DATEN IM INFORMATIKUNTERRICHT

Schlüsselkonzepte des Datenmanagements als Grundlage für die Förderung von Datenkompetenzen im Unterricht

Andreas Grillenberger und Ralf Romeike
Didaktik der Informatik, Freie Universität Berlin
andreas.grillenberger@fu-berlin.de

CERG
Computing Education Research Group Berlin

Freie Universität Berlin

Diese Broschüre baut auf den folgenden Veröffentlichungen der Autoren auf und ist in Teilen mit diesen identisch:

- Grillenberger, A.: Von Datenmanagement zu Data Literacy: Informatikdidaktische Aufarbeitung des Gegenstandsbereichs *Daten* für den allgemeinbildenden Schulunterricht. Dissertation, Freie Universität Berlin, 2019.
- Grillenberger, A. & Romeike, R.: Datenmanagement als Thema für den Informatikunterricht, In: LOG IN, volume 37, 2018.
- Grillenberger, A. & Romeike, R.: Von Klassen und Bäumen – Einstieg in das Thema Data Mining im Informatikunterricht. (in Vorbereitung)
- Grillenberger, A. & Romeike, R.: Empirische Ermittlung der Schlüsselkonzepte des Fachgebiets Datenmanagement, In: Informatische Bildung zum Verstehen und Gestalten der digitalen Welt, 2017.
- Grillenberger, A. & Romeike, R.: Schlüsselkompetenzen im Umgang mit Daten: Ein Thema für den Informatikunterricht, In: INFOS 2015: Informatik allgemeinbildend begreifen, 2015.
- Grillenberger, A. & Romeike, R.: Big-Data-Analyse im Informatikunterricht mit Datenstromsystemen: Ein Unterrichtsbeispiel, In: INFOS 2015: Informatik allgemeinbildend begreifen, 2015.
- Grillenberger, A. & Romeike, R.: Big Data im Informatikunterricht: Motivation und Umsetzung, In: INFOS 2015: Informatik allgemeinbildend begreifen, 2015.
- Grillenberger, A. & Romeike, R.: Teaching Data Management: Key Competencies and Opportunities, In: KEYCIT 2014 – Key Competencies in Informatics and ICT, 2014.

Online-Variante / Weitergabe

Diese Broschüre steht als Onlineversion unter https://dataliteracy.education zur Verfügung. Eine Weitergabe ist explizit erlaubt, solange die Bedingungen der Creative-Commons-Lizenz CC BY-NC-SA 4.0 eingehalten werden. Dies bedeutet insbesondere:

- Sie dürfen die Inhalte verändern, müssen diese aber unter derselben Lizenz zur Verfügung stellen.
- Die Inhalte dürfen nicht kommerziell verwendet werden.
- Sie müssen den Autor als Urheber angeben.

Wir würden uns freuen, wenn Sie uns über Verwendungen des Materials informieren! Abweichungen von der Lizenz, z. B. für die Verwendung der Materialien im Unterricht, können durch die Autoren genehmigt werden (Kontakt per E-Mail, siehe unten).

Verantwortlich für die Herausgabe der Broschüre

Andreas Grillenberger

Didaktik der Informatik, Freie Universität Berlin
Königin-Luise-Straße 24-26, 14195 Berlin

E-Mail: andreas.grillenberger@fu-berlin.de
Telefon: +49 30 838 66103
Web: https://dataliteracy.education
Lehrkräfte-Mailingliste: https://lists.fu-berlin.de/listinfo/lehrkraefte-data-literacy

INHALTSVERZEICHNIS

Daten in der digitalen Welt **4**

Das Fachgebiet Datenmanagement **5**

Historische Entwicklung 5

Zentrale Themen des Datenmanagements 8

Bedeutung von Datenmanagement im Alltag **14**

Schlüsselkonzepte des Datenmanagements **19**

Kerntechnologien 20

Praktiken 20

Entwurfsprinzipien 21

Mechanismen 21

Nutzen für den Unterricht 22

Grundlegende Datenkompetenzen:
Kompetenzmodell für den Informatikunterricht **23**

Inhaltsbereiche des Datenkompetenzmodells 24

Prozessbereiche des Datenkompetenzmodells 24

Exemplarische Kompetenzen 25

Praktische Ideen für den Informatikunterricht **27**

Snap!Twitter 27

Snap!DSS 31

Data Mining: Datenanalyse und Vorhersage 33

Zusammenfassung **38**

Literatur **39**

DATEN IN DER DIGITALEN WELT

Daten werden heute nicht mehr nur als Thema der Informatik wahrgenommen, sondern stellen ein wichtiges Fundament vielfältiger Entwicklungen dar, die oft unter dem Begriff Digitalisierung subsummiert werden. Die Chance, von den damit einhergehenden Möglichkeiten und Chancen zu profitieren, wird oft nicht ergriffen, insbesondere aufgrund des fehlenden Verständnisses von Datenanalysen und deren Auswirkungen: Trotz einer umfangreichen Thematisierung im gesellschaftlichen Diskurs fällt es großen Teilen der Bevölkerung schwer, die Mächtigkeit, Möglichkeiten und Gefahren von Datenanalysen einzuschätzen und sich entsprechend zu positionieren, da wesentliche Grundlagen für das Verständnis der Funktionsweise fehlen. Die Entscheidung, ob aktuelle Entwicklungen, wie die Aufzeichnung des Fahrverhaltens durch KFZ-Versicherungen, geduldet und genutzt werden, kann und muss jeder selbst unter Berücksichtigung des persönlichen Kosten-Nutzen-Verhältnisses und unter Berücksichtigung der Auswirkungen auf die eigene Umwelt und die Gesellschaft treffen. Die Bedeutung einer Kompetenz zum kritisch-reflektierten Umgang mit solchen Themen zeigt sich jedoch insbesondere, wenn entsprechende Verfahren umfassend eingesetzt werden, ohne dass eine Entscheidung für oder gegen eine Teilnahme an diesem System möglich ist: Beispielsweise findet eine Bewertung von Personen anhand von Daten nicht nur durch bekannte Auskunfteien, sondern immer häufiger auch durch staatliche Einrichtungen statt. In China ist diese Entwicklung bereits soweit fortgeschritten, dass dort in Kürze alle Einwohner im Rahmen eines Social-Credit-Systems positiv oder negativ eingestuft werden, mit dem Ziel, diese in Richtung eines gewünschten Verhaltens zu erziehen (Kolany-Raiser & Radtke, 2018). Die mit solchen Möglichkeiten einhergehenden Konsequenzen können ohne ein fundiertes Grundwissen im Umgang mit Daten kaum bewertet werden, da ansonsten das Ausmaß und die Möglichkeiten dieser Analysen im Verborgenen bleiben.

Um auf ein selbstbestimmtes und mündiges Leben in der digitalen Gesellschaft vorzubereiten, muss die Schule – und in diesem Fall speziell der Informatikunterricht – einen Einblick in solche Entwicklungen geben und die Schülerinnen und Schüler befähigen, diese kritisch-reflektiert zu betrachten. Obwohl im Informatikunterricht bereits verschiedene Ansätze zur Förderung eines fundierten Umgangs mit Daten erkennbar sind, findet sich hier aber auch heute noch eine große Lücke.

In dieser Broschüre werden Grundlagen des Themenfelds Daten und Datenmanagement zusammengefasst, mit dem Ziel, Anregungen für die Unterrichtspraxis zu geben. Dazu wird die historische Entwicklung dargestellt, zentrale Themenbereiche erläutert und die Relevanz für Schülerinnen und Schüler anhand von Beispielen erläutert. Die fachliche und fachdidaktische Fundierung der am Ende der Broschüre vorgestellten Unterrichtsbeispiele wird durch das Modell der Schlüsselkonzepte des Datenmanagements und ein an die Bildungsstandards Informatik angelehntes Datenkompetenzmodell vorgenommen.

Zu diesem Thema sind weitere Arbeiten angedacht, die auf der Projektwebseite[1] veröffentlicht werden. Weiterhin sind Lehrkräftefortbildungsmaßnahmen geplant, die dort angekündigt werden. Zusätzlich existiert eine für alle interessierten offene Mailingliste[2], auf der in unregelmäßigen Abständen für Lehrkräfte relevante Informationen bekanntgegeben werden.

---

[1] Projektwebseite: https://dataliteracy.education

[2] Mailingliste: https://lists.fu-berlin.de/listinfo/lehrkraefte-data-literacy

DAS FACHGEBIET DATENMANAGEMENT

Aus fachlicher Perspektive werden alle Aspekte der Verwaltung und Verarbeitung von Daten heute insbesondere durch das Fachgebiet Datenmanagement fundiert. 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. 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, wie beispielsweise 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 Datensicherheit[3], ethisch korrektem Umgang mit Daten[4] sowie Datenschutz[5]. Auch in anderen Bereichen der Informatik, wie der Softwareentwicklung, ist Datenmanagement heute integraler Bestandteil.

HISTORISCHE ENTWICKLUNG

In den letzten zehn bis fünfzehn Jahren ist 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. Abbildungen 1 und 2). Die mit diesem Wachstum einhergehenden steigenden Anforderungen an Datenmanagement stellen eine der zentralen Herausforderungen des Fachgebiets dar und haben wesentlich zu dessen Weiterentwicklung beigetragen.

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 Neues ist. Die erste dokumentierte Arbeit, die bereits dem heutigen *Data Mining* (vgl. Seite 12) 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 zum Einen um

---

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

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

[5] Der Begriff *Datenschutz* ist aus informatischer Sicht klar vom ethisch korrekten Umgang mit Daten zu unterscheiden: Beim Datenschutz geht es insbesondere um den (technischen bzw. organisatorischen) Schutz personenbezogener bzw. -beziehbarer Daten vor Diebstahl, Manipulation und Missbrauch. Datenschutz ist daher ein Aspekt der Datensicherheit, der die sich auf den "*Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten*" (*Europäische Union, 2016*) konzentriert.

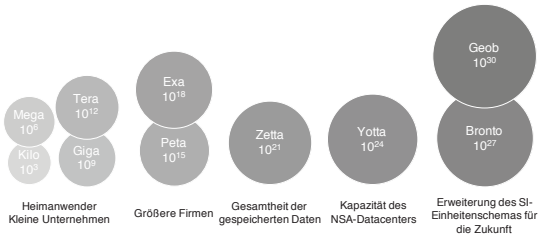


Abbildung 1: Überblick über die Größenordnungen der heute in verschiedenen Bereichen gespeicherten Datenmengen.

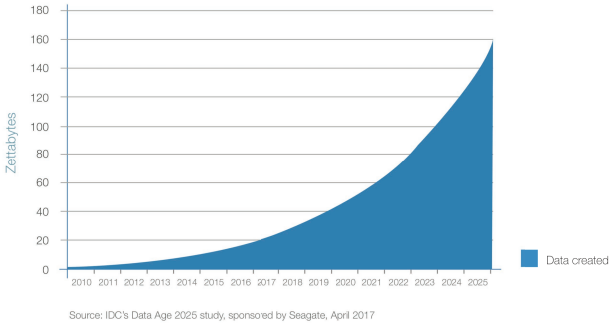


Abbildung 2: Wachstum der Datenmenge der Menschheit (*IDC, 2017*).

Daten, die explizit und strukturiert für diesen Zweck erfasst wurden (z. B. 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 für andere Zwecke erfasst worden sind. Durch Kombination dieser Daten konnte Maury Seerouten durch Ausnutzung von Strömungen und Winden optimieren (vgl. *Maury, 1855*) und so die Reisezeit von Schiffen auf den betrachteten Routen durchschnittlich um ein Drittel reduzieren (vgl. *Mayer-Schönberger und Cukier, 2013*). Obwohl die Datenmenge für heutige Verhältnisse zwar klein war, ist sie für eine händische Datenverarbeitung durchaus beträchtlich. Zugleich zeigen sich hieran zentrale Eigenschaften moderner Datenanalysen: Es werden nicht nur verschiedenste Datenquellen miteinander verknüpft, sondern dabei auch solche Daten miteinbezogen, die sowieso schon für andere Zwecke erfasst wurden und somit schon vorhanden waren. Nicht nur die Datenanalyse von Maury, sondern auch ihre Einflüsse können sich bei zeitgeschichtlicher Betrachtung durchaus mit denen heutiger Datenanalysen messen.

Ein weiteres Beispiel, das die Bedeutung solcher 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 deutliche Korrelation zwischen den Merkmalen *an Cholera erkrankt* und *in Hamburg lebend* ablesen: Abbildung 3 zeigt, dass die politische Grenze zwischen Hamburg und Altona mit wenigen Ausnahmen auch 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 eine zugrundeliegende Kausalität aufgedeckt werden: ein an den politischen Grenzen getrenntes, unterschiedlich weit entwickeltes Wasserversor-gungssystem. Auch dieses Beispiel demonstriert das hohe Potential korrelationsbasierter Datenanalyse.

