei denen direkt Datenmanagementsysteme genutzt werden, d. h. insbesondere in den Phasen Implementierung bis Analyse.

Anknüpfungspunkte in der Informatik

- **Betriebssysteme:** Die Synchronisation spielt in Betriebssystemen insbesondere auf Prozessebene eine zentrale Rolle und wird mit ähnlichen Maßnahmen wie im Datenmanagement gelöst.
- **Parallele Programmierung:** In der parallelen Programmierung ist die Synchronisation zentral, um beispielsweise Zugriffe auf gemeinsam genutzte Variablen (bzw. Ressourcen allgemein) zu kontrollieren.

Begründung der Einordnung

Synchronisation ist eine technische Lösung, die es ermöglicht, Datenmanagementsysteme mit verschiedenen Nutzern nebenläufig zu verwenden und/oder Daten lokal vorzuhalten, ohne mit einem System kontinuierlich kommunizieren zu müssen. Sie kann daher klar als Mechanismus des Datenmanagements eingeordnet werden.

C.14 Partitionierung

Mechanismus

Ähnlich wie bei der Replikation findet bei der Partitionierungvon Daten eine Nutzung mehrerer Datenspeicher zur Speicherung statt. Im Gegensatz zur Replikation werden jedoch nicht dieselben Daten auf mehreren Datenspeichern gespeichert, sondern stattdessen die Daten auf mehrere Speicher (nicht-redundant) verteilt. Dies ist daher insbesondere hilfreich, wenn die Kapazität eines einzelnen Datenspeichers nicht mehr ausreicht oder zur Erhöhung der Performanz des Gesamtsystems auf mehrere Datenspeicher gleichzeitig geschrieben werden soll.

Kernaussagen

- Daten können durch Partitionierung auf verschiedene Datenspeicher verteilt werden.
- Partitionierung kann die Gesamtkapazität eines Systems über die Kapazität eines einzelnen Speichervolumens hinaus erhöhen.
- Durch Partitionierung kann die Geschwindigkeit des Gesamtsystems erhöht werden, indem auf mehrere Datenspeicher parallel geschrieben bzw. von diesen gelesen wird.
- Partitionierung kann für erhöhte Anfragezeiten sorgen, wenn einzelne Datenknoten ausfallen oder langsam reagieren, obwohl das Gesamtsystem verfügbar scheint.
- Bei Partitionierung der Daten werden Maßnahmen zur Sicherung der Integrität des Datenbestandes aufwendiger.

Verwandte Konzepte des Datenmanagements

- Partitionstoleranz
- Integrität
- Verfügbarkeit
- Transport

Relevanz in Verbindung zu Praktiken des Datenmanagements

Da die Partitionierung eher im Hintergrund relevant ist, sollte sie höchstens im Bereich der Implementierung und Optimierung relevant werden, da sie die Nutzbarkeit des Gesamtsystems während dieser Praktiken beeinflusst. Idealerweise tritt sie aber für den Nutzer nicht augenscheinlich zutage.

Anknüpfungspunkte in der Informatik

- **Parallelisierung:** Im Rahmen der Parallelisierung von Aufgaben werden diese in Teilaufgaben (gleicher oder unterschiedlicher Art) zerlegt, die auf mehreren Rechenknoten parallel abgearbeitet werden können.
- **Blockchain:** Bei Blockchain-Algorithmen werden die Informationen über die gesamte Kette dezentral vorgehalten, indem viele Datenblöcke kryptografisch miteinander verknüpft werden.

Begründung der Einordnung

Partitionierung ist eine Maßnahme, die im Hintergrund im Datenmanagementsystem erfolgt und hauptsächlich aus technischen Gründen getroffen wird, da Beschränkungen einzelner Datenspeicher hinsichtlich ihrer Kapazität oder Performanz aufgebrochen werden müssen. Es handelt sich damit bei der Partitionierung klar um einen Mechanismus des Datenmanagements.

C.15 Transport

Mechanismus

Beim Transport handelt es sich um die Übertragung von Daten innerhalb eines Systems oder über dessen Grenzen hinweg. Die Übertragung der Daten muss dabei – je nach Zweck – verschiedenen Kriterien genügen und beispielsweise eine hohe Vertraulichkeit, Geschwin-

digkeit oder Fehlertoleranz sicherstellen. Obwohl die Funktionsweise des Transports eher Thema des Fachgebiets Rechnerkommunikation ist, stellt dieser einen wichtigen Mechanismus für das Datenmanagement dar, da Datenmanagementsysteme erst dadurch sinnvoll nutzbar werden. Gleichzeitig müssen Eigenschaften des Transports von Daten und der Fakt, dass dieser geschieht, auch bei der Implementierung von Datenmanagementsystemen bedacht werden, um dadurch entstehende Fehler und Verzögerungen zu berücksichtigen.

Kernaussagen

- Der Transport erfolgt mit üblichen Protokollen und mittels üblicher Methoden der Rechnerkommunikation.
- Je nach Art und Zweck des Systems kann es sinnvoll sein, dieses möglichst verteilt zu organisieren, sodass sehr viel Transport nötig wird, aber auch es möglichst zentralisiert aufzubauen und Transport soweit möglich zu vermeiden.
- Umso mehr Transport von Daten nötig ist, umso langsamer wird in der Regel das Datenmanagementsystem.
- Daten müssen für den Transport in der Regel verschlüsselt und mit zusätzlichen Informationen, wie Metadaten zur Absicherung des Transports und zur Erhöhung der Fehlertoleranz, versehen werden.
- Die Sicherstellung von Konsistenz sorgt in verteilten Systemen für besonders viel Kommunikation.

Verwandte Konzepte des Datenmanagements

- Partitionierung
- Synchronisation
- Replikation
- Verfügbarkeit
- Integrität

Relevanz in Verbindung zu Praktiken des Datenmanagements

Transport wird in Zusammenhang mit verschiedenen Praktiken relevant, immer dann, wenn Daten mit dem System, innerhalb des Systems oder vom Nutzer mit anderen ausgetauscht werden. Dies ist insbesondere bei der Datenerfassung/-gewinnung, der Implementierung, der Analyse und dem Austausch der Fall, wobei auch bei vielen der anderen Praktiken in Ansätzen Zusammenhänge zu diesem Mechanismus existieren.

Anknüpfungspunkte in der Informatik

- **Programmierung:** In der Programmierung tritt Transport insbesondere in Form von Kommunikation zwischen Modulen auf. Umso höher die Kopplung zweier Module zueinander ist, umso mehr Kommunikation wird für deren Aufgaben nötig.
- **Rechnerkommunikation:** Die Organisation des Transports von Daten ist die zentrale Aufgabe des Fachgebiets Rechnerkommunikation.

Begründung der Einordnung

Obwohl es sich beim Transport prinzipiell um ein Thema des Fachgebiets Rechnerkommunikation handelt, ist dieses für die Realisierung von Datenmanagementsystemen zentral. Ohne den Transport von Daten wäre die Funktion von diesen Systemen undenkbar – somit stellt Transport auch einen Mechanismus des Datenmanagements dar.

C.16 Transaktion

Mechanismus

Transaktionen werden in Datenmanagementsystemen als Mechanismus genutzt, um einzelne Abfragen zu einer Gruppe von Abfragen zu gliedern, die nur gemeinsam ausgeführt werden. Auf diese Weise kann sichergestellt werden, dass eine Änderung, die aus mehreren Teilabfragen besteht, nur entweder komplett durchgeführt wird oder komplett ohne Wirkung verbleibt. Typischerweise erfolgen Maßnahmen zur Beherrschung von Konkurrenz insbesondere auf Transaktionsebene, sodass Transaktionen bevor sie komplett ausgeführt worden sind keinerlei Einfluss auf andere Transaktionen bzw. Abfragen haben und somit Abfragen unabhängig voneinander scheinen.

Kernaussagen

- Transaktionen werden genutzt, um eine dauerhafte Konsistenz des Datenbestands sicherzustellen.
- Transaktionen werden nur komplett oder gar nicht ausgeführt.
- Im Fehlerfall während der Ausführung einer Transaktion wird diese komplett rückgängig gemacht („rollback“).
- Die Koordination konkurrierender Zugriffe erfolgt typischerweise auf Transaktionsebene.

- Die nicht-Atomarität von Transaktionen ermöglicht es, dass zwei nebenläufige Transaktionen gegenseitig voneinander so abhängen, dass dieser Konflikt nicht ohne Weiteres aufgelöst werden kann („deadlock“). In einem solchen Fall wird meist eine Transaktion abgebrochen.

Verwandte Konzepte des Datenmanagements

- Konsistenz
- Isolation
- Dauerhaftigkeit
- Konkurrenz

Relevanz in Verbindung zu Praktiken des Datenmanagements

Die Transaktionalität eines Datenmanagementsystems ist bei der schreibenden Interaktion mit dem System relevant, obwohl sie idealerweise eigentlich für den Nutzer unbemerkbar sein sollte. Damit ist sie insbesondere in Zusammenhang mit den Praktiken Implementierung und Optimierung zentral.