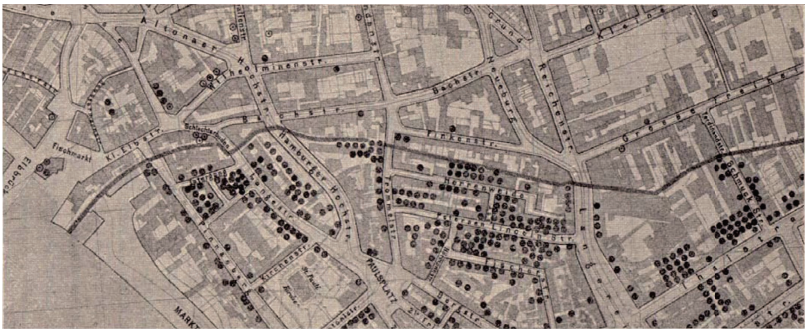


Abbildung 3: Cholerafälle an der Grenze von Hamburg (südlich) und Altona. Aus: Exner (2009).

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 des letzten Jahrzehnts wurde dieses Themengebiet deutlich aufgeweitet, sodass das Fachgebiet Datenmanagement entstand. Der Kristallisationskeim dieser Entwicklung ist unter dem Stichwort *Big Data* bekannt, dessen zentrale Charkteristika die großen Mengen verschiedenster Arten von Daten, die in hoher Geschwindigkeit gespeichert und verarbeitet werden, charakterisiert wird (vgl. *Kemper und Eickler, 2015*). Auch der Wandel von kausalitätsbasierten zu immer häufiger korrelationsbasierten Datenanalysen (oft unter dem Stichwort *Data Mining*) ist 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 (vgl. *Sommer, 2013*): Hierbei werden die Einkäufe der Kunden auf korrelative Zusammenhänge hin analysiert und ggf. weitere Daten, beispielsweise 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 gesellschaftlicher Diskurse, beispielsweise im 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.

Korrelationsbasierte Datenanalysen großer Datenmengen weisen seit jeher hohes Potential auf. Durch neue Möglichkeiten zur Analyse großer und Datenmengen in hoher Geschwindigkeit nimmt dieses jedoch weiter zu.

ZENTRALE THEMEN DES DATENMANAGEMENTS

Im Folgenden werden, um einen ersten Überblick über das Fachgebiet zu geben, zentrale Themen des Datenmanagements herausgegriffen und ausführlich beschrieben. Dabei handelt es sich um wesentliche Themen- und Forschungsbereiche, die so ausgewählt wurden, dass sie die unterschiedlichen Bereiche des Fachgebiets repräsentieren.

Verwaltung und Nutzung großer Datenmengen: Big Data

Der bekannteste Begriff im Zusammenhang mit den Weiterentwicklungen im Datenmanagement ist *Big Data*. Es existiert jedoch keine klare und anerkannte Definition. Unter Bezug auf *McBurney (2013)* schreibt die GI: "*Der Ursprung und die erstmalige Verwen-dung 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 aber Einigkeit darüber, dass Big Data durch die "*drei V*" (vgl. Abbildung 4; *Kemper und Eickler (2013), 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 den *drei V* werden häufig noch weitere genannt, wie die Vertrauenswürdigkeit (*veracity*) und der Wert (*value*) der Daten (vgl. z. B. *Ali-ud-din Khan, Fahim Uddin und Gupta 2014*).

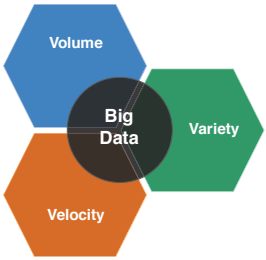


Abbildung 4: Die "drei V" von Big Data

Im Zusammenhang mit Big Data entstanden Technologien, wie parallelisierbare Analysealgorithmen (bspw. Googles Map-Reduce-Algorithmus, vgl. *Dean und Ghemawat (2008)*) oder Datenbanksysteme, die klassische Konzepte auf den Prüfstand stellen und versuchen, den neuen Anforderungen gerecht zu werden. Diese 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 Datenbankkonzepten nicht darum, dass diese kein SQL unterstützen, sondern um eine Abkehr vom relationalen Modell. 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. Beispielsweise lassen dokumentenorientierte Datenbanken dem Nutzer große Freiräume, da sie kein definiertes Datenschema benötigen: So erlauben dokumentenorientierte NoSQL-Datenbanken die Speicherung unterschiedlich strukturierter Dokumente in derselben *Kollektion*[6]. Die Einhaltung vereinbarter Strukturen oder Datenmodelle obliegt daher dem Anwender bzw. der Anwendung. Zur Formulierung von Anfragen an die meisten nichtrelationalen Datenbanken wird nicht die Anfragesprache SQL verwendet, sondern eigene systemspezifische Sprachen, die sich häufig an bekannter Syntax, wie der objektorientierten Syntax von Java,

[6] In dokumentenorientierten Datenbanken entspricht die *Kollektion* im wesentlichen einer Tabelle bei relationalen Datenbanken. Je nach konkretem System werden jedoch auch andere Begriffe genutzt.

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.

Big Data befasst sich mit der Verarbeitung großer Mengen unterschiedlichster Daten in hoher Geschwindigkeit (*volume, variety, velocity*). Diese Analysen werden durch neue Technologien wie NoSQL-Datenbanken ermöglicht, die zum Teil auf bewährte Eigenschaften wie die Sicherstellung von Konsistenz verzichten.

Handhabung großer Datenmengen: Verteilte und Cloud-Datenspeicher

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 anderem 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, insbesondere aufgrund der zunehmenden Datenmengen die heute gespeichert werden, an Bedeutung: 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 einnehmen 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 bei verteilter Datenspeicherung wahren zu können, muss sichergestellt werden, dass beispielsweise Änderungen an den gespeicherten Daten auf allen beteiligten Servern vollzogen wurden, bevor eine neue Transaktion zugelassen wird – es ist offen-sichtlich, dass diese Prüfung die Geschwindigkeit dieser Aktion reduziert und somit bei verteilter Datenspeicherung die Geschwindigkeit einschränkt. Dies wird durch das *CAP-Theorem (Edlich et al., 2011* und *Brewer, 2012*) beschrieben, laut dem die drei Eigenschaften Konsistenz (**c**onsistency), Verfügbarkeit (**a**vailability) und Partitionstoleranz (**p**artition tolerance}) unvereinbar sind. Nur zwei dieser Eigenschaften können gleichzeitig erreicht werden. Das CAP-Theorem verdeutlicht damit eine der 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[7]-Paradigma gehorchen, aber nur eingeschränkt für verteilte Datenspeicherung geeignet sind, vernachlässigen nicht-relationale Datenbanken

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

typischerweise die Konsistenz des Datenbestandes. Sie genügen dem BASE[8]-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 herzustellen. Solche Datenbankvarianten werden typischerweise in Situationen eingesetzt, in denen die dauerhafte Konsistenz der Daten weniger kritisch ist, beispielsweise bei sozialen Netzwerken oder Suchmaschinen. Bekannte Vertreter dieser Datenbanken stellen die verteilte und auf Leistung optimierte Datenbank *Google BigTable* (*Chang et al., 2008*), die unter anderem für Google Books genutzt wird, die graphbasierte Datenbank *Neo4J* oder die heute weit verbreitet eingesetzte dokumentenorientierte Datenbank *MongoDB* dar.

Mit der zunehmenden Verbreitung verteilter Datenspeicherung steigt auch die Bedeutung von verteilten Analysen: Um eine hohe Analysegeschwindigkeit zu erzielen, ist es heute in vielen Fällen essenziell, Analyseprozesse auf verschiedenen Rechenknoten parallel durchzuführen. Dabei ist beispielsweise der *Map-Reduce*-Algorithmus von (vgl. Abbildung 5) zentraler Bedeutung: Dieser teilt die Verarbeitung in vier Schritte[9] auf, von denen die zwei den Namen des Algorithmus bestimmenden durch den Nutzer spezifiziert und jeweils parallel auf mehreren Rechenknoten ausgeführt werden. Dies ermöglicht eine hochgradig parallele Datenverarbeitung und somit eine hohe Flexibilität. Das Map-Reduce-Verfahren wird heute in einer Vielzahl von Anwendungsfällen eingesetzt und stellt ein wichtiges Beispiel für die Datenverarbeitung im Big-Data-Zeitalter dar.

Zur Speicherung großer Datenmengen muss auf verteilte Datenspeicher zurückgegriffen werden. Zur Erhöhung ihrer Performanz, muss gegebenenfalls auf andere als zentral erachtete Eigenschaften von Datenspeichern verzichtet werden, z. B. Konsistenz (*CAP-Theorem*). Um die Verarbeitungsgeschwindigkeit zu erhöhen, wird hingegen häufig Parallelisierung eingesetzt (z. B. *Map-Reduce*).

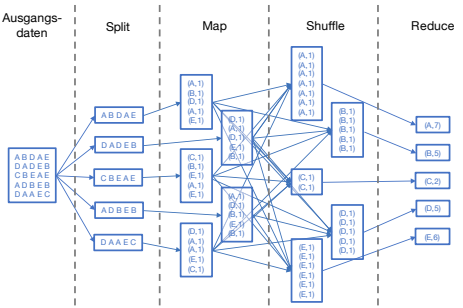


Abbildung 5: Veranschaulichung des Map-Reduce-Algorithmus am Beispiel einer Wortzählung.

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

[9] Je nach Darstellung wird teils auf den ersten Schritt, *split*, verzichtet, sodass der Algorithmus auch häufig mit drei Schritten dargestellt ist.

Strukturierung und Beschreibung von Daten durch Metadaten

Ein weiteres zentrales Thema des Datenmanagements, das zugleich auf einer eher technischen aber auch auf einer eher nutzerorientierten Ebene angesiedelt sein kann, stellen Metadaten dar: Diese werden zur internen Datenorganisation und -strukturierung innerhalb von Datenmanagementsystemen eingesetzt, beispielsweise in Form von Prüfsummen oder (Such-)Indizes, treten aber andererseits auch auf Anwendungsebene und damit im Wahrnehmungsbereich des Nutzers zutage und spielen selbst im gesellschaftlichen Diskurs eine Rolle: Die Möglichkeiten und Gefahren der Erfassung von Metadaten und deren Einfluss auf unsere Privatsphäre wird immer häufiger diskutiert.

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

- *Deskriptive Metadaten* beschreiben oder identifizieren Informationen und Informationsquellen und machen diese (einfacher) zugänglich/auffindbar. Dabei handelt es handelt sich z. B. um den Namen einer Datei, den Entstehungsort eines Fotos, die Autoreninformationen bei Dokumenten, Informationen über den Client bei einer HTTP-Anfrage 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 vorgesehenes Ablauf- bzw. 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 Funktionen schließen sich nicht gegenseitig aus: Beispielsweise hat ein einem Foto mitgegebenes Schlagwort eine deskriptive sowie 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 erst möglich: Ohne Metadaten wäre die Synchronisation von Daten genauso wenig denkbar, wie der Zugriff auf Fotos durch Suche nach deren Entstehungsort oder die Organisation von Dateien in Ordnerstrukturen. Gleichzeitig schafft die oft im Hintergrund stattfindende Erzeugung von Metadaten auch verschiedene Gefahrenpotentiale: So wurden Fälle bekannt, in denen durch Nutzung von Metadaten Kommunikations- oder Bewegungsprofile erstellt oder durch ein fehlendes Bewusstsein über Metadaten Ausschnitte geheimer Dokumente veröffentlicht wurden. Für einen verantwortungsbewussten und selbstbestimmten Umgang mit Daten ist daher heute ein Bewusstsein für die Allgegenwärtigkeit von Metadaten unabdingbar.

Ein weiteres Beispiel für die Aussagekraft von Metadaten stellt die *Strava Heat Map* (*Strava, 2017*) dar: Auf dieser Karte stellt der Fitnesstracker-Hersteller Strava die Aktivitäten seiner Nutzerinnen und Nutzer dar. Die Erfassung der Positionsdaten war dabei nicht der primäre Zweck der eingesetzten Fitnesstracker, diese wurden den Aktivitätsdaten als vom Nutzer kaum wahrgenommenes Metadatum beigefügt. Obwohl die über diese Heat Map gewonnen aten auf den ersten Blick harmlos erscheinenen, erlauben sie auch sensible in nicht-öffentliche Bereiche wie Militärbasen *(The Verge, 2018)*, die zum Teil 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.

Metadaten werden zur Beschreibung von Daten, zu deren Strukturierung und für administrative Zwecke eingesetzt. Sie liefern zusätzliche, über die eigentlichen Daten hinausgehende Informationen, die auch eine wertvolle Quelle für Analysen darstellen.

Korrelationsbasierte Datenanalyse: Data Mining

In engem Zusammenhang mit den 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 zu entdecken oder Trends zu erkennen, werden die Daten mit statistischen Verfahren fast schon willkürlich miteinander in Beziehung gesetzt. Dabei geht es zumeist nicht um die Ermittlung von Kausalzusammenhängen, sondern rein um Korrelationen, auch wenn diese ggf. 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 solcher Data-Mining-Analysen ist die Analyse von Suchanfragen durch Suchmaschinenbetreiber. Beispielsweise entwickelte Google unter dem Namen *Flu Trends* ein Modell zur Vorhersage von Grippewellen. Die Herangehensweise an das Problem ist typisch für Data Mining: Statt sich auf rein logische Begriffe wie “Grippe”, auf Symptome oder auf Medikamentennamen zu fokussieren, wurden willkürlich wirkende Suchwörter mit den Statistiken der Gesundheitsbehörden abgeglichen und dabei nach Korrelationen gesucht (*Ginsberg et al., 2009*). Das Ergebnis war ein Katalog von Suchbegriffen und ein Modell, das, trotz zeitweiser Fehler in der absoluten Zahl an Grippefällen, die Verläufe von Grippewellen gut vorhersagte (*Valdivia et al., 2010*). Während die Begriffe mit der höchsten Korrelation erklärbar waren, da es sich oft um die logisch erwarteten handelte, und diese korrelationsbasierte Analyse somit zur Bestätigung von Kausalitäten beitrug, zeigten auch weitere Begriffe erstaunlich hohe Korrelation mit Grippewellen, obwohl kein logischer Zusammenhang erkennbar ist. Auch wenn Flu Trends mittlerweile eingestellt wurde, bietet Google mit *Trends* auch weiterhin ein Werkzeug zur Suchdatenanalyse an, mit dem jeder die Häufigkeit von Suchbegriffen auswerten und eigene Analysen durchführen kann. Damit lassen sich beispielsweise Vermutungen über die Verbreitung von bestimmten Geräten/Diensten oder die Beliebtheit verschiedener Personen in verschiedenen Regionen aufstellen, die zeitliche Entwicklung der Popularität von Suchbegriffen nachvollziehen und damit in Zusammenhang stehende Themen finden.

Data Mining befasst sich mit der explorativen Auswertung, oft von bereits existierenden Daten. Meist steht dabei die Suche nach korrelativen Zusammenhängen im Vordergrund, die aber häufig auch zur Entdeckung von Kausalzusammenhängen beitragen.

Datenauswertung in Echtzeit: Datenstromsysteme

Eine neue Möglichkeit zur Datennutzung stellen heute Datenstromsysteme dar: Im Gegensatz zu Datenbanken, bei denen Daten dauerhaft vorgehalten und mehrfach für Analysen herangezogen werden können, verarbeiten Datenstromsysteme Daten in Echtzeit. Dazu werden vorher definierte Analysen sofort durchgeführt und die Daten nicht dauerhaft, sondern höchstens kurzzeitig, gespeichert. Dies steht klar im Gegensatz zum

Grundprinzip einer Datenbank, bei der die Dauerhaftigkeit essenziell ist. Diese Vorgehensweise führt insbesondere dazu, dass Analysen schneller und effizienter als bei Datenbanken durchführbar sind. Gleichzeitig büßen sie jedoch an Flexibilität ein, da einmal analysierte Daten nicht für weitere Analysen zur Verfügung stehen, sodass insbesondere Data-Mining mit diesen Systemen nicht möglich ist. Datenstromsysteme verarbeiten einen Eingabedatenstrom[10] und erzeugen einen aus den Analyseergebnissen bestehenden und dem Eingabedatenstrom gleichgetakteten Ergebnisdatenstrom. Obwohl Echtzeitanalysen im Allgemein ein eher komplexes Thema sind, können Datenstromsysteme auf ein einfaches Prinzip reduziert werden: 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 Unterricht die Möglichkeit zu geben, selbst Datenanalysen in Echtzeit durchzuführen und dahinterstehende Prinzip zu verstehen.

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 Analyse mit einer herkömmlichen Datenbank scheitert daher schon an der in kurzer Zeit anfallenden großen Datenmenge. Gleichzeitig ist es aufgrund der hohen Dynamik und der nicht in wenigen Millisekunden durchführbaren Analysen auf solchen Datenmengen kaum möglich, sinnvolle Analysezeitpunkte zu bestimmen, die nötig wären, da eine klassische Datenanalyse mit Datenbanken nicht kontinuierlich stattfinden kann. Mit einem Datenstromsystem sind solche Analysen jedoch relativ einfach und effizient möglich. Entsprechende Unterrichtsideen werden am Ende dieser Broschüre beschrieben.

Datenstromsysteme sind im Gegensatz zu Datenbanken nicht auf eine dauerhafte Speicherung sondern eine schnelle Analyse von Daten, idealerweise in Echtzeit, optimiert. Sie agieren als Filter für Datenströme und generieren so kontinuierlich Ergebnisse auf anhand bereits vorher definierter Abfragen.

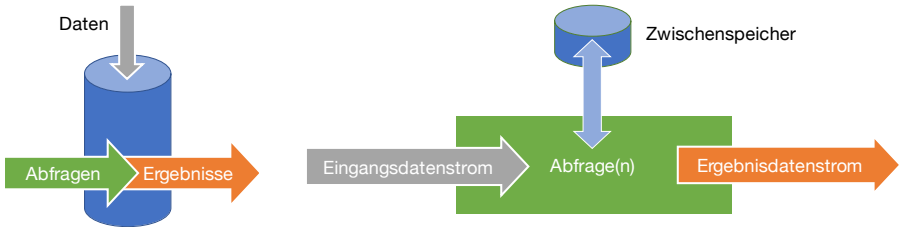


Abbildung 6: Vergleich des Funktionsprinzips von Datenbanksystemen (links) und Datenstromsystemen (rechts).

[10] *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.*"

BEDEUTUNG VON DATENMANAGEMENT IM ALLTAG

Aufgrund des deutlichen und weiterhin zunehmenden Einflusses von Daten auf das tägliche Leben wird der verantwortungsbewusste Umgang mit eigenen und fremden Daten ein immer integralerer Bestandteil unseres 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. Um die Notwendigkeit dafür erkennen und solche Möglichkeiten fundiert einsetzen zu können, ist es notwendig, den Wert der eigenen Daten und die Möglichkeiten moderner Datenanalyse zu erkennen.

Im Kontext der aktuellen Innovationen auf diesem Gebiet wird Datenmanagement immer stärker zu einem Thema des Alltags. Somit kann eine Vielzahl verschiedener Beispiele gefunden werden, die Themen des Datenmanagements mit dem Alltag verknüpfen, von denen im Folgenden einige charakterisiert werden.

Datenspeicherung

Die Speicherung von Daten ist heute im Alltag allgegenwärtig. Jeder speichert große Datenmengen, beispielsweise in Form von Textdokumenten, Musik, Videos oder E-Mails. Dabei gibt es gibt verschiedene Möglichkeiten, diese Daten zu strukturieren, zu speichern und zu organisieren. Gleichzeitig können sie u. a. lokal, im privaten Netzwerk oder in der Cloud, als Datei oder in einer Datenbank, als einfache Datei mit beliebigem Inhalt oder unter Nutzung von spezialisierten Werkzeugen, wie z. B. Medienbibliotheken, gespeichert werden (vgl. Abbildung 7). In den meisten Fällen passiert dies heute weiterhin lokal als Datei, wobei die Auslagerung von (Kopien dieser) Dateien in entfernte Cloud-Datenspeicher immer häufiger geschieht. Diese Möglichkeit ist flexibel genug, damit sie für vielfältige Arten von Daten eingesetzt werden kann. Im Gegensatz dazu unterstützen spezialisierte Werkzeuge meist nur eine bestimmte Art von Daten, bieten aber bei deren Verwaltung höheren Komfort: E-Mails werden genauso wie Kalender kaum direkt als Dateien, sondern in spezialisierten Programmen verwaltet, Musik kann in Medienbibliotheken um umfangreiche Informationen angereichert, und damit einfacher zugreif- und organisierbar verwaltet werden. Bei Wahl eines ungeeigneten Datenspeichers treten Unzulänglichkeiten auf, mit denen der Anwender konfrontiert wird und für deren Lösung ein grundlegendes Verständnis der dahinterstehenden informatischen Grundlagen hilfreich ist.

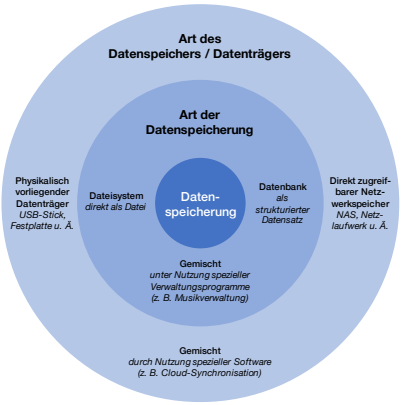


Abbildung 7: Verschiedene Möglichkeiten zur Datenspeicherung.

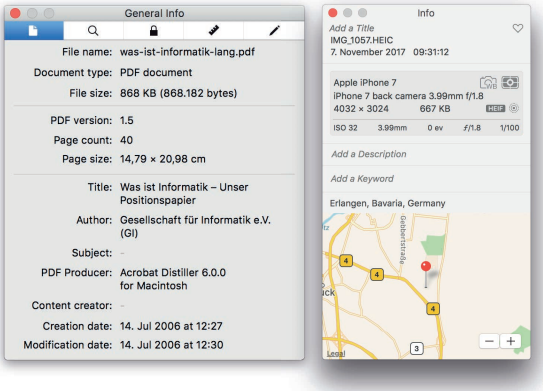


Abbildung 8: Übersicht über die Metadaten eines PDF-Dokuments und eines Fotos.

Metadaten

Ergänzend zu den eigentlichen Daten werden heute immer auch, oft umfangreiche, Zusatzinformationen als sogenannte Metadaten bzw. Rahmendaten erzeugt und gespeichert. Es handelt sich beispielsweise um einfache Dateiattribute, wie den Zeitstempel der letzten Änderung oder den Dateinamen, aber auch um komplexere Informationen, wie nachverfolgte Änderungen in einem Office-Dokument, Logdateien, den Entstehungsort eines Fotos, einer Datei zugeordnete Stichworte, Informationen über den Autor und Ähnliches (vgl. Abbildung 8). Die Masse und Ausprägung dieser Daten hängt stark von der genutzten Anwendung und dem Dateiformat ab. Obwohl die meisten dieser Zusatzinformationen prinzipiell für Nutzer erkennbar sind, werden sie von diesen oft nicht beachtet oder wahrgenommen: Während beispielsweise Firmen oft strikt darauf achten, (Bearbeitungs-)Kommentare und Notizen aus finalen Dokumenten zu entfernen, wird jedoch häufig vergessen, das Titel- oder Autor-Attribut des Dokuments zu überprüfen. Auch diese Metadaten können jedoch sensible und nicht zur Veröffentlichung bestimmte Informationen, wie einen frühen Namen des Dokuments, Informationen über den Autor und Ähnliches enthalten. In vielen Fällen führten Metadaten zur Veröffentlichung sensibler bzw. unter Geheimhaltung stehender Informationen: Beispielsweise veröffentlichten die Vereinten Nationen 2005 einen Bericht über einen Anschlag in Syrien, in dessen öffentlicher Version keine Namen mutmaßlicher Attentäter genannt wurden. In den vor der Veröffentlichung nicht entfernten Metadaten, die die Versionierungsfunktion der verwendeten Textverarbeitung speicherte, waren die zensierten Textstellen jedoch enthalten und konnten somit wiederhergestellt werden (*Zeller, 2005*). Ein weiteres Beispiel stellt deren Erzeugung durch Smartphones dar: Standardmäßig versehen diese Fotos oft automatisch mit Informationen zum Aufnahmeort und der verwendeten Kamera bzw. des verwendeten Smartphones. Werden solche Fotos mit anderen Personen geteilt, können somit unbeabsichtigte Informationen weitergegeben werden.