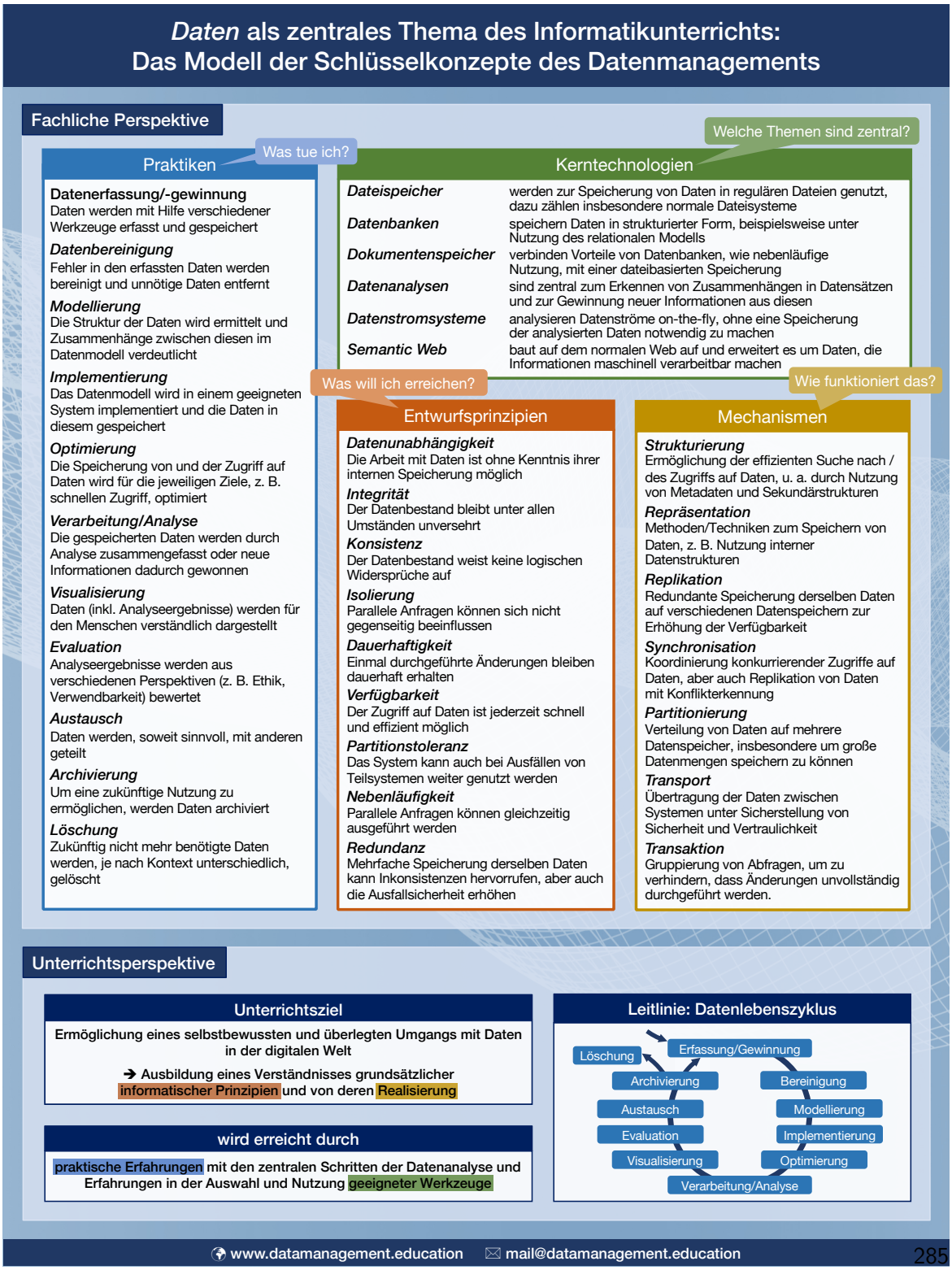
Anknüpfungspunkte in der Informatik

- **Parallelisierung:** Auch bei der Parallelisierung anderer Aufgaben tritt ein den Transaktionen ähnliches Konzept zutage: Beispielsweise gibt es bei der parallelen Programmierung kritische Abschnitte, deren Anweisungen nicht unterbrochen werden dürfen, d. h. dass diese, wie Transaktionen, nur gemeinsam oder gar nicht ausgeführt werden.
- **Rechnerkommunikation:** In der Rechnerkommunikation tritt ein ähnliches Konzept auf, wenn nach einer Anfrage an einen anderen Rechner nur Teile der Antwort übermittelt werden können, da dann entweder versucht werden muss, diese Teile erneut anzufordern oder die Anfrage im Gesamten als fehlerhaft verworfen und neu ausgeführt werden muss.

Begründung der Einordnung

Transaktionen stellen ein Konzept von Datenmanagementsystemen dar, das im Hintergrund von diesen Systemen verwendet wird und auf technischer Ebene die Erfüllung verschiedener Entwurfsprinzipien ermöglicht. Es ist daher für die Funktionsweise von Datenmanagementsystemen aus technischer Sicht zentral und kann somit als Mechanismus des Datenmanagements betrachtet werden.

Anhang D: Poster zu den Schlüsselkonzepten des Datenmanagements



Anhang E: Untersuchung der Inhalte der Data Science

Die im Folgenden skizzierte Studie zur Ermittlung der Inhalte der Data Science wurde zur Schaffung einer Grundlage für die Entwicklung des Data-Literacy-Kompetenzmodells (vgl. Kapitel 9) durchgeführt. Dabei war nicht das Ziel, eine alleinstehende Charakterisierung der Data Science zu entwickeln, die diese Wissenschaftdisziplin vollumfänglich beschreiben kann, sondern eine wichtige Grundlage für dieses Kompetenzmodell zu schaffen. Entsprechend wurde die Data Science nur so tiefgehend wie nötig und daher beispielsweise auch weniger detailliert als das Fachgebiet Datenmanagement betrachtet. Die folgende Beschreibung ist weitgehend identisch mit dem dazu veröffentlichten Konferenzbeitrag *(Grillenberger und Romeike, 2018c)*, der die Untersuchung ausführlicher darstellt.

Die heute oft als eigene Disziplin betrachtete Datenwissenschaft wurde insbesondere im Rahmen des EDISON-Projekts bereits durch einen *Body of Knowledge (Demchenko, Belloum und Wiktorski, 2017)* und ein Kompetenzframework *(Demchenko, Manieri und Belloum, 2017)* charakterisiert. Dabei wurde der Untersuchungsschwerpunkt jedoch nicht auf fachlich fundierte Arbeiten gelegt, stattdessen wurde eine externale Sicht auf das Fachgebiet eingenommen und insbesondere untersucht, welche Erwartungen bzw. Anforderungen Unternehmen in Stellenanzeigen für Datenanalysten stellen. Um eine fachliche Fundierung aus informatischer Perspektive sicherzustellen, kann diese Arbeit daher nicht herangezogen werden, ohne deren Validität zuvor gesondert sicherzustellen. Stattdessen werden als Basis für die Entwicklung des Data-Literacy-Kompetenzmodells im Folgenden die Inhalte der Data Science unter Berücksichtigung der wissenschaftlichen bzw. fachlichen Perspektive ermittelt. Dazu wurden die Modulhandbücher bereits etablierter Data-Science-Studiengänge umfassend hinsichtlich der thematisierten Inhalte untersucht. Es wird erwartet, dass diese, von Experten gestalteten, Dokumente die als zentral erachteten Inhalte klar darstellen.

E.1 Voruntersuchung und Ziel der Untersuchung

Im Rahmen einer ersten Voruntersuchung wurden einzelne Modulhandbücher gesichtet, um die Machbarkeit der angestrebten Analyse zu evaluieren. Im Allgemeinen stellen diese Modulhandbücher unter anderem die in den Modulen enthaltenen Inhalte aber auch angestrebte Kompetenzen dar. Es hat sich jedoch gezeigt, dass die Kompetenzen insbesondere bei Betrachtung verschiedener Studiengänge, aber teils auch innerhalb eines Studiengangs, stark unterschiedlich formuliert und detailliert sind. Entsprechend wäre eine Betrachtung der vermittelten Kompetenzen aufgrund der notwendigen Interpretation stark subjektiv geprägt und kaum valide. Gleichzeitig hat sich jedoch auch gezeigt, dass die Inhalte, auf denen im Folgenden der Fokus liegen soll, aus den Modulhandbüchern klar deutlich werden.

Um eine internationale Validität der Studie sicherzustellen, wurden bereits im Rahmen der Vorstudie deutschsprachige und internationale Studiengänge kontrastiert: Dabei zeigte sich, dass, vermutlich aufgrund unterschiedlicher rechtlicher Rahmenbedingungen, die charakteristischen Dokumente nicht nur einen sehr unterschiedlichen Aufbau haben, sondern sich auf im Detailgrad teils stark unterscheiden. Gleichzeitig waren jedoch kaum Unterschiede in den inhaltlichen Schwerpunktsetzungen erkennbar. Um eine möglichst nachvollziehbare und objektive Methodik anwenden zu können, wird der Fokus dieser Untersuchung auf Studiengänge im deutschsprachigen Raum gelegt, bei denen eine relative ähnliche Ausgestaltung der Dokumente zu erwarten ist. Um die Validität der Ergebnisse auch außerhalb des deutschsprachigen Raums zu überprüfen, werden die Ergebnisse der Untersuchung jedoch im Nachgang mit internationalen Studiengängen kontrastiert.

Die zentrale Forschungsfrage dieser Untersuchung lautet daher basierend auf den Vorüberlegungen: *Welche zentralen Inhalte charakterisieren Data-Science-Studiengänge im deutschsprachigen Raum?*

E.2 Untersuchungsmethode: Qualitative Inhaltsanalyse

Um dieser Frage nachzugehen, wurde ein empirischer Ansatz basierend auf einer qualitativen Inhaltsanalyse nach *Mayring (2010)* gewählt. Diese erlaubt u. a. einen Literaturkanon systematisch zu explorieren und basierend auf diesem ein Kategoriensystem aufzubauen. Dieses entspricht hier der angestrebten inhaltlichen Charakterisierung der Data Science. Obwohl dieses Kategoriensystem im Allgemeinen sowohl induktiv, aus dem Material heraus, oder deduktiv, auf Basis bereits existierender Arbeiten, aufgebaut werden kann, ist zur Erreichung des hier angestrebten Ziels der induktive Ansatz klar zu bevorzugen. Auf diese Weise kann das analysierte Material adäquat repräsentiert werden, ohne dass es durch Einordnung in ein bereits 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 der Analyse, auch die Kodiereinheit und ein Analysekriterium 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.

E.2.1 Festlegung der analysierten Materialien

Als Basis für die Analyse wurden Modulhandbücher verschiedener Data-Science-Studiengänge aus dem deutschsprachigen Raum ausgewählt. Um einen Überblick über die Studiengänge in diesem Bereich zu gewinnen, wurde der Hochschulkompass[70] zum Analysezeitpunkt (März 2018) herangezogen. Unter diesen sind jedoch verschiedene Studiengänge, die Themen der Data Science nur am Rande anschneiden und andere Schwerpunkte setzen, beispielsweise Informatikstudiengänge die lediglich ein bis zwei Data-Science-Module verpflichtend beinhalten oder die nur die Möglichkeit bieten, sich in einigen Wahlpflichtmodulen auf Data Science zu spezialisieren. Um eine Beeinflussung der Ergebnisse durch diese zu vermeiden, wurden solche Studiengänge nicht einbezogen.

Innerhalb der so 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- oder Wahlmodule blieben 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 konnten Studiengänge der folgenden Hochschulen als Basis für die Analyse berücksichtigt werden: Beuth Hochschule Berlin, Hochschule Darmstadt, Technische Universität Dortmund, Universität Jena, Universität München / LMU, Universität Mannheim[71], Hochschule Albstadt-Sigmaringen, Universität Salzburg, Hochschule der Medien Stuttgart (alle M. Sc.), Universität Marburg, Universität Stuttgart (beide B. Sc.).

E.2.2 Festlegung der Kodiereinheit und Analysekriterien

Die Kodiereinheit wurde auf semantische Weise so definiert, dass jede kodierte Einheit, unabhängig von ihrer Länge im Text, sich nur jeweils auf genau einen inhaltlichen Aspekt beziehen soll. Als Auswahlkriterium wurde festgelegt, dass alle in den Dokumenten genannten Inhalte betrachtet werden. Da der Schwerpunkt der Analyse auf den Inhalten lag, die die Data Science charakterisieren, wurden solche Aspekte, die den allgemeinen informatischen oder mathematischen Grundlagen zuzuordnen und nicht Data-Science-spezifisch sind, nur in geringem Detailgrad erfasst. Somit wurden beispielsweise *endliche Automaten* und *elementare Statistik* nicht im Detail berücksichtigt, sondern nur die Kategorien *Statistik* bzw. *Theoretische Informatik* eingeführt. Hingegen wurde beispielsweise die *Klassifikation*, die in der Datenanalyse eine wichtige Rolle spielt und in der Betrachtung in

[70] `http://www.hochschulkompass.de`, zuletzt geprüft: 6.3.2018

[71] Der *Mannheim Master in Data Science* ist in den folgenden Darstellungen nicht erkennbar, da keine Pflichtmodule definiert wurden und somit im Sinne der festgelegten Kriterien kein Modul berücksichtigt werden konnte.

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

E.3 Analyse der Dokumente und Strukturierung der Ergebnisse

Im Anschluss wurden die ausgewählten Dokumente unter Berücksichtigung der festgelegten Kriterien systematisch analysiert. Statt direkt ein hierarchisches Kategoriensystem aufzubauen, wurde, wie zuvor beschrieben, eine Liste aller genannten Inhalte erstellt. Es wurden jedoch, soweit dies bereits eindeutig ersichtlich war, Kodierungen vermieden, die zu detailliert sind: Beispielsweise wurden bei gemeinsamer Nennung einer großen Anzahl verschiedener Methoden zum überwachten Lernen nicht alle einzeln aufgenommen, sondern gesammelt unter einem 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 genannt wurden, sondern in unterschiedlichen Bereichen oder Dokumenten. Die dadurch möglichen Zusammenfassungen wurden entsprechend 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.

Um diese Überschneidungen zu beheben und die letztlich ausgewählten Begriffe auf ähnliches Niveau zu bringen, wurde diese Liste strukturiert und zusammengefasst. Dazu wurde entschieden, dass auf der obersten Ebene nur relativ abstrakte Begriffe genannt werden sollen, die die großen Themenbereiche der Data Science repräsentieren. Zusätzlich wurden die beiden Bereiche *informatische Grundlagen*, *mathematische Grundlagen* eingeführt, unter denen alle informatischen Aspekte bzw. mathematischen Grundlagen zusammengefasst werden, die eher übergreifend bzw. nicht speziell der Data Science zugehörig sind, jedoch der Vollständigkeit halber nicht unerwähnt bleiben sollen. Auf den tieferen Ebenen wurden jeweils Begriffe unter einem passenden Oberbegriff zusammengefasst, die von ihrer Bedeutung zusammengehörig sind und auch häufig gemeinsam genannt wurden, jedoch nur, wenn durch die Zusammenfassung keine relevanten Details verloren gehen. Somit wurden beispielsweise die Methoden der Datenanalyse in vier Kategorien subsumiert: *Methoden des unüberwachten Lernens* (darunter insbesondere *Clustering, Assoziation*), *Methoden des überwachten Lernens* (darunter insbesondere *Klassifikation, Entscheidungsbäume, Regression*), *Komplexere Methoden* (darunter beispielsweise *Neuronale Netze*) sowie *Verknüpfung von Methoden/Ensemble-Learning*. Auf die prinzipiell mögliche weitere Zusammenfassung, beispielsweise unter dem Begriff *Methoden der Datenanalyse*, wurde bewusst verzichtet, da dadurch die Unterscheidung und deren unterschiedliche Zwecke und Charakteristika in den Hintergrund rücken würde 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 die 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 Tabelle E.1 zusammen mit der Abdeckung in den verschiedenen Studiengängen dargestellt.

E.4 Diskussion der Ergebnisse

Das entstandene Kategoriensystem beschreibt die Data Science aus informatischer Sicht durch die vier großen Bereiche *Datenanalyse und Maschinenlernen*, *Big Data*, *Datenschutz, Ethik* und *Datenspeicher*. Hinzu kommen die *informatischen* und *mathematischen Grundlagen*. Ohne detaillierter auf die eigentlichen Inhalte einzugehen, kann durch die ermittelten Bereiche bereits wesentlich zur Charakterisierung der Data Science beigetragen 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.

Die mathematischen Grundlagen wurden in dieser Analyse zwar nur oberflächlich untersucht wurden, trotzdem 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 der Data Science würde jedoch den Rahmen der 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. In dem Bereich der Grundlagen besteht daher zwar ein grundlegendes Fundament, über das große Einigkeit herrscht, aber gleichzeitig eine breite Vielfalt an potenziellen Inhalten die, je nach Studienort, von einem Data Scientist beherrscht werden sollen.

Inhaltliche Charakterisierung der Data Science. Neben den Grundlagen konnten vier Inhaltsbereiche der Data Science ermittelt werden. Diese umfassen *Datenspeicher*, insbesondere Aspekte des Fachgebiets Datenbanken bzw. Datenmanagement wie (relationale) Datenbanken, Datenmodellierung und Abfragesprachen, den Bereich *Datenschutz und Ethik*, der bei Datenanalysen aus gesellschaftlicher aber auch informatischer Sicht eine wichtige

	HdM Stuttgart	Univ. Salzburg	HS Albst.-Sigmar.	Univ. Stuttgart	LMU München	Univ. Marburg	Univ. Jena	TU Dortmund	HS Darmstadt	Beuth HS Berlin	Anz. Nennungen	Prozent 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%

Tabelle E.1: Kategoriensystem zur Beschreibung der Data Science zusammen mit der Abdeckung in den analysierten Studiengängen.

Rolle spielt, sowie *Datenanalyse und Maschinenlernen* und *Big Data*, die insbesondere Aspekte beinhalten, die in der Informatik sonst nur am Rande thematisiert werden und in der Data Science eine neue Bedeutung erlangen.

Insbesondere *Datenanalyse und Maschinenlernen* sind in der Datenwissenschaft zentral, dieser Bereich wird 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 die des *überwachten Lernens* (80 %), zu denen u. a. Klassifikation, Entscheidungsbäume und Regression zählen. Durch die stark analytisch geprägte Sichtweise dieses Bereichs stellt er gleichzeitig den Bezug zur hauptsächlichen Aufgabe von Datenwissenschaftlern dar, die sich weniger um die konkrete Datengewinnung und 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 Zeiträume stellt eine wesentliche Herausforderung der 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 das Themenfeld *Big Data* jedoch auch anderen Bereichen der Informatik, insbesondere dem eher auf Datenspeicherung ausgerichteten Datenmanagement, zugeordnet werden. Dies liefert eine mögliche Erklärung für die vergleichsweise geringere Repräsentation in den Studiengängen, die entsprechende Kenntnisse möglicherweise als bereits vorher zu erwerbende Grundlagen voraussetzen.

E.5 Internationaler Vergleich

Zum internationalen Vergleich der Ergebnisse wurden drei Studiengänge ausgewählt, die klar definierte Inhalte vorweisen und somit für den Vergleich geeignet scheinen: 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 die entsprechenden Modulhandbücher bzw. andere Dokumente, die diese beschreiben, in einer qualitativen Inhaltsanalyse untersucht. Dabei lagen dieselben Grundsätze wie zuvor zugrunde. Statt jedoch induktiv ein Kategoriensystem aufzubauen, wurde die zuvor aufgebaute Charakterisierung deduktiv an die Materialien herangetragen, sodass ein Vergleich möglich wurde. Die Ergebnisse dieser Analyse sind in Tabelle E.2 abgebildet[72].

[72] Zu den Ergebnissen des Berkeley-Studiengangs ist anzumerken, dass keine Information über die dabei thematisierten Datenanalysemethoden auffindbar war, sodass hier eine detaillierte Einstufung nicht möglich war. Datenanalysemethoden im Allgemeinen wurden jedoch explizit genannt, was in Tabelle E.2 durch die spezielle Markierung berücksichtigt wurde.

Bei der Auswertung der drei Studiengänge wurden keine Begriffe ermittelt, die in die zuvor entwickelte Charakterisierung nicht sinnvoll eingeordnet werden konnten. Es kann daher eine hohe Vollständigkeit dieser Charakterisierung angenommen werden. Gleichzeitig zeigt sich, dass diese Studiengänge zwar nicht die identischen Schwerpunkte legen, die Abweichungen aber relativ gering sind und geringfügig unterschiedlichen Schwerpunktsetzungen sowie (insbesondere hinsichtlich der Vorkenntnisse) dem unterschiedlichen Bildungssystem geschuldet sein dürften. 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.