Andererseits vereinfachen Metadaten auch den Umgang mit Daten oder ermöglichen diesen erst: Indem Informationen hinzugefügt werden, können Dateien besser organisiert oder wiedergefunden werden, selbst die Zuordnung eines Dateinamens entspricht bereits dem Hinzufügen eines Metadatums. Insbesondere Suchfunktionen profitieren von solchen Zusatzinformationen, beispielsweise indem eine effiziente Suche nach Fotos stattfinden kann, selbst wenn der Dateiname nicht bekannt ist, indem nach in Metadaten

gespeicherten Informationen wie Aufnahmezeitpunkt oder –ort gesucht wird. Dass alle Betriebssysteme Metadaten für Suchzwecke nutzen, erleichtert die Organisation der eigenen Daten stark. Metadaten müssen jedoch immer zusammen mit den eigentlichen Daten gepflegt werden, damit die Konsistenz gewahrt bleibt.

Redundanz und Konsistenz

Bei der Verwaltung von Daten besteht häufig die Gefahr der Entstehung von Inkonsistenzen, beispielsweise wenn im Dateisystemen gespeicherte Dateien mehrere Themen betreffen und daher – zum einfacheren Zugriff – in zwei verschiedenen Ordnern als Kopien abgelegt werden. Sobald eine der beiden Dateien bearbeitet, aber die andere vergessen wird, entsteht eine Inkonsistenz. Redundant gespeicherte Daten sollten daher in vielen Fällen möglichst vermieden werden, beispielsweise indem statt eines Duplikats eine Verknüpfung an einem zweiten Ort abgelegt wird. Dies entspricht auch der im Informatikunterricht üblichen Darstellung im Kontext der Datenbanksysteme: Hier werden Redundanzen und Inkonsistenzen durch Normalisierung der Datenschemata vermieden. Neben Redundanzen können jedoch auch andere Ursachen für plötzlich inkonsistente Daten verantwortlich sein, beispielsweise Datensynchronisation: Heute besitzt im Durchschnitt jeder 2,9 mobile Geräte, wie Laptops, Smartphones oder Tablets (*Truong, 2013*), zwischen denen große Datenmengen synchronisiert werden. Da die Daten auf allen Geräten nicht nur gelesen, sondern auch bearbeitet werden, birgt dies die Gefahr der Entstehung von Inkonsistenzen, sobald nicht auf allen Geräten dieselbe aktuelle Version vorliegt. Abhängig von der Art der Daten und der eingesetzten Synchronisationswerkzeuge können die Konsequenzen dieser Inkonsistenz und die damit verbundenen für den Nutzer erkennbaren Phänomene stark variieren: Im besten Fall kann der Konflikt automatisch aufgelöst werden, ggf. muss der Nutzer manuell eingreifen, im schlimmsten Fall geht eine der beiden Versionen verloren, ohne dass der Nutzer es rechtzeitig bemerkt.

Obwohl Redundanz, auch in der Schulinformatik, meist insbesondere hinsichtlich ihrer Vermeidung betrachtet wird, gibt es auch Nutzungsszenarien, in denen diese gezielt eingesetzt wird. Neben der redundanten Speicherung von Daten beispielsweise in verteilten Datenspeichern, die genutzt wird um die Ausfallsicherheit und Verfügbarkeit zu erhöhen, wird Redundanz auch in deutlich alltagsnäheren Beispielen eingesetzt: Für die Erstellung von Datensicherungen ist sie essenziell, da jede Sicherungskopie im Prinzip eine redundante Speicherung der Daten darstellt. Obwohl prinzipiell dieselben Gefahren wie bei Redundanzen in Datenbanken bzw. allgemein bestehen, werden Inkonsistenzen einer Datensicherung zu ihrem Original teils sogar gezielt genutzt: Gerade bei manuellen Backups ist sich jeder bewusst, dass diese regelmäßig ausgeführt werden müssen, damit ein aktueller Stand des Backups (und damit die Konsistenz zum Original) sichergestellt ist. Die Inkonsistenz eines Backups zu dessen Original ist jedoch hilfreich, wenn Änderungen zurückgenommen und Daten auf einen vorherigen Stand zurückgesetzt werden sollen. Bei der Versionierung von Daten wird daher eine Menge an redundanten Kopien erzeugt, die jeweils einen unterschiedlichen zum Original inkonsistenten Stand darstellen. Der ambivalente Einsatz von Redundanz ist dabei typisch für verschiedene Prinzipien der Informatik, die sich je nach Anwendungsfall unterschiedlich äußern können. In Zusammenhang mit Datensicherungen können dabei verschiedene Phänomene auftreten, die insbesondere mit den unterschiedlichen Möglichkeiten zur Erstellung von Backups in Zusammenhang stehen: So ist es beispielsweise möglich, Backups zeitlich kontinuierlich oder diskret zu erstellen, es muss zwischen inkrementellen, differentiellen oder kompletten Backups unterschieden werden, Backups können während der Laufzeit

des Systems (online, *hot backup*) oder außerhalb dieser (offline, *cold backup*) stattfinden. Diese Varianten unterscheiden sich insbesondere in der Ressourcennutzung (z. B. Speicherplatz), der Geschwindigkeit und der Häufigkeit ihrer Erstellung. Für jeden Nutzungsfall muss gezielt entschieden werden, welche Möglichkeit genutzt wird. Durch diese Entscheidungsvielfalt wird Datensicherung in den Blickpunkt der Nutzerinnen und Nutzer gerufen, die geeignete Backupstrategien erkennen und damit umgehen müssen.

Datensicherheit und Verschlüsselung

Heute werden Daten kontinuierlich erfasst und damit auch ein großer Teil des eigenen Lebens in Form von Daten abgebildet. Dieser Anteil nimmt in Zeiten von Smartphones, anderen mobilen Geräten, Heimautomation und Ähnlichem immer stärker zu. Während die Erfassung teils offensichtlich ist, beispielsweise bei der Nutzung von sozialen Netzwerken, werden auch viele Daten im Hintergrund gewonnen, z. B. durch Sensoren, Logdateien oder Smartphones. Daten werden dabei meist nicht nur lokal auf dem eigenen PC gespeichert, sondern üblicherweise auch über (oft ungeschützte) Kommunikationskanäle wie das Internet übertragen oder auf mobilen Datenträgern, meist ohne geeignete Schutzmaßnahmen, gespeichert. Dadurch entstehen deutliche Probleme für den Schutz der Privatsphäre und der eigenen Daten, so besteht die Gefahr des Identitätsdiebstahls, finanzieller Verluste oder des Verlustes von Geschäftsdaten. Eine sichere Speicherung und Übertragung privater, persönlicher und vertraulicher Daten ist daher heute oft wichtiger als von vielen Personen erwartet. In diesem Zusammenhang existiert eine Vielfalt an auftretenden Phänomenen: Beispielsweise können von vielen Geräten die eigentlich interessanten Daten weiterhin ausgelesen werden, obwohl die Geräte mit Passwort (oder einer anderen Authentifizierungsmethode) geschützt sind – dass dieser Schutz den unbefugten Zugriff auf die Daten allenfalls erschwert, wird von den Nutzenden kaum wahrgenommen. Sie müssen sich daher den Unterschied bewusst machen zwischen dem Schutz des Geräts selbst und dem Schutz der gespeicherten Daten, der nach dem Stand der Technik, nur durch Verschlüsselung gewährleistet werden kann.

Ein weiterer Aspekt der Datensicherheit ist die Entscheidung, ob man der Korrektheit bestimmter Daten vertrauen kann oder nicht. Beispielsweise kann das Autoren-/Von-Attribut von E-Mails oder Dokumenten typischerweise einfach verändert/gefälscht werden. Weder dieses Attribut, noch der eigentliche Nachrichteninhalt, sind vor Manipulation geschützt. Dies führt zu einem weiteren Phänomen: Spam- bzw. Phishing-Mails können die eigene E-Mail-Adresse als Absender tragen ohne dass ein eigenes Gerät kompromittiert wurde. Während das Wissen darüber, dass E-Mails mit gefälschten Absendern in Umlauf sind, mittlerweile weit verbreitet ist, wird dies bei anderen Daten oft vergessen, obwohl diese oft nicht stärker gegen Manipulation geschützt sind. Methoden zur Sicherstellung der Authentizität von Daten werden im Kontext der immer stärker zunehmenden Digitalisierung rechtlich relevanter Vorgänge z. B. Behördengänge immer wichtiger. Eine mögliche Lösung stellt hierbei die digitale Signierung dar, wodurch nicht nur der Inhalt, sondern auch die zugehörigen Metadaten geschützt werden.

Datenanalyse

Obwohl immer größere Datenmengen öffentlich und frei zugänglich sind (*Open Data*), ist nur ein kleiner Personenkreis dazu in der Lage, diese wirklich zu nutzen. Prinzipiell können Datenanalysen jedoch eigentlich von jedem durchgeführt werden: Neben Daten werden zunehmend auch Werkzeuge für diesen Zweck oft kostenfrei angeboten, sodass jeder die Möglichkeit hat, z. B. durch Kombination mehrerer Datenquellen, umfangreiche Informationen zu gewinnen. Beispielsweise kann die Wahl eines Hotels auch unter Zu-

hilfenahme von Datenanalysen stattfinden: Zusätzlich zu Informationsquellen, wie Bewertungen auf Buchungsportalen, können auch Umgebungsbedingungen betrachtet werden: Die Stadt New York stellt z. B. einen Datensatz zur Verfügung, der alle Anrufe bei der Servicenummer 311 beinhaltet, bei denen es u. a. um Beschwerden über Lärm oder schlechten Straßenzustand geht. Indem dieser Datensatz mit den Ergebnissen von Restaurantüberprüfungen, die auch als Open Data zur Verfügung stehen, verknüpft wird und Analysemethoden wie Clustering, Kategorisierung und Assoziation genutzt werden, können subjektive begründete Entscheidungen über den persönlichen Hotelaufenthalt getroffen werden. Diese können auch automatisiert werden, indem Webplattformen diese Auswertungen durchführen und für den Reisenden nachvollziehbar aufbereiten. Diese Analysen eröffnen somit sowohl neue Gefahren, denen sich Schülerinnen und Schüler bewusst werden müssen, aber auch vielfältige neue Möglichkeiten.

Datenschutz und Datenspuren

Durch neue Möglichkeiten zur Speicherung und Analyse großer Datenmengen entstehen auch verschiedene Gefahren für den Datenschutz und die eigene Privatsphäre. Wie zuvor dargestellt, können beispielsweise Metadaten solche Gefahren mit sich bringen, wenn Nutzer sich deren Erzeugung nicht bewusst sind oder nicht angemessen mit diesen umgehen. Diese Gefahren verstärken sich durch die Verwendung moderner Geräte, Anwendungen und Dienste, die häufig unkontrolliert und kontinuierlich Daten über die Umgebung, den Nutzer, das Gerät und dessen Nutzung erfassen und an den Hersteller oder andere Dienste weitergeben. Diese Erfassung geschieht oft im Hintergrund und ohne Wissen des Nutzers, die erfassten Daten sind dabei außerdem häufig wesentlich umfangreicher als durch den Nutzer vermutet oder als es diesem offen kommuniziert wird. Durch die Analyse solcher Daten entstehen vielfältige Gefahren für die Privatsphäre: Beispielsweise können aus Daten oft große Teile des Privatlebens rekonstruiert und gegebenenfalls sogar für kriminelle Zwecke (wie gezielten Einbruch oder Identitätsdiebstahl) eingesetzt werden. Gleichzeitig führen Datenanalysen heute immer öfter auch zu beeindruckenden und kritisch diskutierten Ergebnissen, die ohne grundlegende Kenntnisse über die eingesetzten Methoden unerklärbar scheinen: Beispielsweise wurden durch die US-amerikanische Supermarktkette in Rahmen einer groß angelegten Analyse der über ihre Kunden vorliegenden Daten versucht, Korrelationen in diesen zu ermitteln, um besser auf verschiedene Personengruppen zugeschnitten werben zu können. Unter anderem konnten in diesem Zusammenhang ca. 25 Produkte ermittelt werden, die ausreichen, um mit hoher Wahrscheinlichkeit zu ermitteln, dass eine Kundin schwanger ist (*Hill, 2012*). Bei der Diskussion der Möglichkeiten von Datenanalysen ist es daher wichtig, ein Bewusstsein einerseits für die Chancen dieser neuen Technologien, andererseits aber auch für deren Gefahren, zu schaffen. Obwohl eine vollständige Verhinderung der Datensammlung heute nicht ohne wesentliche Einschränkungen des täglichen Lebens möglich ist, kann an vielen Stellen zumindest einer Verbreitung der Daten entgegengewirkt oder die Menge der herausgegebenen Daten beschränkt werden, indem die vorhandenen Privatsphäreeinstellungen der Anwendungen und Dienste genutzt werden. Die vier wichtigsten Fragen in diesem Zusammenhang sind daher: *Wer speichert welche Daten über mich? Wer kann auf welche Daten über mich zugreifen? Wem kann ich meine Daten anvertrauen? Was kann ich unternehmen, um weniger Daten über mich preiszugeben?*

SCHLÜSSELKONZEPTE DES DATENMANAGEMENTS

Zur verantwortungsbewussten Nutzung von Daten und datenbasierten Anwendungen und Technologien, ist ein Verständnis der zugrundeliegenden Prinzipien notwendig. Erst mit diesem können die zentralen Zusammenhänge innerhalb dieses Fachgebiets und auch die Folgen von Handlungen bewusst werden. Auch aus fachdidaktischer Sicht bietet sich eine Fokussierung auf die zentralen Aspekte des Fachgebiets an, schließlich soll allgemeinbildender Unterricht nicht möglicherweise nur kurzzeitig bedeutsamen Innovationen folgen, sondern sich auf Aspekte konzentrieren, die aus fachlicher und didaktischer Sicht langfristig relevant sind. In dem im Folgenden skizzierten Modell des Datenmanagements werden neben den Prinzipien auch die für das Fachgebiet relevanten Praktiken sowie die teils veränderlichen Kerntechnologien dargestellt:

Kerntechnologien stellen konkrete Anwendungen bzw. Technologien dar, die einen klaren Bezug zu Datenmanagement haben. Andererseits repräsentieren sie aber 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 aber auch als Entscheidungskriterien bei der Auswahl eines Systems für einen konkreten Anwendungszweck herangezogen werden.

Mechnismen 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.

Einen Überblick über das Modell der Kerntechnologien, Praktiken und zentralen Prinzipien des Datenmanagements gibt Abbildung 9:

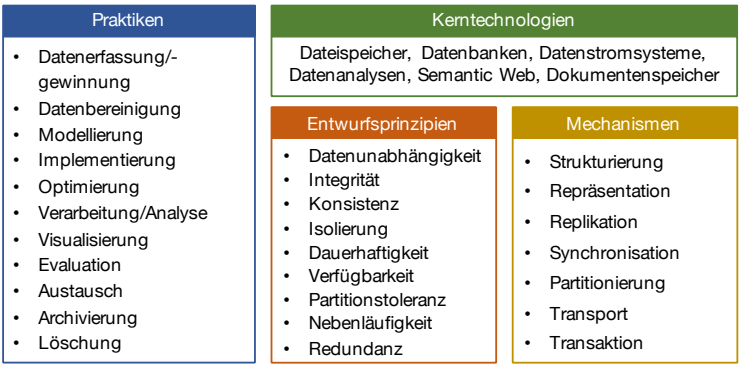


Abbildung 9: Modell der Schlüsselkonzepte des Datenmanagements.

Die im Modell enthaltenen Begriffe werden im folgenden kurz charakterisiert, eine detailliertere Darstellung inklusive einer ausführlicheren Beschreibung aller Begriffe und der jeweiligen Bedeutung für den Unterricht steht im Internet zur Verfügung:
https://dataliteracy.education/schluesselkonzepte

KERNTECHNOLOGIEN

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

Datenbanken werden unterteilt in relationale und nicht-relationale Datenbanksysteme: Die Nutzung des relationalen Modells setzt eine klare Strukturierung der Daten voraus und erzwingt deren Einpassung in ein Datenschema, sodass die Erfüllung des ACID-Prinzips, das die traditionellen Anforderungen an Datenbanken berücksichtigt, sichergestellt wird. Bei nichtrelationalen Modellen werden u. a. Integritätsanforderungen weniger stark gewichtet. Im Gegenzug lassen nichtrelationale Modelle jedoch verschiedene andere Strukturierungsmöglichkeiten für Daten zu und eröffnen mehr Freiräume.

Datenstromsysteme sind nicht für die dauerhafte Speicherung von Daten, sondern für eine schnelle und effiziente Analyse optimiert. Durch Nutzung des Datenstromprinzips erlauben sie eine sofortige Analyse neuer Daten, während gleichzeitig wenig Overhead entsteht, der die Analyse ausbremst, sodass Ergebnisse üblicherweise sofort bereitstehen.

Datenanalysen bzw. die 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.

Semantic Web beschreibt die systematische Anreicherung von im Web zugreifbaren Daten 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, indem auch die inneren Strukturen der Dokumente zumindest zum Teil zur Strukturierung und zum Zugriff verwendet werden. Die Struktur wird aber nicht vorher (z. B. in Form eines Schemas) definiert und es wird keine einheitliche Struktur der Dokumente erzwungen, sodass eine geringere Strukturierung als bei relationalen Datenbanken vorliegt.

PRAKTIKEN

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, die Zugänglichmachung durch Strukturierung, aber auch die Recherche nach und Abfrage von geeigneten Datensätzen sein. Gegebenenfalls können Daten mehrerer Quellen zusammengeführt werden.

Datenbereinigung ist nötig, um eine hohe Datenqualität sicherzustellen und beispielsweise ungültige (z. B. Werte außerhalb des zulässigen Wertebereichs), falsche (z. B. erkennbare Messfehler) oder ungeeignet formatierte Daten (z. B. Datumsangaben als Klartext oder im falschen Format) zu auszufiltern und/oder zu korrigieren.

Modellierung wird insbesondere zur klaren und verständlichen Strukturierung von Daten und zur Verdeutlichung von Zusammenhängen zwischen Datensätzen genutzt, aber auch um einen Überblick über Datensätze und deren Struktur zu bekommen.

Implementierung des Datenmodells in einem Datenmanagementsystem ermöglicht die Nutzung und Speicherung von Daten und ist damit grundlegend für weitere Praktiken.

Optimierung findet insbesondere zur Erhöhung des Effizienz bei der Speicherung der Daten und beim Zugriff auf diese, aber auch zur Erhöhung der Analysequalität statt.

Verarbeitung/Analyse umfasst die Aggregation von Daten, aber auch die Erzeugung neuer Informationen aus Daten unter Nutzung verschiedener Datenanalysemethoden.

Visualisierung von Daten wird genutzt, um die Analyseergebnisse verständlich und gut erfassbar für den Menschen aufzubereiten.
Evaluation der Ergebnisse umfasst, neben der Beurteilung der eigentlichen Ergebnisse, auch die Einschätzung der Qualität des Datensatzes und des Analyseansatzes.
Austausch von Daten kann die Analyseergebnisse aber auch die Originaldaten umfassen und auf verschiedenen Wegen stattfinden.
Archivierung von Daten wird genutzt, um diese für zukünftige und ggf. nicht vorhersehbare Zwecke zur Verfügung zu stellen. Durch Archivierung wird der Datenlebenszyklus für eine gewisse Zeit pausiert, aber spätere Nutzungen ermöglicht.
Löschung von Daten kann bspw. erfolgen, um Speicherplatz zu gewinnen oder das Persönlichkeitsrecht von Personen zu wahren. Durch die Löschung wird eine spätere Verwendung der Daten unterbunden und der Datenlebenszyklus beendet.

ENTWURFSPRINZIPIEN

Datenunabhängigkeit: Die Arbeit mit Daten ist ohne Kenntnis ihrer internen Speicherung möglich, da Details der Implementierung durch Abstraktion 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 beeinflussen sich nicht gegenseitig.
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: Die mehrfache Speicherung derselben Daten wird zur vermieden, um Inkonsistenzen zu verhindern, kann aber auch die Ausfallsicherheit des Systems erhöhen.

MECHANISMEN

Strukturierung von Daten beschreibt Maßnahmen, die die Suche nach und den Zugriff auf Daten ermöglichen und schnell und effizient gestalten. Es handelt sich beispielsweise um die Anreicherung von Daten mit Metadaten, die Einordnung in 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 von Daten auf verschiedenen Medien bzw. Datenspeichern, u. a. 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. Gleichzeitig wird darunter aber auch die Replikation von Daten verstanden, wenn diese auch eine Konflikterkennung beinhaltet.
Partitionierung befasst sich mit der Speicherung über verschiedene Datenspeicher hinweg. Dabei zu speichernde Daten beispielsweise zur Erhöhung der Zugriffsgeschwindigkeit oder aufgrund großer Datenmengen, auf verschiedene Teilsysteme bzw. Datenspeicher verteilt, statt sie, wie bei der Replikation, in Kopie auf diesen zu speichern.

Transport beschreibt die Übertragung von Daten innerhalb eines Systems und über Systemgrenzen hinweg. Zentrale Aspekte stellen dabei beispielsweise die Sicherheit und Vertraulichkeit der Übertragung dar.
Transaktionen bezeichnen Gruppen von Abfragen, die entweder alle erfolgreich durchgeführt oder alle abgebrochen werden. Sie werden insbesondere zur Isolation nebenläufiger Anfragen und zur Erhöhung der Fehlertoleranz eingesetzt.

NUTZEN FÜR DEN UNTERRICHT

Die im dargestellten Entwurfsprinzipien und Mechanismen stellen wichtige Themen für den Informatikunterricht dar, die nicht nur zeitbeständig, sondern oft auch in verschiedenen Bereichen der Informatik einsetzbar sind[11]. Sie sind daher auch Kandidaten für fundamentale Ideen der Informatik (vgl. Schwill, 1993). Gleichzeitig wird durch diese jedoch auch Sinn Zweck von Datenmanagement deutlich: neben der Gewinnung von Informationen sind die Sicherheit und die Verfügbarkeit von Daten übergeordnete Ziele. Die Kerntechnologien können insbesondere zur Kontextualisierung eingesetzt werden, da sie Bezüge zur Lebenswelt der Schüler herstellen: Mit den Technologien und ihren Auswirkungen kommen Schüler am offensichtlichsten im Alltag in Berührung und entwickeln daher Vorstellungen über bzw. Erklärungen zu diesen. Durch Einbeziehung der Praktiken des Datenmanagements kann ein handlungsorientierter Unterricht ermöglicht werden, in dem Lernende selbst professionelle Tätigkeiten im Unterricht auf didaktisch angemessenem Niveau nachvollziehen können. Die Praktiken können als Lebenszyklus der Daten aufgefasst werden (vgl. Abbildung 10), der als Leitlinie für den datenorientierten Informatikunterricht dient.

Insgesamt eignet sich das vorgestellte Modell somit zur fachlichen Fundierung des datenorientierten Informatikunterrichts, als Basis zur Curriculumsentwicklung, aber auch zur Gewinnung von Ideen für die Unterrichtsgestaltung, die je nach Zielsetzung an einer oder mehreren der vier Kategorien ausgerichtet sein kann. Das Modell stellt dabei jedoch eine fachliche Perspektive in den Vordergrund, es kann und soll daher nicht das Ziel des Informatikunterrichts sein, alle Aspekte dieses Modells in vollem Umfang abzubilden. Für eine weitergehende Betrachtung des Themas ist es daher nötig, eine informatikdidaktische Perspektive einzunehmen, wie es im Folgenden geschieht.

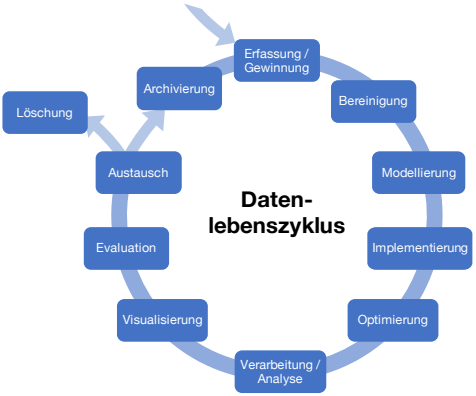


Abbildung 10: Datenlebenszyklus, abgeleitet von den Praktiken des Datenmanagements.

[11] Diese Bezüge werden in der ausführlichen Darstellung im Web expliziert.

GRUNDLEGENDE DATENKOMPETENZEN: KOMPETENZMODELL FÜR DEN INFORMATIKUNTERRICHT

In immer mehr Kontexten werden heute Datenkompetenzen gefordert, die u. a. im Informatikunterricht erworben werden müssen. Aus informatischer Perspektive sind diese, oft als *Data Literacy* bezeichneten Kompetenzen insbesondere durch das Fachgebiet Datenmanagement, aber auch durch die Data Science[12], fundiert. Unter *Data Literacy* wird nach der verbreitet anerkannten Definition des Hochschulforums Digitalisierung (2017) die Fähigkeit verstanden, "*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.*" An diesen Datenkompetenzbegriff setzt eine stärker didaktisch orientierte Betrachtung des Themengebiets an: Durch die Strukturierung und Explizierung von Kompetenzbereichen mit Fokus auf den allgemeinbildenden Schulunterricht kann eine Reduktion der Komplexität stattfinden, aber auch eine adäquate Hervorhebung zentraler Bereiche erreicht werden.