	Berkeley Univ.	Columbia Univ.	Univ. of Washington	Anzahl Nennungen	Prozent der Studiengänge
Mathematik					
Analysis					0%
Lineare Algebra					0%
Statistik	×	1		2	66,7%
Informatische Grundlagen					
Algorithmen und Datenstrukturen		×		1	33,3%
Betriebssysteme					0%
IT-Security					0%
Programmierung	×	1		2	66,7%
Rechnerarchitektur					0%
Rechnerkommunikation					0%
Software Engineering					0%
Theoretische Informatik					0%
Verteilte Systeme					0%
Datenanalyse und Maschinenlernen					
Analyseprozess	×			1	33,3%
Datenvorverarbeitung		×	1	2	66,7%
Ensemblelernen	×	1	1	3	100%
Komplexere Methoden	×	1		2	66,7%
Methoden überwachtes Lernen	o	×	1	2	66,7%
Methoden unüberwachtes Lernen	o		×	1	33,3%
Modellauswahl, -beurteilung, -anpassung		×		1	33,3%
Text Mining					0%
Visualisierung			×	1	33,3%
Big Data					
Algorithmen und Methoden der Big-Data-Verarbeitung		×		1	33,3%
Big-Data-Architekturen und -Systeme	×	1	1	3	100%
Prinzipien der Big-Data-Analyse	×		1	2	66,7%
Webdaten		×		1	33,3%
Datenschutz, Ethik					
Datenschutz			×	1	33,3%
Ethische Aspekte			×	1	33,3%
Sicherheit					0%
Datenspeicher					
Daten(bank)modellierung		×	1	2	66,7%
Datenbanken (insb. relational)	×	1	1	3	100%
Datenbanksprachen		×	1	2	66,7%

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

Anhang F: Unterrichtskonzept „Datenanalyse und Vorhersage“

Datenanalyse und Vorhersage mit Klassifikationsbäumen

Ein Unterrichtskonzept für die Sekundarstufe II

Lehrerversion / Gesamtkonzept
Stand: 31. Mai 2018

Andreas Grillenberger
Friedrich-Alexander-Universität Erlangen-Nürnberg
Professur für Didaktik der Informatik
Kontakt: andreas.grillenberger@fau.de

Erstellt mit Unterstützung von *StRin RS Anne-Katrin Jäger*

Einführung für die Lehrkraft

Ziele

Im Rahmen des Unterrichts soll den Schülern die Möglichkeit gegeben werden, zu erkennen wie die heute allpräsenten korrelationsbasierten Datenanalysen funktionieren. Diese versuchen, aus einem großen Berg an Daten Informationen zu gewinnen, ohne dass der konkrete Weg der Analyse vorher klar ist.

Es wird dabei angestrebt, dass die Schüler einen kritischen Blick auf Datenanalysen entwickeln und sich der Grenzen dieser Analysen bewusst werden.

Folgende Lernziele werden daher angestrebt: Die Schüler…

- erklären anhand eines Beispiels den Unterschied zwischen Kausalität und Korrelation bezogen auf Datenanalysen.
- beschreiben den Ablauf einer typischen korrelationsbasierten Datenanalyse (ggf. unter Zuhilfenahme eines Diagramms).
- beschreiben das Konzept „Klassifikationsbaum“ und erstellen einen solchen für gegebene Regeln.
- erstellen anhand eines „Klassifikationsbaums“ eine Prognose für einen Datensatz.
- beurteilen Analysen hinsichtlich ihrer Qualität anhand der auftretenden Fehl-Zuordnungen.
- führen einfache korrelationsbasierte Datenanalysen mit einem geeigneten Werkzeug am Computer selbst durch.
- beurteilen reale und fiktive Beispiele von korrelationsbasierten Datenanalysen hinsichtlich ihres Nutzens und ihrer Gefahren.

Zum Material

 Alle im Folgenden genannten Aufgaben sind am Ende des Konzepts auch in einer Schülerversion zu Arbeitsblättern zusammengefasst.

Überblick

Zeitansatz: je nach Detailgrad werden ca. zwei bis vier Doppelstunden veranschlagt.

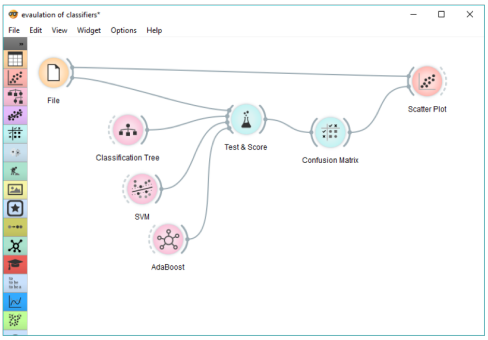
- Einblick in die Nutzung von Datenanalysen anhand eines Realweltbeispiels; Versuch dieses zu erklären, indem mögliche Kausalzusammenhänge diskutiert werden
- Begrifflichkeit: Kausalität vs. Korrelation als „mit gesundem Menschenverstand erklärbar“ vs. „unerklärbar, aber anhand von Daten richtig erscheinend“; Diskussion der damit einhergehenden Gefahr durch Fehleinstufungen
- Überblick über den Datenanalyseprozess und Feststellung wo Kausalität vs. Korrelation dabei zentral ist
- Händische Durchführung einer einfachen korrelationsbasierten Analyse zum Erlernen der Grundsätze
  - Finden von Regeln im Datensatz
  - Darstellung der Regeln als Klassifikationsbaum
  - Nutzung des Klassifikationsbaums zur Prognose von Attributen weiterer Datensätze
  - Diskussion der Qualität
- Durchführung am Computer zum Erkennen des Potentials und der Grenzen bei Verwendung größerer Datensätze
  - Aufstellen einer Vermutung, welche Attribute des Datensatzes für die Bestimmung des vorherzusagenden Attributs relevant sind
  - Nutzung eines Datenanalysewerkzeugs zur Erstellung eines Klassifikationsbaums; Verifikation der vorherigen Hypothesen über relevante Attribute

[1] `https://creativecommons.org/licenses/by-nc-sa/4.0/`

  - Erstellen einer automatisierten Vorhersage anhand des erstellten Klassifikationsbaums; erste Einschätzung der Ergebnisse
  - Nutzung einer Confusion Matrix um Fehler in den Vorhersagen systematisch zu erkennen; Beurteilung der Analysequalität
  - Überprüfung von Möglichkeiten zur Verbesserung der Analysequalität
- Diskussion verschiedener fiktiver und realer Beispiele zur Verwendung von Datenanalysen hinsichtlich ihres Nutzens und ihrer Gefahren

Verwendetes Datenanalysewerkzeug

Es wird das Datenanalysetool Orange3 genutzt, das an der Universität Ljubljana entwickelt wird. Dieses erlaubt einen einfachen Zugang, da die Datenanalysen nicht durch textbasierte Programmierung durchgeführt, sondern in einer Art Datenflussdiagramm aufgebaut werden. Damit sind alle Analysemöglichkeiten direkt sichtbar, andererseits gibt die visuellen Orientierung auch eine bessere Übersicht über den Ablauf der Analyse. Trotzdem empfiehlt es sich, eine reduzierte Version des Werkzeugs zu verwenden, die einen Großteil der möglichen Funktionen ausblendet, um den kognitiven Aufwand zur Erfassung der Möglichkeiten dieses Tools möglichst gering und die Bedienoberfläche möglichst übersichtlich zu halten. Eine solche Version kann durch die DDI-FAU (Kontakt: `mailto:andreas.grillenberger@fau.de`) zur Verfügung gestellt werden.



Arbeitsblatt 1 (L): Logik oder scharfes Hinsehen?

Zum Einstieg wird die Nutzung eines für die Schüler interessanten und gleichzeitig Fragen aufwerfenden Realweltbeispiels vorgeschlagen. Es bietet sich hier beispielsweise die folgende Geschichte an, die sich in den USA ereignet haben soll (z. B. präsentieren mit Beamer o. Ä.):

> *Dem US-Einzelhandelsriesen Target gelang es durch die Analyse herauszufinden, welche Kundinnen schwanger sind. Duhigg schreibt, dies sei für das Unternehmen sehr wichtig gewesen, denn werdende Eltern seien so etwas wie der „Heilige Gral" für Unternehmen wie Target. In einer Schwangerschaft änderten sich die Gewohnheiten, und wer vorher keine gute Kundin des Einzelhändlers gewesen sei, könne es danach werden - wenn man ihr zu richtigen Zeit die richtige Werbung zusendet.*
> *Die Statistiker von Target, so berichtet es Duhigg, identifizierten etwa 25 Produkte, die darauf hinweisen, dass Kundinnen schwanger sind. Genauer gesagt, wenn sie sich im zweiten Trimester ihrer Schwangerschaft befinden. Denn zu diesem Zeitpunkt fingen sie an, sich neue Sachen zu kaufen, und Target schickte ihnen dann schon Werbung. Zu den identifizierten Produkten gehörten parfümfreie Körperlotion, große Mengen an Watte und Nahrungsergänzungsmittel wie Kalzium, Magnesium und Zink. Target habe in der Kundendatenbank gesucht und Zehntausende Frauen gefunden, die mit großer Wahrscheinlichkeit bald Mutter würden.*
> *Der Autor Duhigg berichtet darüber, wie die Werbung für Schwangerschaftsprodukte den Vater einer Tochter in Rage versetzte. Er beschwerte sich in einem Target-Markt in der Nähe von Minneapolis darüber, dass seine Tochter - noch ein Teenager - Werbung für Babykleidung erhalten habe. Ob man sie dazu animieren wolle, schwanger zu werden, fragte er den Manager des Ladens. Dieser entschuldigte sich, doch als er später noch einmal sein Bedauern zum Ausdruck bringen wollte und den Vater anrief, stellte sich heraus, dass die Tochter wirklich schwanger war. Target hatte es nur vor dem Vater der jungen Frau gewusst.*
>
> — *Frankfurter Neue Presse, 13.09.2014*

Verfügbar unter: `http://www.fnp.de/art673,1029989`

Dieser Unterrichtseinstieg wirft die Frage auf, wie der Supermarkt Target die entsprechenden Produkte erraten konnte und wie solche Analysen im Allgemeinen funktionieren. Dies kann im Unterrichtsgespräch diskutiert werden.

Während bei diesem Beispiel noch vermutet werden kann, dass findige Personen sich die Kriterien zur Erkennung einer Schwangerschaft überlegt und anhand der Kundendaten überprüft haben, zeigt sich spätestens im Folgenden zweiten Beispiel, dass dies nicht immer auf diese Weise funktionieren kann:

Für Schüler/-innen, z. B. auf Arbeitsblatt

Im Unterricht hast du bereits einen Artikel darüber gesehen, wie Daten heute im Einzelhandel verwendet werden, um Kunden auf sie zugeschnitte Werbung präsentieren zu können. Onlineshops gehen heute jedoch schon weiter und versuchen, ihren Kunden viele Produkte möglichst schnell liefern zu können:

> *Noch bevor ein Kunde überhaupt den Button "Kaufen" anklickt, soll die für ihn passende Ware schon auf dem Weg in Richtung seiner Wohnung sein. Dem Versandhändler Amazon wurde ein Patent zugesprochen, das einen „vorausschauenden Versand“ („anticipatory shipping“) ermöglichen soll. Das heißt: Bestimmte Waren werden schon einmal an ein Versandzentrum geschickt, in dessen Nähe sich ein oder mehrere Kunden höchstwahrscheinlich für das Produkt interessieren. Wird es dann schließlich bestellt, ist es umso schneller beim Empfänger.*
>
> *— Spiegel Online, 18.01.2014*

Verfügbar unter: `http://www.spiegel.de/netzwelt/web/a-944252.html`

Eine weitere Kontextualisierung kann durch Beispiele wie Same-Day-Delivery, wie sie verschiedene Onlinehändler anbieten, geschehen.