Bei der Entwicklung des Datenkompetenzmodells wurde entschieden, die bewährte und häufig als hilfreich angesehene Struktur der die GI-Empfehlungen für Bildungsstandards (vgl. *AK Bildungsstandards, 2008* und *AK Bildungsstandards SI, 2016*) prägenden Kompetenzmodelle zu übernehmen, die die informatischen Kompetenzen in zwei Dimensionen aufteilen: "*Die Inhaltsbereiche charakterisieren mindestens zu erwerbende fachliche Kompetenzen. Die Prozessbereiche beschreiben, auf welche Art und Weise die Schülerinnen und Schüler mit den genannten Fachinhalten umgehen sollen.*" (*AK Bildungsstandards, 2008*) Diese Betrachtung betont insbesondere den hohen Wert sowohl der zu erwerbenden inhalts- als auch der prozessorientierten Kompetenzen, die immer in enger Beziehung zueinander stehen.

Auf Basis des Modells der Schlüsselkonzepte des Datenmanagements sowie einer inhaltlichen Charakterisierung der Data Science, wurde das in Abbildung 11 dargestellte

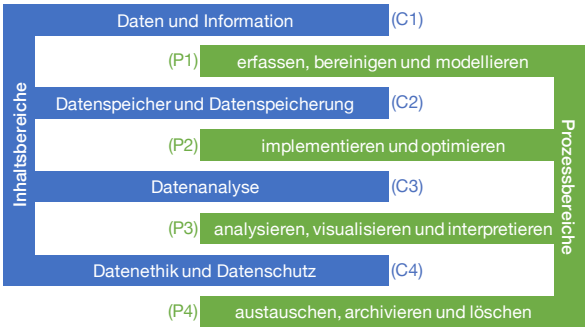


Abbildung 11: Das entwickelte Datenkompetenzmodell.

[12] *Data Science* wird häufig als eigene Wissenschaft von der systematischen Verarbeitung und Analyse von Daten betrachtet, die insbesondere mathematische und informatische Konzepte miteinbezieht und diese jeweils in konkreten Anwendungskontexten verwendet.

Datenkompetenzmodell entwickelt, dessen Inhalts- und Kompetenzbereiche im Folgenden ausführlicher beschrieben werden.

INHALTSBEREICHE DES DATENKOMPETENZMODELLS

Daten und Information vereint grundlegende fachliche Aspekte im Zusammenhang mit Daten und Information: Er berücksichtigt u. a. die Unterscheidung dieser beiden Begriffe, die unterschiedliche Aussagekraft verschiedener Daten (im Sinne des Informationsgehaltes nach *Shannon, 1984*), Methoden zur Erfassung von Information in Form von Daten, aber auch zur Strukturierung von Daten durch die Anreicherung mit zusätzlicher Information beispielsweise durch Metadaten, sowie die Unterscheidung großer und kleiner Datenmengen bezüglich deren Aussagekraft/Informationsgehalt.

Datenspeicher und Datenspeicherung berücksichtigt vielfältige Aspekte aus dem Bereich des Datenmanagements: Es beinhaltet grundlegende Mechanismen von Datenspeichern, wie die Repräsentation von Daten bis hin zur physikalischen Ebene, deren Replikation, Synchronisation oder Transport. Gleichzeitig werden 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.

Datenanalyse befasst sich mit 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. allgemein deskriptive Methoden, sondern auch explorative, korrelationsbasierte und auf Vorhersagen ausgerichtete Datenanalysen, die oft im Zusammenhang mit maschinellem Lernen stattfinden und Methoden aus diesem Bereich miteinbeziehen.

Datenethik und Datenschutz 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 durch Maßnahmen wie Verschlüsselung miteinbezieht.

Zusammenfassend berücksichtigen die vier Inhaltsbereiche eine Vielfalt von Aspekten der Data Literacy, aber auch generell der Informatik: Während der zweite Inhaltsbereich 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. Vielfältige Bezüge zu verschiedensten Themen der Informatik finden sich jedoch auch in den beiden eher spezialisierten Bereichen: Beispielsweise spielen im zweiten Inhaltsbereich, bei der Thematisierung von Zugriffsmöglichkeiten für Daten, auch Aspekte der Rechnerkommunikation eine nicht zu vernachlässigende Rolle. Im dritten Bereich sind hingegen klare Bezüge zur Algorithmik erkennbar.

PROZESSBEREICHE DES DATENKOMPETENZMODELLS

erfassen, bereinigen und modellieren: Dieser Prozessbereich beschäftigt sich mit den frühen Phasen der Arbeit mit Daten. Dabei werden drei Bereiche betrachtet, die 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 stattfinden und 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 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?*

implementieren und optimieren: Diese beiden Tätigkeiten können zu verschiedenen Zeitpunkten im Datenlebenszyklus stattfinden, insbesondere bei der Implementierung des Datenmodells in einem Datenspeicher sowie bei der Speicherung der Daten. Andererseits ist, je nach gewählter Methode und Werkzeug, gegebenenfalls auch bereits bei der Datenerfassung oder später bei der Analyse eine entsprechende Implementierung und Optimierung, beispielsweise von einfachen Algorithmen, nötig. Entsprechend können auch hier verschiedene Ziele verfolgt werden, die alle anderen Bereiche des Datenlebenszyklus betreffen. 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?*

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 Daten einfacher erfassbar zu machen und zur Unterstützung der Interpretation, werden verschiedene Visualisierungstechniken eingesetzt. Teilweise kann auch die eigentliche Datenanalyse mit visuellen Methoden durchgeführt werden. Der dritte Prozessbereich befasst 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?*

austauschen, archivieren und löschen: Der vierte Prozessbereich ist während des gesamten Erfassungs-, Speicherungs- und Analyseprozesses von Bedeutung. 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 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?*

EXEMPLARISCHE KOMPETENZEN

Wie auch in den GI-Empfehlungen für Bildungsstandards in Informatik, werden Kompetenzen im Sinne des Datenkompetenzmodells immer zugleich einem Prozess- und einem Inhaltsbereich zugeordnet. Dabei sind nicht nur nebeneinanderliegende Bereiche miteinander kombinierbar, sondern grundsätzlich jeder Bereich mit jedem.

Um einen Eindruck des Breite der durch das Datenkompetenzmodell erfassten Kompetenzen zu vermitteln, werden in Tabelle 1 exemplarische Datenkompetenzen dargestellt und dem Modell zugeordnet.

	P1 erfassen, modellieren und bereinigen	P2 implementieren und optimieren	P3 analysieren, visualisieren und interpretieren	P4 austauschen, archivieren und löschen
C1 Daten und Information	• wählen Sensoren zur Datenerfassung aus • strukturieren Daten in Hinblick auf die geplante Analyse • überprüfen, ob die erfassten Daten die korrekte Information widerspiegeln	• implementieren die Erfassung der Daten • implementieren Algorithmen zur Abfrage Daten von Web-APIs • diskutieren Möglichkeiten zur Optimierung der Erfassung	• kombinieren Daten um neue Informationen zu gewinnen • verdeutlichen Informationen durch Visualisierungen • interpretieren Daten und Analyseergebnisse	• entscheiden, ob Originaldaten mit anderen geteilt werden • wählen Daten zur Archivierung aus • entscheiden über die Löschung nicht mehr benötigter Daten
C2 Datenspeicher und Datenspeicherung	• wählen ein Datenmodell aus • strukturieren Daten, um sie adäquat zu speichern und abzurufen • stellen das Datenmodell in geeigneter Form dar	• wählen einen Datenspeicher aus • nutzen Möglichkeiten zur Erhöhung der Zugriffseffizienz • erhöhen die Speicherplatzeffizienz durch Nutzung von Kompression	• binden vorhandene Datenquellen an ein Analysewerkzeug an • nutzen geeignete Austauschformate • speichern Analyseergebnisse auf angemessene Art und Weise	• bestimmen, wer zu welchem Zweck Zugriff auf Daten erhält • bestimmen Zugriffsrechte auf Daten • diskutieren Problem, die durch Löschung von Daten entstehen
C3 Datenanalyse	• überprüfen, welche Daten die Qualität der Ergebnisse beeinflussen • strukturieren Daten in geeigneter Weise für deren Analyse • kombinieren Daten aus unterschiedlicher Datenquellen	• implementieren Analysealgorithmen • nutzen Stellschrauben zur Beeinflussung der Analysequalität • optimieren Datenanalysen um deren Qualität zu erhöhen	• wählen geeignete Analysemethoden aus • visualisieren Daten und Analyseergebnisse • interpretieren die Ergebnisse von Datenanalysen	• entscheiden, mit wem welche Analyseergebnisse geteilt werden • diskutieren eine Archivierung von Originaldaten • entscheiden, welche Informationen über den Analyseprozess geteilt werden
C4 Datenethik und Datenschutz	• berücksichtigen ethische Aspekte bei der Datenerfassung • entscheiden, ob es vertretbar ist, Daten aus verschiedenen Quellen zusammenzuführen • diskutieren eine kontinuierliche Datenerfassung hinsichtlich Ethik und Datenschutz	• diskutieren Wege zur Anonymisierung und Pseudonymisierung von Daten • schließen bestimmte Daten von der Speicherung auf Basis ethischer Überlegungen aus • vergeben Zugriffsrechte unter Berücksichtigung des Datenschutzes	• diskutieren ethische Einflüsse der durchgeführten Analyse • entscheiden, ob Analyseergebnisse ausreichend anonymisiert sind • berücksichtigen durch die Analyse hervorgerufene Datenschutzprobleme	• diskutieren, ob eine Datenspeicherung für potentielle zukünftige Nutzung vertretbar ist • löschen von personenbezogene Daten auf sichere Art und Weise • finden Methoden zur angemessenen Anonymisierung

Tabelle 1: Exemplarische Datenkompetenzen, eingeordnet in die Bereiche des Datenkompetenzmodells.

PRAKTISCHE IDEEN FÜR DEN INFORMATIKUNTERRICHT

Basierend auf dem Datenkompetenzmodell wurden drei Beispiele für die Förderung von Datenkompetenzen im Informatikunterricht entwickelt:

1. **Grundlegende Datenanalysen mit Snap!Twitter**
   Das Werkzeug Snap!Twitter erlaubt es unter Nutzung blockbasierter Programmierung, einfache Analysen des Twitterdatenstroms durchzuführen, ohne Kenntnisse in der Programmierung oder bezüglich Datenanalysen vorauszusetzen. Dieses Beispiel eignet sich sehr gut für einen einfachen und motivierenden Einstieg in die Thematik der Datenanalysen. Dabei können auch Grundsätze von Datenstromsystemen (vgl. Seite 12) thematisiert werden.

2. **Fortgeschrittenere Datenstromanalysen mit Snap!DSS**
   Während Snap!Twitter auf die Datenquelle Twitter eingeschränkt ist, erlaubt das Snap!-Datenstromsystem *Snap!DSS* die Nutzung beliebiger Datenquellen, insofern sie in der Programmiersprache Snap! bzw. einem ihrer Derivate abfragbar sind. Es ist somit beispielsweise die Analyse von Sensordaten möglich, die von einem Arduino-Mikrocontroller unter Nutzung von Snap4Arduino abgerufen werden.

3. **Einführung in das Data Mining: Datenanalyse und Vorhersage**
   Dieses Unterrichtskonzept, das für mindestens 3-4 Doppelstunden ausgelegt ist, erlaubt einen ersten Einblick in die korrelationsbasierte Datenanalyse und die Vorhersage anhand der Analyseergebnisse. Dabei werden unter anderem Aspekte wie Klassifikation und Klassifikationsbäume, Korrelation vs. Kausalität, Analysequalität und ethische Aspekte thematisiert.

Diese drei Unterrichtsbeispiele werden im folgenden etwas ausführlicher vorgestellt und ein erster Einblick in die jeweiligen Möglichkeiten gegeben. Weitere Informationen zu den Unterrichtsbeispielen sind auf der Projektwebseite zu finden:

http://dataliteracy.education/unterrichtsmaterial

SNAP!TWITTER

Überblick

Um Datenstromanalysen 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. Mit Snap!Twitter wurde ein solches Werkzeug entwickelt: Um eine 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, das Werkzeug nicht 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 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 gewählt, da dieser eine kostenfreie Möglichkeit zum Zugriff auf ausreichend umfangreiche Datenmengen zur Verfügung stellt, die live erzeugt werden, sich somit kontinuierlich verändern und entsprechend einen typischen Einsatzzweck für Datenstromanalysen darstellen. Durch Nutzung des Filter-Datenstroms von Twitter konnte die Chance erhöht werden, 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 Datenanalyse in Snap! zu implementieren, wurde ein Ansatz gewählt, der auch auf andere Datenquellen übertragen werden kann. Anstatt den eigentlichen Quellcode von Snap! zu modifizieren, wurden alle neuen Funktionen nativ implementiert, indem die zur Verfügung gestellten Blöcke und insbesondere die Möglichkeit, JavaScript zu nutzen, verwendet wurden. Somit wird eine flexible Portabilität der Erweiterung auch auf andere Snap!-Varianten und eine hohe Kompatibilität zu zukünftigen Updates sichergestellt. Im Gegensatz zu einer, ersten Variante von Snap!Twitter wurde die aktuelle Version als Node.JS-Anwendung konzipiert und kann auf diese Weise auf einem beliebigem Server in Betrieb genommen werden. Ein durch den Autor betreuter Server steht zur Verfügung. Dieser kann auch für Erprobungen oder den Schulunterricht genutzt werden – falls daran Interesse besteht einfach Kontakt aufnehmen.