Den Schülern kann nun eine erste Aufgabe gegeben werden, in der das Ziel sein sollte, zu erkennen, dass diese Art der „Voraussage“ von Kundenverhalten nicht mehr rein auf logischen Schlüssen geschehen kann. Die Aufgabe könnte daher wie folgt lauten:

Aufgabe 1

Um herauszufinden, was ein Kunde als nächstes bestellen könnte, müssen die Versandhändler umfangreiche Daten über ihre Kunden sammeln und analysieren.

a) Was wissen Onlinehändler über ihre Kunden? Woher haben diese die jeweilige Information?

Information über den Kunden	**Quelle**
Vor- und Nachname	*Registrierung*
Adresse	*Registrierung*
Geburtsdatum	*Registrierung*
Beliebte Artikel	*Einkäufe*
Bekannte	*Versandadressen*
Aufenthaltsorte	*Versand- & IP-Adressen*
...	...

b) Wahrscheinlich sind nicht alle Informationen, die ein Onlinehändler über seine Kunden hat auch wichtig, wenn er herausfinden möchte, welchen Artikel der Kunde als nächstes bestellen könnte. Markiere in der Tabelle oben die Zeilen, von denen du denkst, dass sie für diese Zweck wichtig sind, indem du ein + neben die wichtigen Zeilen machst.

Die Ergebnisse können nicht auf Korrektheit überprüft werden, geben den Schülern aber die Gelegenheit sich Gedanken über das Konzept zu machen und zu erkennen, dass es eigentlich nur wenige Daten gibt, die dafür wirklich hilfreich erscheinen. Erst bei Betrachtung der Gesamtmenge an Daten können aus diesen anscheinend relevante Schlüsse gezogen werden. Ziel ist es, über mögliche Attribute zu diskutieren. Das ist gut, denn es

zeigt, dass keine logischen Schlüsse gezogen werden können und leitet damit hin zu korrelativen Analysen. Auf dieser Basis können dann korrelative Analysen besprochen werden. Die Ergebnisse können im Lückentext gesichert werden:

Aufgabe 2
Fülle folgenden Lückentext aus: Es gibt bei der Datenanalyse zwei Möglichkeiten, wie wir Vorhersagen treffen können:

1. Wenn wir bereits etwas über die zu analysierenden Daten wissen, dann können wir uns erklären wie etwas funktioniert und damit Schlussfolgerungen ziehen. Es gibt also logische Zusammenhänge, sog. Kausalzusammenhänge, die wir zur Vorhersage nutzen können.
   Beispiel:
   `WENN ein Kunde in den letzten 5 Einkäufen Chips gekauft hat, DANN wird er auch beim nächsten Mal welche kaufen.`

2. In anderen Bereichen erkennen wir keinerlei logische Zusammenhänge. Stattdessen können wir nach Mustern in den Daten suchen. Diese liefern uns auch Zusammenhänge, wir können sie uns aber oft nicht erklären. Solche Zusammenhänge bezeichnen wir als korrelative Zusammenhänge.
   Beispiel:
   `WENN ein Kunde den Artikel X gekauft hat und er in Y wohnt und mindestens 35 Jahre alt ist, DANN wird er auch Z kaufen.`

Kausalzusammenhänge helfen uns zwar dabei Dinge zu verstehen, sie sind aber für Datenanalysen oft relativ wenig interessant: Sie sind oft offensichtlich und bekannt, sodass sie nur wenig neue Informationen hervorbringen. Wir können uns aber logisch erklären, dass sie richtig/wahr sind. Die korrelativen Zusammenhänge sind daher oft spannender, da sie neue Informationen eröffnen. Sie haben aber den Nachteil, dass sie nicht unbedingt logisch nachvollziehbar sind: Wie genau Wohnort und Alter das Kaufverhalten prägen, können wir uns meist nicht logisch erklären. Außerdem müssen wir sie erst finden, was relativ schwierig ist.

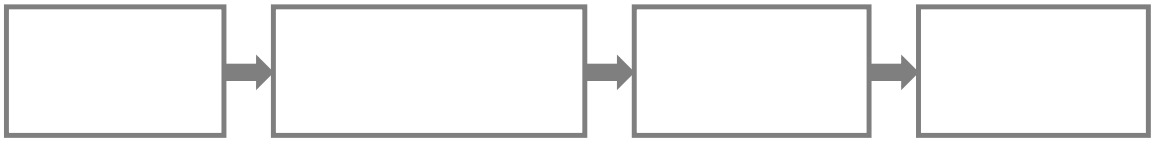
Am Ende dieser Unterrichtsstunde haben die Schülerinnen und Schüler bereits einen ersten Einblick in die Ziele und Problematik der Datenanalyse erhalten. Auf dieser Basis wird in der nächsten Unterrichtsstunde eine erste händische Datenanalyse durchgeführt.

Arbeitsblatt 2 (L): Händische Datenanalyse (Teil I)

Auf dieser Basis kann mit den Schülern der Prozess der Datenanalyse thematisiert werden, dabei bietet es sich an, ein Modell des Prozesses vorzustellen, das auch als Art Advance Organizer dienen kann und den weiteren Lernprozess strukturiert. Dieser kann an der Tafel gemeinsam erarbeitet werden, beispielsweise unter Nutzung von Karteikarten für die vier Teilprozesse:

Zur Erlangung eines grundlegenden Verständnisses über die Datenanalysen wird ein erster Musterdatensatz zur Verfügung gestellt, der in einem Durchgang durch den gesamten Prozess händisch analysiert wird. Es wird dabei im Bereich der Erstellung des Vorhersagemodells versucht, sowohl die kausalitäts- als auch die korrelationsbasierte Datenanalyse zu thematisieren und deren Unterschiede herauszustellen.

Aufgabe 1
Im Unterricht wurde bereits gemeinsam erarbeitet, sie eine Datenanalyse abläuft. Vervollständige den folgenden Lückentext und das folgende Ablaufmodell:
Als erster Schritt der Datenanalyse, müssen die Daten erfasst/gewonnen und gespeichert werden. Aus diesen Daten wählt man sich üblicherweise eine kleine Teilmenge aus, um aus dieser das Vorhersagemodell zu erstellen, d. h. um Regeln zu finden, die die Vorhersage der gesuchten Eigenschaften ermöglichen. Diese Regeln können dann genutzt werden, um die Vorhersagen zu erstellen. Als letzter Schritt jeder Datenanalyse sollte die Bewertung der Ergebnisse erfolgen, mit dem Ziel eine möglichst gute Qualität der Ergebnisse sicherzustellen.



Der erste Schritt für die Schüler ist die Erstellung eines geeigneten Vorhersagemodells, dies erfolgt in der Folgenden Aufgabe:

Aufgabe 2
Ein Onlineshop hat über seine Kunden verschiedene Daten gesammelt und möchte nun seine Kunden durch den Versand von individuellen Gutscheinen zu weiteren Käufen anregen. Dazu will er herauszufinden, welche Produktkategorie für jeden Kunden jeweils besonders interessant ist.
Der Shop hat bereits folgende Daten über jeden seiner Kunden gesammelt: `Alter, Familienstand, Anzahl der Kinder, präferierte Zahlungsart, Kategorien der letzten vier eingekauften Produkte (Film, Sport, Software, Elektronik, Kleidung, Musik, Bücher oder Auto).`
Um jedem Kunden einen Gutschein zu schicken, den dieser wahrscheinlich einlöst, möchte der Onlineshop herausfinden, welche Kategorie für den Käufer besonders interessant ist. Welche WENN-DANN-Regeln vermutest du, die dem Onlineshop dabei helfen könnten? *Hinweis: natürlich kannst du mehrere Bedingungen mit „und“ verknüpfen, z. B. „Kategorie 1 = Elektronik und Anzahl Kinder = 0“.*

- `WENN XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX DANN ist der nächste Kauf aus der Kategorie XXXXXXXXXXXXXXXXXXXX`
- `WENN XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX DANN ist der nächste Kauf aus der Kategorie XXXXXXXXXXXXXXXXXXXX`
- `WENN XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX DANN ist der nächste Kauf aus der Kategorie XXXXXXXXXXXXXXXXXXXX`

Die Schüler werden möglicherweise zu verschiedenen Regeln kommen, im Endeffekt wird sich aber herausstellen, dass zu wenig Information verfügbar ist, um sinnvolle und stichhaltige Kausalzusammenhänge zu erkennen. Als nächstes kann daher folgende Aufgabe gestellt werden:

Aufgabe 3

Nachdem die Datenwissenschaftler des Unternehmens erkannt haben, dass keine stichhaltigen Zusammenhänge in den bisher vorliegenden Informationen erkennbar sind, wurde entschieden, es anders zu versuchen: Der Onlineshop hat daher einige seiner Kunden befragt, was für sie als nächstes interessant ist. Dabei sind folgende Daten herausgekommen. Welche Zusammenhänge erkennt du in der unten dargestellten Tabelle?
Beispiel:
WENN Kauf 1 ein Film ist und mit Girokarte bezahlt wurde, DANN interessiert der Kunde sich als nächstes für Artikel der Kategorie Auto.
Kurz: „Kauf1"=„Film" und „bezahlt mit"=„Giro" ⇒ „Interesse"=„Auto"

Alter	**Kauf 1**	**Kauf 2**	**Kauf 3**	**Kauf 4**	**verheiratet**	**Kinder**	**bezahlt mit**	**Interesse**
25-50	Film	Software	Film	Sport	Ja	1	Giro	Auto
25-50	Elektronik	Musik	Film	Software	Nein	1	VISA	Bücher
<18	Film	Elektronik	Sport	Sport	Nein	0	Giro	Auto
<18	Film	Musik	Kleidung	Sport	Nein	0	Giro	Auto
18-25	Bücher	Musik	Film	Haushalt	Nein	1	Giro	Bücher
<18	Bücher	Film	Film	Bücher	Nein	0	VISA	Bücher
25-50	Film	Film	Sport	Sport	Ja	1	Giro	Auto
25-50	Musik	Film	Film	Spielzeug	Nein	1	Giro	Bücher
25-50	Musik	Musik	Film	Haushalt	Nein	1	Giro	Bücher
25-50	Elektronik	Musik	Bücher	Software	Ja	1	Master	Elektronik
25-50	Software	Elektronik	Film	Spielzeug	Nein	1	Master	Bücher
25-50	Film	Film	Sport	Sport	Ja	0	Master	Elektronik
25-50	Musik	Elektronik	Bücher	Elektronik	Ja	1	Master	Elektronik

- `WENN XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX DANN ist der nächste Kauf aus der Kategorie XXXXXXXXXXXXXXXXXXXX`
- `WENN XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX DANN ist der nächste Kauf aus der Kategorie XXXXXXXXXXXXXXXXXXXX`
- `WENN XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX DANN ist der nächste Kauf aus der Kategorie XXXXXXXXXXXXXXXXXXXX`

Die Schüler können hier verschiedene Regeln manuell aus den Daten herauslesen, die willkürlich wirken (und sind). Diese sind daher nicht logisch erklärbar, wie es für korrelationsbasierte Analysen oft üblich ist.