Snap!Twitter stellt für die Nutzer verschiedene Möglichkeiten zur Verfügung:

- Eine Bibliothek an Blöcken zum Zugriff auf alle Attribute, die ein Tweet mitbringt
- Möglichkeit zur Nutzung von Kartendarstellungen, z. B. um Tweets dort anzuzeigen
- Eine flexibel einsetzbare Diagramm-Bibliothek

Einen Einblick in diese Möglichkeiten gibt Abbildung 12.

Einsatzmöglichkeiten

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 Grundzüge von Datenstromanalysen kennenlernen und verstehen. Das Ziel im Unterricht ist insbesondere, eigene einfache Datenanalysen durchführen und Schlüsse aus den gewonnenen Ergebnissen ziehen zu können. Dies kann im Unterricht vielfältig kontextualisiert werden: Insbesondere sind als Kontext Datenanalysen gut geeignet, denen Schülerinnen und Schülern heute bereits im Alltag begegnen, beispielsweise bei Verwendung sozialer 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 z. B. Preisanpassungen für bestimmte Kundengruppen durchzuführen. Um zu verstehen, wie

solche Analysen und Vorhersagen überhaupt funktionieren, welches Potenzial und welche Risiken diese bergen und um fundiert zu entscheiden, 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:

Anhand einfacher Klassifikationsaufgaben können Lernende 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 12) erstellt werden.

Abbildung 12: Visualisierung von Tweets auf einer Karte und im Diagramm, eingefärbt nach der jeweiligen Sprache.

Die Diskussion möglicher aus der Klassifikation gewinnbarer Aussagen offenbart den Schülerinnen und Schülern die Grenzen dieser Methode: Beispielsweise 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.

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 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, das Prinzip der Clusterbildung nachzuvollziehen, ohne die mathematischen Grundlagen verstehen zu müssen. 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?*

Wenn anhand von Daten Entscheidungen oder Vorhersagen getroffen werden sollen, müssen entsprechende Regeln definiert bzw. gefunden werden. Diese werden in Form von Assoziationen aufgestellt, die Zusammenhänge zwischen Attributen 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 eine entsprechende Kartendarstellung kann 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.

Ein erster Einstieg in die Datenanalyse kann beispielsweise durch das automatisch in Snap!Twitter angezeigte Tutorial oder einfache Fragestellungen erfolgen, zum Beispiel:

- *In welchem Land / welcher Region sind die Twitternutzer am aktivsten?*
- *Über welchen Präsidentschaftskandidat in den USA wird am meisten getwittert?*
- *Fällt die Antwort auf Frage 2) regional unterschiedlich aus?*
- *Welche der Sprachen Deutsch, Englisch oder Französisch sind auf Twitter am beliebtesten?*
- *Vermutlich stellen Sie in Aufgabe 4 fest, dass nur wenige deutsche Tweets vorhanden sind. Eine Erklärung könnte sein, dass in Deutschland Twitter kaum genutzt wird. Überprüfen Sie diese Hypothese. Gibt es weitere mögliche Erklärungen?*
- *Sind Twitter-User mit mehr Followern auch selbst aktiver?*
- *Warum reicht ein Datenstromsystem alleine nicht aus, um die Trends auf der Startseite von Twitter zu generieren?*

Später kann dann, versucht werden, eine komplexere Fragestellung auf empirische Weise zu beantworten. Verschiedene Ideen hierzu sind die folgenden Fragen bzw. Thesen, die mit Hilfe von Snap!Twitter (z. T. eingeschränkt) überprüft werden können, sodass es sicherlich ein wichtiger Aspekt ist, die durchgeführte Analyse kritisch zu reflektieren:

- *Eine Studie behauptet, dass Menschen in Japan wesentlich weniger soziale Kontakte im realen Leben pflegen als solche in den USA.*
- *Menschen in den USA haben oft die Lieblingsfarbe blau – in Europa ist hingegen oft grün.*
- *Im Internet sprechen die meisten Personen eher Englisch als ihre Muttersprache.*
- *Thema A ist derzeit mehr diskutiert als B.*
- *Personen aus England sind im Internet oft wesentlich beliebter.*
- *Die Nutzung von sozialen Medien wie Twitter ist regional geprägt – in welchen Ländern ist Twitter weniger beliebt?*

SNAP!DSS

Überblick

Basierend auf den Ergebnissen der Erprobung von Snap!Twitter wurde eine weitere Variante des Werkzeugs entwickelt, die den Nachteil aufgreift, dass Snap!Twitter auf eine Datenquelle eingeschränkt ist, sodass eine in Bezug auf die Datenquelle flexibler einsetzbare Alternative entstand. Dafür wird die Möglichkeit eröffnet, verschiedenste Datenquellen zu nutzen, solange diese durch *reporter*-Blöcke, d. h. Blöcke die Werte zurückliefern, 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, deren Nutzung nicht ohne weiteren Programmieraufwand möglich ist. Zusätzlich zur Unabhängigkeit von der Datenquelle implementiert das Snap!-Datenstromsystem Snap!DSS alle zentralen Funktionen von Datenstromsystemen und der üblichen Abfragesprache *CQL* (*Continuous Query Language*). Im Vergleich zu Snap!Twitter werden die Abfragen in dieser deklarativ statt imperativ formuliert, wodurch alle Anfragen

als eigene Einheit betrachtet werden können, die auch verschachtelbar sind. Trotz dieser umfangreichen Möglichkeiten wird das Werkzeug jedoch, genau wie Snap!Twitter, 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 Mikrocontrollern bietet, wird dies z. B. in Snap4Arduino ermöglicht.

Das implementierte Werkzeug bietet daher folgende Kernfunktionalitäten:

- Nutzung mit jedem Derivat von Snap! möglich, das keine funktionalen Einschränkungen vornimmt.
- Implementierung aller 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, d.h. einer besonders bei der Datenstromanalyse verbreitete Technik, die es ermöglicht, in kontinuierlich ausgeführten Abfragen relative anstatt absoluter Selektionskriterien anzugeben.
- Nutzung von Ausgangsdatenströmen einer Analyse als Eingangsdatenstrom für eine weitere (Verschachtelung von Anfragen).
- Kontinuierliche Auswertung der Anfragen im Hintergrund.
- Eine flexibel einsetzbare Diagramm-Bibliothek.

Einen Einblick in diese Möglichkeiten gibt Abbildung 13.

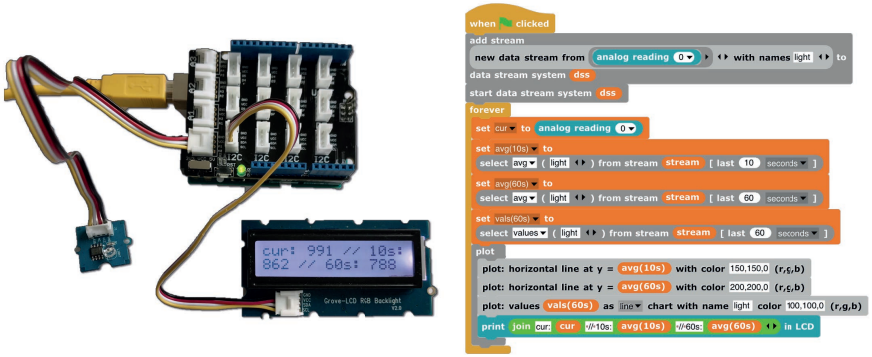


Abbildung 13: Arduino-Mikrocontroller mit Lichtsensor und Display sowie Code zur Auswertung der Sensordaten in Snap4Arduino mit der Erweiterung Snap!DSS.

Einsatzmöglichkeiten

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 ein wichtiges Funktionsprinzip moderner Technologien 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 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 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* 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 nur ineffizient lösbar (Speicherung aller Werte) oder eher komplex (z. B. durch Nutzung eines Ringpuffers als Datenstruktur) wäre. In der obigen Abbildungen 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.

Dieses Beispielprojekt zeigt 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.

DATA MINING: DATENANALYSE UND VORHERSAGE

Während die zuvor vorgestellten Ideen allgemein für die Thematisierung einfacher Datenanalysen ausgelegt sind, werden in der im Folgenden beschriebenen Unterrichtsidee zwei besonders spannende und einflussreiche Thema des modernen Datenmanagements

konkret aufgegriffen, *Data Mining* und die *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 Entscheidungen darüber treffen zu können, ob man konkrete Informatiksysteme und/oder Dienste nutzt, die Daten in großem Umfang sammeln und analysieren, ist es jedoch 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, 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.

Überblick über das Unterrichtskonzept

Um die zuvor charakterisierte Lücke im aktuellen Informatikunterricht anzugehen, zielt das im Folgenden vorgestellte Unterrichtskonzept darauf ab, ein Grundverständnis von Datenanalysen und Prognosen zu fördern. Der geplante Unterrichtsverlauf ist auf vier Doppelstunden ausgelegt, kann aber gegebenenfalls individuell angepasst werden, da an verschiedenen Stellen eine deutlich vertieftere oder etwas oberflächlichere Betrachtung möglich ist. Die Unterrichtssequenz wurde so gestaltet, dass auf keinerlei Vorwissen aufgebaut werden muss, sodass diese flexibel in den Unterricht integrierbar ist. Die Thematik ist sowohl für Sekundarstufe I als auch Sekundarstufe II geeignet. In der Unterrichtssequenz wurden insbesondere folgende, in das zuvor dargestellte Kompetenzmodell eingeordnete, Kompetenzen angestrebt:

- erläutern, warum und wie aus gespeicherten Daten verschiedene und ggf. neue Informationen gewonnen werden können (C1/P3)
- charakterisieren des Unterschieds zwischen korrelations- und kausalitätsbasierten Zusammenhängen in Daten sowie der jeweiligen Aussagekraft (C1/P3, z. T. C4/P3)
- skizzieren des Ablaufs einer (korrelationsbasierten) Datenanalyse (C3/P3)
- charakterisieren einer typischen Analysemethode und erläutern des zugrundeliegenden Prinzips an einem geeigneten Beispiel (C3/P3)
- durchführen einfacher Datenanalysen unter Nutzung einer üblichen Methode, händisch sowie unter Nutzung eines geeigneten Softwarewerkzeugs (C3/P3)
- prognostizieren fehlender Attribute eines Datensatzes unter Rückgriff auf eine selbst durchgeführte Datenanalyse (C3/P3)
- bewerten des Ergebnisses der Vorhersage und erläutern von Ideen zur Verbesserung (C3/P3)
- reflektieren der Ergebnisse unter Einbeziehung ethischer und gesellschaftlicher Gesichtspunkte (C4/P3)

Der Schwerpunkt der Unterrichtssequenz liegt aus inhaltlicher Perspektive auf den Themen, die im Informatikunterricht bislang wenig thematisiert werden, und klammert daher den Inhaltsbereich C2, der im Rahmen von Datenbanken häufig vorkommt, aus.

Aufgrund der zeitlichen Knappheit, die notwendig war, um eine flexible Einbettung der Sequenz in den Unterricht zu ermöglichen, musste auch bei den Prozessbereichen eine Schwerpunktsetzung stattfinden, sodass mit bereits erfassten und vorstrukturierten Daten gearbeitet wurde, sodass die Prozessbereiche P1 und P2 außen vor gelassen wurden. Auch der Prozessbereich P4 stellt allenfalls einen Randaspekt dar.

Um die angestrebten Ziele zu erreichen, wurde der Unterrichtsverlauf in zwei Blöcke geteilt: Anfangs erfolgt eine Einführung in den Datenanalyseprozess ohne Nutzung digitaler Werkzeuge. Dabei findet ein erster Kontakt mit den zur Analyse und Vorhersage genutzten Methoden und Prinzipien, aber auch dabei auftretenden Schwierigkeiten statt. Um die Automatisierbarkeit und das damit einhergehende Potential einschätzen zu können, wird daraufhin ein Softwarewerkzeug zur Durchführung einer Datenanalyse am PC genutzt, der ein größerer Datensatz zugrunde liegt. Dieser wurde so gewählt, dass die Analyse die Schülerinnen und Schüler direkt betrifft, um eine kritische Diskussion der Ergebnisse und der Analyse selbst anzustoßen.