An dieser Stelle kann diskutiert werden, ob Regeln wie „kein Kind" ⇒ „Interesse" = „Elektronik" aufgenommen werden sollten – aus den Daten ergibt sich das, es kann jedoch vermutet werden, dass diese auf nur einem Datensatz basierende Regel wenig stichhaltig ist.

Damit diese Regeln besser nutzbar sind, werden sie üblicherweise als Klassifikationsbaum dargestellt. Fachlich handelt es sich dabei um einen (nicht zwingend binären) Entscheidungsbaum, dessen Blättern die getroffenen Entscheidungen darstellen. Eine Aufgabe zur Überführung der Regeln in einen Baum könnte wie folgt aussehen:

Arbeitsblatt 3 (L): Händische Datenanalyse (Teil 2)

Aufgabe 1

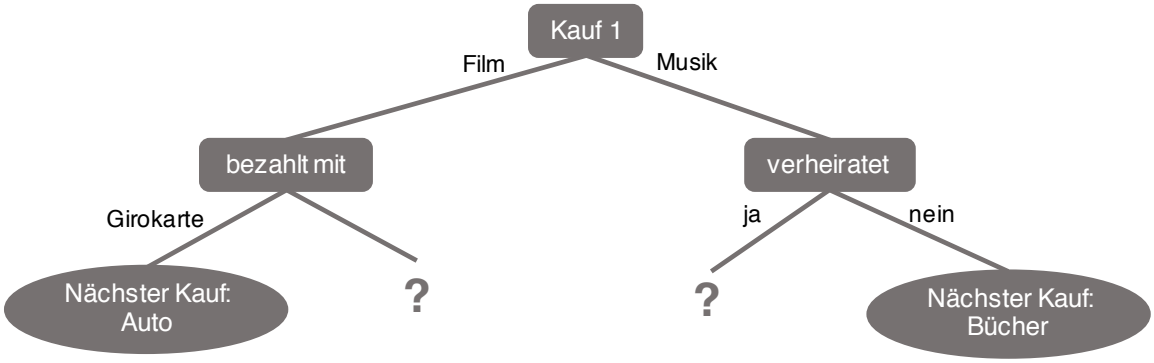
Wenn der Onlineshop nun Vorhersagen treffen will, dann sortiert er die Kunden in verschiedene Kategorien ein - dies nennt man „Klassifikation“.

Um eine größere Menge an Regeln einfacher durchschaubar und anwendbar zu machen, stellt man diese als „Klassifikationsbaum“ dar. Dieser Baum symbolisiert die Entscheidungen, die anhand der Regeln getroffen werden.
Beispiel: Die Regeln

1. *„Kauf 1“ = „Film“ und „bezahlt mit“ = „Giro“ ⇒ nächster Kauf = „Auto“*
2. *„Kauf 1“ = „Musik“ und nicht verheiratet ⇒ nächster Kauf = „Bücher“*

können als Klassifikationsbaum wie unten abgebildet dargestellt werden.

- Markiere im Baum den Weg, den du gehen musst, wenn du herausfinden möchtest, was ein Kunde als nächstes gekauft hat, der als „Kauf 1“ einen Film gekauft und mit einer Girokarte bezahlt hat.
- Wir kennen zusätzlich die folgende Regel: *Kauf 1 = „Film“ und „bezahlt mit“ ist „Mastercard“ und „verheiratet“ = „ja“ ⇒ nächster Kauf = „Elektronik“*
  Um diese im Baum zu berücksichtigen, musst du eine weitere Entscheidung ergänzen. Überlege dir, wo das sinnvoll ist und ergänze die Entscheidung.
- Überprüfe deine Ergänzung, indem du den Weg farbig markierst, den du durch den Baum gehen musst, um herauszufinden, was ein Kunde als nächstes kauft, der als „Kauf 1“ einen Film gekauft hat und mit „Mastercard“ bezahlt hat.



Mit diesem Klassifikationsbaum kann nun gut beschrieben werden, welche Regeln in den bisher bekannten Daten vorherrschen. Um eine Prognose zu treffen, müssen diese bekannten Informationen genutzt werden, um damit Informationen über einen Kunden vorherzusagen (eine Prognose zu generieren), die erst im Nachgang (wenn überhaupt) überprüft werden kann. Dazu kann den Schülern folgende Aufgabenstellung vorgelegt werden:

Aufgabe 2
Verwende den vorherigen Klassifikationsbaum, um zu entscheiden, an welcher Produktkategorie die folgenden Kunden wahrscheinlich als nächstes interessiert sind. Wenn diese Entscheidung anhand der beiden Regeln bzw. anhand des Baums nicht getroffen werden kann, schreibe ein **?** in das Feld „vsl. interessiert an“.

Alter	**Kauf 1**	**Kauf 2**	**Kauf 3**	**Kauf 4**	**verheiratet**	**Kinder**	**bezahlt mit**	**vsl. interessiert an**
25-50	Film	Film	Sport	Sport	Ja	1	Giro	*Lösung: Auto*
25-50	Musik	Elektronik	Sport	Sport	Nein	1	Giro	*Lösung: Bücher*
>50	Film	Musik	Kleidung	Sport	Nein	1	Giro	*Lösung: Auto*
<18	Film	Musik	Film	Haushalt	Ja	1	Master	*Lösung: Elektronik*
>50	Bücher	Software	Film	Sport	Ja	1	VISA	*Lösung: ?*

Im Beispiel wurde absichtlich eine Stelle eingebaut, an der die Zuordnung mehrdeutig ist: Bei diesem Kunden ist es nicht möglich, eine eindeutige Vorhersage zu treffen. Dies kann genutzt werden, um eine Diskussion einzuleiten, was in solchen Fällen geschehen soll – und wie gut diese Analyse überhaupt sein kann: Was verursacht ein einzelner konträrer Datensatz der noch dazu kommt? Wie können wir die Analyse verbessern? Wie wichtig ist in diesem Fall eine möglichst gute Analyse?...

Arbeitsblatt 4 (L): Datenanalyse am Computer

Natürlich werden solche Datenanalysen in real nicht händisch, sondern rechnergestützt durchgeführt. Es soll den Schülern daher auch die Möglichkeit gegeben werden, das übliche Vorgehen auszuprobieren und damit auch mit etwas mehr Daten zu arbeiten, als bei der händischen Analyse vorher.

Als Datenbasis kann beispielsweise ein Datensatz gewählt werden, der die Schülerdaten von über 600 Schülern aus Portugal enthält, und in dem anhand dieser die Endnote/-punktzahl der Schüler prognostiziert werden soll. Diese Analyse kann beispielsweise wie folgt kontextualisiert werden. Es wurde hier absichtlich ein Kontext gewählt, der die Schüler direkt betrifft und der bei diesen sehr umstritten sein dürfte, um zu demonstrieren, dass auch die Schüler selbst von solchen Analysen prinzipiell direkt betroffen sein könnten.

Für Schüler/-innen, z. B. auf Arbeitsblatt
Es wäre für eure Lehrerin sicherlich eine sehr praktische Sache, die Idee der Onlineshops zu nutzen, um eure Schulnoten vorherzusagen: Dann würde es ausreichen jedes Mal nur ein paar Arbeiten zu korrigieren und die Noten aller anderen „vorherzusagen“. Doch wie (gut) funktioniert das wirklich?

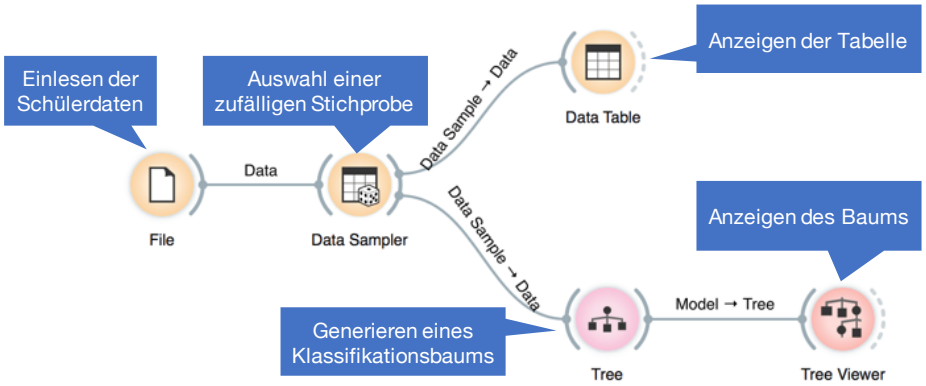
Diese Fragestellung soll im Folgenden mit den Schülern ausführlich thematisiert werden, indem anhand des vorliegenden Datensatzes eine Analyse durchgeführt wird, die es erlauben wird, grundlegende Rückschlüsse auf die Qualität und die Möglichkeiten solcher Datenanalysen zu ziehen. Es bietet sich dabei an, den gleichen Prozess wie vorher händisch durchlaufen, auch an dieser Stelle wieder aufzugreifen. Als erstes steht die Erstellung von Klassifikationsregeln und eines Klassifikationsbaumes an. Um sich in den vorgegebenen Datensatz hineinzudenken, wird den Schülern zuerst folgende Aufgabe an die Hand gegeben:

Aufgabe 1
Zunächst ohne Computer: Von welchen der in der Tabelle unten dargestellten Eigenschaften der Schülerinnen/Schüler vermutest du, dass sie für das „Erraten“ bzw. „Berechnen“ der Endnoten wichtig sind? Markiere diese in der folgenden Tabelle.

Attribut	**Beschreibung**	**relevant?**
Schule	Kürzel der Schule: “GP” oder “MS”	
Geschlecht	“M” oder “W”	
Alter	Zahlenwert	
Wohnumfeld	“urban” oder “ländlich”	
Familiengröße	“$\leq$3” oder “$>$3”	
Bildung Mutter	Grundschule; Hauptschule; Realschule/Gymnasium; Universität; keine	
Bildung Vater	vgl. Bildung Mutter	
Beruf Mutter	Gesundheitsbranche; Lehrerin; Hausfrau; Öffentlicher Dienst; sonstige	
Beruf Vater	vgl. Beruf Mutter	
Fahrzeit	Tägliche Fahrzeit des Schülers zur Schule: “<15min”; “15-30min”; “30-60min”; “$>$60min”	
Lernzeit	Zeit die der Schüler außerhalb des Unterrichts pro Woche zum Lernen aufwendet: “<2h”; “2-5h”; “5-10h”; “$>$10h”	
Wiederholte Schuljahre	Anzahl der Schuljahre die der Schüler wiederholt hat: 0; 1; 2 oder 3	
Unterstützung Familie	Ob der Schüler durch die Familie Unterstützung bei seinen schulischen Pflichten erhält: “Ja” oder “Nein”	
Nachhilfe	Ob der Schüler Nachhilfeunterricht nimmt: “Ja” oder “Nein”	
Außerunterrichtliche Schulaktivitäten	Nimmt der Schüler an Aktivitäten in der Schule außerhalb des Unterrichts teil: “Ja” oder “Nein”	
Internetzugang	Ob Zuhause ein Internetzugang für den Schüler nutzbar ist: “Ja” oder “Nein”	
familiäre Beziehungen	Als wie gut schätzt der Schüler seine familiären Beziehungen ein: “sehr schlecht”; “schlecht”; “mittelmaessig”; “gut”; “sehr gut”	
Freizeit	Menge an Freizeit: “sehr wenig”; “wenig”; “mittel”; “viel”; “sehr viel”	
Weggehen	Wie wichtig ist es dem Schüler; mit Freunden wegzugehen: “sehr wichtig”; “wichtig”; “mittel”; “unwichtig”; “sehr unwichtig”	
Gesundheit	Die Gesundheit des Schülers: “sehr schlecht”; “schlecht”; “mittelmaessig”; “gut”; “sehr gut”	
Abwesenheiten	Wie oft war der Schüler im aktuellen Schuljahr abwesend vom Unterricht: Zahlenwert	
Punkte 1	Punkte im ersten Test: 0 bis 20	
Punkte 2	Punkte im zweiten Test: 0 bis 20	
Punkte 3	Punkte im dritten Test (zu schätzen): 0 bis 20	

Nachdem die Schüler anhand der Aufgabe ihre Vermutungen geäußert und den Datensatz kennengelernt haben, kann nun eine Analyse am PC durchgeführt werden. Zur Reduzierung des Zeitaufwandes bietet es sich an, ein vorgefertigtes Analyseschema an die Hand zu geben, auf dem die Schüler aufbauen können. Eine erste Aufgabe kann dann die Arbeit mit dem Klassifikationsbaum sein:

Aufgabe 2
Nun werden wir das Ganze am Computer ausprobieren. Starte dazu am Computer das Programm „Orange3“ und lade das Projekt „Schulnoten“. Ein Teil der Analyse ist dort bereits vorbereitet:



Das Programm lädt also die Datei mit den Schülerdaten (*File*). Daraus wird ein kleiner Anteil der Daten (Standard: 30 %) ausgewählt (*Data Sampler*), der sozusagen den „korrigierten Arbeiten“ entspricht. Aus dieser Stichprobe werden automatisch Regeln gesucht und als Klassifikationsbaum gespeichert (*Tree*). Damit dieser angezeigt werden kann, wird er an den *Tree Viewer* übergeben.
Lasse dir nun den Klassifikationsbaum mit Hilfe des *Tree Viewer* anzeigen. Der Baum sieht etwas komplizierter aus, als der im letzten Arbeitsblatt. Kannst du Unterschiede zu den von dir erwarteten Attributen feststellen? Erscheinen die Kriterien, nach denen ein Schüler eine bestimmte Note bekommt, für dich logisch und sinnvoll?
Hinweis: Der Baum sieht bei dir möglicherweise anders aus als bei deinem Nachbarn. Das liegt daran, dass die 30 % der Schülerdaten auf jedem Computer getrennt zufällig ausgewählt werden. Du kannst auch bei dir eine neue Stichprobe auswählen, indem du im Data Sampler den Befehl „Sample Data“ nutzt. Es wird dann automatisch auch ein neuer Baum erzeugt.

Anhand der Aufgabe lernen die Schüler das verwendete Programm grundsätzlich kennen und können den Aufbau der Analyse sowie den entstehenden Klassifikationsbaum verstehen, was eine wichtige Grundlage für das weitere Vorgehen bildet, in dem die Schüler die Analyse nun erweitert und schlussendlich bewerten sollen.

Daher steht als nächstes die Verwendung des generierten Modells zur Vorhersage der Noten aller Schüler an:

Arbeitsblatt 5 (L): Datenanalyse am Computer (Teil 2)

Aufgabe 1
Natürlich wollen wir den Klassifikationsbaum verwenden um die Punkte der Schüler automatisch vorhersagen zu können. Dies kannst du machen, indem du von links das *Prediction*-Symbol nach rechts ziehst. Diese Funktion benötigt zwei Eingaben: Den *Baum*, anhand dessen es die Vorhersagen treffen soll, sowie die *Daten*, zu denen es etwas vorhersagen soll. Ziehe daher eine Verbindung vom Halbkreis rechts neben dem *Tree* (dieser Halbkreis entspricht dem Ausgang/Rückgabewert dieser Funktion) zum Eingang der *Prediction*-Funktion sowie vom Ausgang des *File* zum Eingang der *Prediction*.
Um nun anzuzeigen, welche Vorhersagen Orange3 getroffen hat, können wir auf die *Prediction* doppelklicken. Du siehst dann eine Tabelle, die wie folgt aussieht:

	Tree	Note 3	Schule	Geschlecht	Alter	Wohnumfeld	Familiengroesse
1	4.0	4.0	GP	W	18.0	urban	>3
2	4.0	4.0	GP	W	17.0	urban	>3
3	3.0	3.0	GP	W	15.0	urban	<=3
4	3.0	3.0	GP	W	15.0	urban	>3
5	3.0	3.0	GP	W	16.0	urban	>3
6	3.0	3.0	GP	M	16.0	urban	<=3
7	3.0	3.0	GP	M	16.0	urban	<=3

Wie sieht es aus - war deine Vorhersage perfekt? Wie gut würdest du sie in Schulnoten einschätzen (ankreuzen)?

① — ② — ③ — ④ — ⑤ — ⑥

Wenn du die echten Punktzahlen und die vorhergesagten vergleichst: Wie stark ist die maximale Abweichung, die du findest?
Bonusfrage: Es wurde als Daten an dieser Stelle wieder das File verwendet, nicht wie vorher der Data Sampler. Warum wäre es hier sinnlos, den Data Sampler als Eingabe zu nehmen?

Da die Auswertung anhand der Tabelle relativ mühsam und inakkurat ist, bietet es sich an, den Schülern noch eine Möglichkeit zu präsentieren, mit der das Ganze systematischer stattfinden kann. Es bietet sich daher an, eine sog. Confusion Matrix als Hilfmittel zu verwenden. Diese zweidimensionale Matrix hat als eine Dimension den eigentlichen Wert, als andere den vorhergesagten. Somit erlaubt sie es, einen Einblick in die Validität der Vorhersage zu gewinnen und Ausreißer und deren Ausmaß schnell und einfach zu erkennen.

Aufgabe 2
Wenn wir unsere Analyse beurteilen wollen, ist es sehr aufwändig, nur die Tabelle anzusehen. Stattdessen können wir eine *Confusion Matrix* nutzen, die uns zeigt, wie „verwirrt“ die Analyse war. Diese kannst du (nachdem du das Symbol von der Liste links in den Arbeitsbereich rechts gezogen hast) direkt mit der Prediction verbinden. Wenn du die Confusion Matrix doppelklickst zeigt sie dir eine Tabelle, an der links die echten Punktzahlen stehen und oben die vorhergesagten Punktzahlen. In der Tabelle steht für jede dieser Kombinationen, wie viele Noten dort einsortiert wurden:

- Markiere im Diagramm die perfekten Schätzungen. Wo findest du diese?
  Auf der Diagonalen von links oben nach rechts unten
- Bei wie vielen Schülern hat die Analyse richtig geschätzt?
  Abweichend je nach Schueler wegen Sampling
- Bei wie vielen Schülern war die Vorhersage nur wenig falsch, d. h. bei wie vielen hat sie sich maximal um zwei Punkte verschätzt?
  Abweichend je nach Schueler wegen Sampling

Nachdem die Schüler jetzt wahrscheinlich gesehen haben, dass die Analyse keinesfalls als perfekt angenommen werden kann, ist es sinnvoll, sich zu überlegen, wie diese verbessert werden kann. Dazu wird den Schülern eine wichtige Stellschraube eröffnet: die Größe des für die Erzeugung des Analysemodells genutzten Datensatzes, d.h. der gewählten Stichprobe. Um möglichst gute Ergebnisse zu erreichen müsste diese natürlich möglichst groß sein (idealerweise 100 %). Das ist aber oft nicht sinnvoll, da noch ein Teil der Daten benötigt wird, um das Analysemodell zu testen.

Aufgabe 3
Wir können die Analyse noch etwas verbessern. Dazu kann die Anzahl der Schüler, die für die Erstellung des Baums verwendet werden, angepasst werden. Doppelklicke dazu auf den Data Sampler und ändere die Prozentzahl der Schülerdaten ab.

- Die Analyse verbessert sich...
  - ☐ beim Erhöhen der Samplegröße
  - ☐ beim Verringern der Samplegröße
- Mit welchem Prozentsatz der Schülerdaten wird die Analyse am besten?
  Mit möglichst vielen Daten, d. h. 100 %
- Ergibt es Sinn, diesen Prozentsatz an Daten für die Erstellung des Modells zu nutzen? Was wären dabei mögliche Probleme?
  Nein, da dafür dann (im konkreten Beispiel) alle Klausuren korrigiert sein müssen, was den Sinn der Vorhersage zerstört.
- Würdest du dich dabei wohl fühlen, wenn deine Lehrerin diese Möglichkeit nutzt, um deine Arbeiten zu bewerten?
  ☐ Ja ☐ Nein
- Falls es deiner Lehrerin gelingen würde, die Qualität der Analyse zu steigern, sodass nur noch wenige Schülerinnen bzw. Schüler falsch (besser oder schlechter) bewertet werden, wäre das dann eine ausreichend faire Lösung für dich?
  ☐ Ja ☐ Nein

Als letzter Schritt bei der automatisierten Analyse bietet es sich an, zu diskutieren, was passiert, wenn wir gesamt nur eine Schulklasse mit 30 Schülern ansehen. Als Lehrerdemo kann daher schnell das Modell umgebaut werden, sodass statt der gesamten 600 Schüler nur eine Stichprobe von 30 Schülern ausgewählt wird (von denen wiederum nur ein kleiner Anteil zur Erstellung des Modells genutzt wird). Es zeigt sich damit für die Schüler noch stärker, dass große Datenmengen sinnvoll sind, wenn Vorhersagen getroffen werden sollen, während kleine Datenmengen teils enorm fehlerträchtige Analyseergebnisse nach sich ziehen. Dies zeigt, warum jegliche datenbasierte Geschäftsmodelle darauf angewiesen sind, viele Daten über ihre Kunden zu sammeln. An dieser Stelle bietet sich ggf. auch, je nach Vorwissen der Schülerinnen und Schüler, ein Vergleich mit der Wahrscheinlichkeitsrechnung bzw. dem Gesetz der großen Zahlen an.

Arbeitsblatt 6 (L): Diskussion der Ergebnisse

Als letzter Teil der Unterrichtssequenz sollte eine Diskussion der Bedeutung des gelernten stehen. Es ist an dieser Stelle wichtig, dass die Schüler sich folgende Aspekte bezüglich Datenanalysen bewusstmachen:

- Mit zunehmender Anzahl an Datensätzen wird eine Vorhersage typischerweise genauer.
- An manchen Stellen sind selbst kleinste Fehler bei der Vorhersage unerwünscht, während selbst viele Fehler an anderen Stellen tolerierbar sind.
- Dadurch, dass wir die Regeln auf denen die Vorhersagen basieren, zum Teil nicht logisch nachvollziehen können, erscheinen uns diese Vorhersagen oft als gefährlich.
- Selbst wenn wir vermuten, dass wir von Datenanalysen nur profitieren können, kann es sein, dass wir (warum auch immer) anders eingestuft werden - es sollte also kritisch hinterfragt werden, wie wir zu konkreten Nutzungsszenarien stehen.

Zu diesem Zweck bietet sich ein Gruppenpuzzle an, dass wie folgt aufgebaut wird:

Teil I: Auftrag an die Stammgruppen:
Euch stehen fünf Beispiele aus dem Bereich zu Datenanalysen zur Verfügung, die ihr im Rahmen dieser Unterrichtsphase kennenlernen sollt.
Wählt euch jede/-r genau eines der fünf Beispiele aus, mit dem ihr euch im Folgenden etwas intensiver beschäftigen wollt. Geht dann in die Expertengruppen, in denen alle zusammenkommen, die sich mit demselben Beispiel beschäftigen.
Beispiele

- Analyse von Kreditkartendaten/-nutzung:
  `http://bit.ly/2FpcRZx` – `http://bit.ly/2HcxEMP` – `http://bit.ly/2FrBzZg`
- Datenanalysen und Smart Cars:
  `http://bit.ly/2Fvso9Y` – `http://bit.ly/2oKN5oS` – `http://bit.ly/2oJa9EE`
- Datenanalysen durch KFZ-Versicherungen:
  `http://bit.ly/2Fw4ag4` – `http://bit.ly/2FrrKKD` – `http://bit.ly/2Fg5wvV`
- Beurteilung von Personen anhand von Datenanalysen: `http://bit.ly/2ssGJcA` – `http://bit.ly/2I3MxT4` – `http://bit.ly/2FqDm0u`
- Datenanalysen im Smart Home:
  `http://bit.ly/2FYpqbM` – `http://bit.ly/2uBQHd0` – `http://bit.ly/2m0eaCs`

Teil II: Auftrag an die Expertengruppen

Ihr erhaltet zu einem Kontext von Datenanalysen und -vorhersagen ein Beispiel wie dieses real oder fiktiv eingesetzt werden könnte.
Versucht dieses Beispiel gemeinsam nachzuvollziehen. Besprecht dazu miteinander folgende Fragen und macht euch dazu Notizen. Bereitet euch darauf vor, eure Ergebnisse in den Stammgruppen kurz vorzustellen und zu diskutieren.

- Welche Daten werden genutzt?
- Wie werden die Daten analysiert?
- Was ist das Ziel dieser Analyse?
- Ist das Beispiel überhaupt praktisch machbar? Warum bzw. warum nicht?
- Sind mögliche Fehler bei der Datenanalyse tolerierbar oder nicht?
- Sehr ihr eine solche Analyse als hilfreich/sinnvoll/nützlich an oder eher als gefährlich? Sollte es zulässig sein, diese Art der Analyse zu nutzen?
- Würdet ihr eure Daten für diesen Zweck freiwillig hergeben?
- Wofür könnten die Daten zukünftig - wenn sie schon einmal da sind - noch genutzt werden?

Teil III: Auftrag an die Stammgruppen:
Nachdem ihr nun in den Expertengruppen die Beispiele diskutiert habt, sollt ihr diese nun gemeinsam vergleichen. Dazu habt ihr in eurer Gruppe nun einen Experten für jedes der fünf Beispiele.
Damit alle über die Beispiele Bescheid wissen, erklärt ihr diese einander kurz. Geht dabei insbesondere auf die in den Expertengruppen diskutierten Aspekte ein.
Diskutiert nun die Gemeinsamkeiten und Unterschiede der Beispiele: Handelt es sich bei den Analysen um von euch gewünschte und sinnvolle Arten der Datennutzung? Wenn ja, welche Vorteile hat das für euch? Wenn nein, wie könnt ihr diesen möglicherweise entfliehen und verhindern, dass eure Daten so genutzt werden?

Anhang G: Beobachtungsbogen zur Unterrichtserprobung

Notizen									
Kontrolle	Lehrer								
	Schüler								
Methode	Diskussion								
	Gruppen-arbeit								
	Partner-arbeit								
	Demo								
	Arbeits-auftrag								
	Lehrer-vortrag								
Phase	Diskussion								
	Festigung								
	Übung								
	Heran-führung								
	Motivation								
Zeit									

Anhang H: Schülerfragebogen zur Unterrichtserprobung

Fragebogen zur Unterrichtseinheit Datenanalyse und Vorhersage

Wie sehr interessieren dich folgende Themen der Informatik bzw. Informationstechnologie? (ankreuzen)

	sehr interessant	etwas interessant	kaum interessant	gar nicht interessant	ist mir unbekannt
Programmierung	👍	👌	👇	👎	?
Entwickeln von Computerspielen	👍	👌	👇	👎	?
Technisches Zeichnen / CAD	👍	👌	👇	👎	?
Datenspeicherung	👍	👌	👇	👎	?
Datenanalyse	👍	👌	👇	👎	?
Künstliche Intelligenz	👍	👌	👇	👎	?
Computergrafik	👍	👌	👇	👎	?
Entwickeln von Smartphone-Apps	👍	👌	👇	👎	?
Office-Programme	👍	👌	👇	👎	?

Wie sehr stimmst du den folgenden Aussagen zum Datenanalyse-Unterricht zu? (ankreuzen)

	stimme zu	stimme eher zu	stimme teils zu	stimme kaum zu	stimme nicht zu
Die Aufgaben waren für mich gut lösbar.	👍	👌	👇	👇	👎
Das Programm Orange 3 war gut bedienbar.	👍	👌	👇	👇	👎
Ich verstehe jetzt, was mit Daten gemacht werden kann.	👍	👌	👇	👇	👎
Ich kann jetzt selbst Daten analysieren.	👍	👌	👇	👇	👎
Das Thema hat mir mehr Spaß gemacht als viele andere Themen im IT-Unterricht.	👍	👌	👇	👇	👎
Ich möchte im Unterricht noch mehr über Datenanalysen erfahren.	👍	👌	👇	👇	👎
Ich möchte nicht, dass jeder Daten bzw. Informationen über mich bekommt.	👍	👌	👇	👇	👎
Das Thema Datenanalyse und Vorhersage war kompliziert.	👍	👌	👇	👇	👎
Wenn möglich vermeide ich es, Daten über mich beispielsweise im Internet preiszugeben.	👍	👌	👇	👇	👎
Mit meinen Daten kann sowieso niemand etwas anfangen.	👍	👌	👇	👇	👎
Ich war überrascht, wie einfach gute Vorhersagen erstellt werden können.	👍	👌	👇	👇	👎

Was hast du im Unterricht zu Datenanalysen/Vorhersagen gelernt? (Stichworte reichen aus)
Ich habe gelernt / bemerkt / entdeckt / weiß jetzt, …

-
-
-
-
-

Welche Schritte müssen durchgeführt werden, um anhand von Daten eine Vorhersage zu treffen?

Was bedeutet der Begriff „Klassifikation“?

Wie kann die Qualität von Vorhersagen oft verbessert werden?

Datenanalysen im Gesundheitssystem
Immer häufiger wird vorgeschlagen, dass Ärzte ihre Patienten zukünftig insbesondere anhand von Daten untersuchen sollen. Dies wird jedoch sehr umstritten.

Welche Daten könnten Ärzte dafür möglicherweise nutzen? Woher kommen diese?

Welche Probleme könnten auftreten, wenn Ärzte zukünftig Patienten anhand von Daten untersuchen?

Im Internet wird behauptet, dass Datenanalysen besser sind als normale Untersuchungen durch Ärzte, weil sie sich nicht irren können. Stimmt das? Welche Probleme können bei solchen „digitalen Untersuchungen“ auftreten?