In der ersten Doppelstunde stehen die Grundzüge des Analyseprozesses im Vordergrund. Dazu wird die Bedeutung von Datenanalysen für den Alltag anhand eines Zeitungsartikels motiviert, der den Versuch eines US-Einzelhändlers beschreibt, zu erkennen, ob Kundinnen schwanger sind, um ihnen gezielt Werbung zuzusenden (*Hill, 2012*). Dieser Artikel erlaubt eine Diskussion über mögliche Attribute, anhand derer eine Schwangerschaft vorhergesagt werden könnte, sodass die Schülerinnen und Schüler schnell einen Zugang zum Thema bekommen. Daraufhin wird der Wert und Nutzen von Daten für verschiedene Zwecke, Unternehmen und Geschäftsmodelle diskutiert, hier können die Schülerinnen und Schüler eigene Beispiele einbringen. Dieses Unterrichtsgespräch wird in Richtung der Begriffe *Kausalzusammenhang*, *Korrelation* und *Prognose* gelenkt, die daraufhin eingeführt werden. Aus den Ideen der Lernenden wird ein Ablaufmodell der Datenanalyse erstellt.

Darauf basierend wird in der zweiten Doppelstunde der Weg vom Datensatz zur Prognose händisch durchlaufen, sodass verschiedene Prinzipien dieser Analysen erkennbar werden. Dazu wird ein vorgegebener einfacher und fiktiver Datensatz aus dem Kontext Onlinehandel auf Zusammenhänge hin untersucht und diese als Regeln (im Sinne von Assoziationen) formuliert. Um sie einfacher erfassbar und anwendbar zu machen, wird ein (binärer) Entscheidungsbaum als Darstellungsform eingeführt und als Klassifikationsbaum genutzt. Die gefundenen Regeln/Assoziationen werden auf einen weiteren Datensatz mit unbekannten Attributwerten angewandt. Durch die nicht immer eindeutige Möglichkeiten zur Regelanwendung und entsprechende Entscheidungsspielräume wird eine Reflektion über die angestrebte Analysequalität angestoßen. Zum Abschluss der Stunde wird in die zentrale Aufgabe der nächsten Doppelstunde eingeführt, in der das Ziel ist, Schulnoten aus echten Personendaten und vorherigen Noten vorherzusagen.

In der dritten Doppelstunde wird ein frei zugänglicher Datensatz (*Cortez und Silva, 2008*) mit (anonymisierten) Daten von über 600 portugiesischen Schülerinnen und Schülern (verschiedene persönliche Daten und Punktzahlen in drei Prüfungen) analysiert. Mithilfe des Werkzeugs *Orange* (vgl. unten) werden, basierend auf einer Stichprobe des Datensatzes, Regeln zur Vorhersage der dritten Punktzahl aus den restlichen Daten ermittelt (Abbildung 14) und diese auf den Rest der Schülerinnen und Schüler angewendet, mit dem fiktiven Ziel, den Korrekturaufwand dieser Prüfung für die Lehrkraft durch Nutzung von Analyse- und Prognosemöglichkeiten signifikant zu senken.

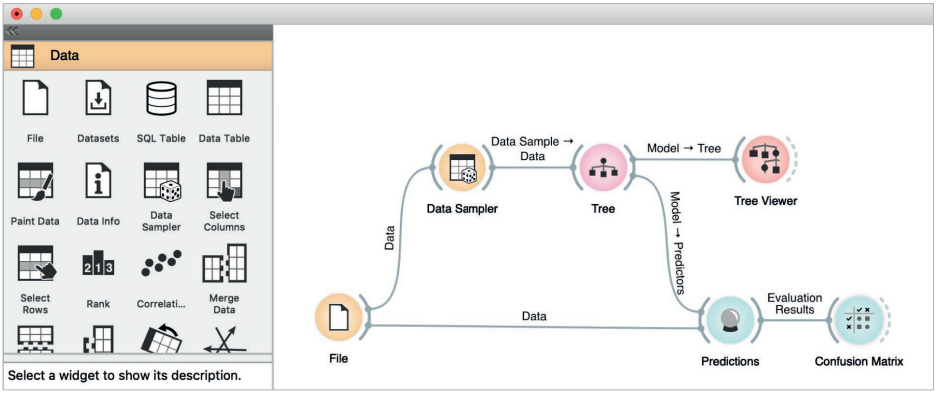


Abbildung 14: Modellierung der Datenanalyse im verwendeten Analysewerkzeug Orange.

Der Entscheidungsbaum (Abbildung 15) erlaubt einen Einblick in die Entstehung der getroffenen Vorhersagen. Die relativ hohe Analysequalität (Abbildung 16) und die direkte Betroffenheit der Schülerinnen und Schüler eröffnet ein hohes Potential zur kritischen Hinterfragung von Datenanalysen allgemein und insbesondere der Analysequalität.

Für die vierte Doppelstunde ist geplant, die bisher erworbenen Kompetenzen auf weitere Kontexte zu übertragen und somit beispielsweise die Datennutzung in der Medizin, durch Versicherungen und Banken zu hinterfragen und im Rahmen eines Gruppenpuzzles rechtliche, ethische und moralische Aspekte dieser Analysen zu diskutieren.

Werkzeugauswahl: Das Data Mining Tool *Orange*

Um die Mächtigkeit und das Potential automatisierter Datenanalysen im Unterricht für die Schülerinnen und Schüler erlebbar zu machen, wurde ein Werkzeug für den Unterricht gesucht, das eine intuitive Nutzung ohne detailliertes Vorwissen (weder im Bereich Datenanalyse, noch in der Programmierung) erlaubt. Entsprechend war z. B. die Verwendung einer klassischen Programmiersprache wie Python nicht möglich. Besonders einfach verwendbar sind grafisch orientierte Analysewerkzeuge, bei denen die Anwender die Analyse als Datenflussmodell beschreiben, zum Beispiel *rapidminer*[13] und *Orange*[14], die beide alle für den Schulunterricht notwendigen Funktionalitäten bereitstellen. Zur Erprobung des Unterrichtskonzepts wurde aus zwei Gründen das Werkzeug *Orange* gewählt und erfolgreich eingesetzt: Einerseits war es hier möglich, die Komplexität weiter zu reduzieren, da es sich um ein Open-Source-Programm handelt, in dem nicht benötigte Module somit ausgeblendet werden konnten[15]. *rapidminer* hingegen wäre nicht entsprechend anpassbar gewesen und durch das kommerzielle Geschäftsmodell weniger einfach nutzbar.

[13] https://rapidminer.com/educational-program

[14] https://orange.biolab.si

[15] Die Ausblendung der Module hat sich als nicht notwendig erwiesen, da das Programminterface von den Schülerinnen und Schülern wesentlich intuitiver erfassbar war, als zunächst angenommen.

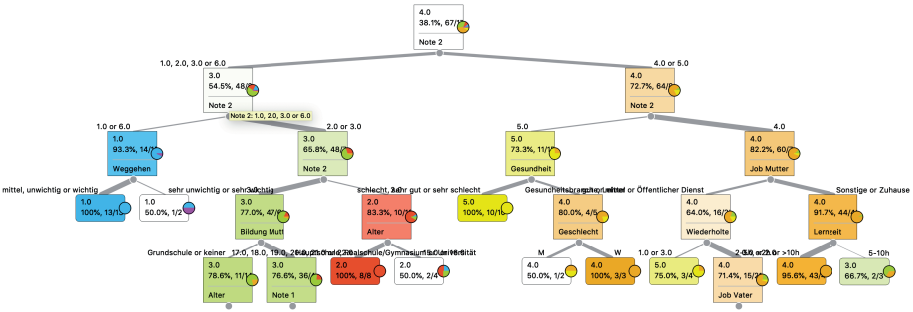


Abbildung 15: Darstellung eines Klassifikationsbaums in Orange.

predicted / actual	1.0	2.0	3.0	4.0	5.0	6.0
1.0	34	10	2	0	0	0
2.0	1	52	31	1	0	0
3.0	0	1	182	34	0	0
4.0	0	0	5	212	19	0
5.0	0	0	0	9	40	0
6.0	0	0	0	9	7	0

Abbildung 16: Konfusionsmatrix zur Beurteilung der Analysequalität in Orange.

ZUSAMMENFASSUNG

Wie in dieser Broschüre gezeigt wurde, hat das Themenfeld *Daten und Datenmanagement* heute viele Facetten, die über die bislang übliche Unterrichtsthema *Datenbanken, Datenmodellierung* und *Datenschutz* hinausgehen. Eine Thematisierung dieser Themen im Informatikunterricht ist aber wichtig, um Schülerinnen und Schülern das Rüstzeug für ein selbstbestimmtes Leben in der digitalen Welt mitzugeben, ihnen zu ermöglichen, fundierte Entscheidungen bezogen auf den Umgang mit ihren persönlichen Daten, aber den Daten anderer, zu treffen und schlussendlich auch – beispielsweise durch eigene Datenanalysen – von den heute vorhandenen Möglichkeiten zu profitieren.

Die im Rahmen dieser Broschüre kurz präsentierten Unterrichtsideen zeigen, dass dies möglich ist: Im Rahmen von Erprobungen im Informatikunterricht, aber auch in Lehrkräftefortbildungen in mehreren Bundesländern, erhielt ich größtenteils positive Resonanz und konnten feststellen, dass diese Themen nicht nur im Unterricht umsetzbar sind, sondern ein sowohl aus Sicht der Lehrkräfte, aber auch aus Perspektive der Schülerinnen und Schüler, ein interessantes und wichtiges Thema für den Informatikunterricht darstellen.

Um die in dieser Broschüre angesprochenen Themen im Unterricht voranzubringen, wird sich der Autor auch weiterhin umfassend mit diesem Thema befassen. In diesem Rahmen sind auch weitere Lehrkräftefortbildungen geplant, die dazu beitragen sollen, dass Lehrkräfte ihre Ideen zur Umsetzung sowie ihr Fachwissen in diesem Bereich weiter ausbauen können und somit die Möglichkeit bekommen, diese Themen umfassend zu diskutieren. Neben kleineren Workshops für Lehrkräfte, beispielsweise an den in den meisten Bundesländern stattfindenden Tagungen der Landeslehrergruppen der Gesellschaft für Informatik, ist angedacht, Ende 2019 eine größer angelegte Lehrkräftefortbildung anzubieten. In deren Rahmen soll nicht nur ein knapper Einblick in das Thema angeboten werden, sondern intensiv mit den beteiligten Lehrkräften am Thema gearbeitet werden, mit dem Ziel, die Themen in die Breite zu tragen. Prinzipiell interessierte Lehrkräfte, die das Potential sehen, diese Themen in ihren Unterricht einzubauen bzw. auch im eigenen Lehrkräfteumfeld weiterzutragen, sind gerne eingeladen, beim Autor dieser Broschüre ihr Interesse zu bekunden und sich in die u. g. Mailingliste einzutragen. Die geplante größer angelegte Lehrkräftefortbildung wird mit Mitteln der *Google PD Educator Grants* gefördert, sodass die Kosten der Teilnehmenden minimiert werden können.

Weitere Informationen zum Projekt:

Web: https://dataliteracy.education
Lehrkräfte-Mailingliste: https://lists.fu-berlin.de/listinfo/lehrkraefte-data-literacy

LITERATUR

Ali-ud-din Khan, M., Fahim Uddin, M., & Gupta, N. (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.

Brewer, E. (2012). CAP Twelve Years Later: How the "Rules" Have Changed. Computer, 45(2).

Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., … Gruber, R. E. (2008). Bigtable: A Distributed Storage System for Structured Data. ACM Trans. Comput. Syst., 26(2).

Codd, E. F. (1970). A Relational Model of Data for Large Shared Data Banks. Commun. ACM, 13(6).

Cortez, P., & Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. In A. C. Brito & J. M. F. Teixeira (Eds.), Proceedings of 5th Annual Future Business Technology Conference, Porto, 2008.

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified Data Processing on Large Clusters. Comm. ACM, 51(1).

Dorschel, J. (2015). Praxishandbuch Big Data. Springer Gabler.

Edlich, S., Friedland, A., Hampe, J., Brauer, B., & Brückner, M. (2011). NoSQL. Hanser Verlag.

Exner, M. (2009). Die Entdeckung der Cholera-Ätiologie durch Robert Koch 1883/84. In: *HygMed* 34 (4).

Fischer, S. (2014). Big Data: Herausforderungen und Potenziale für deutsche Softwareunternehmen. Informatik-Spektrum, 37(2).

AK Bildungsstandards. (2008). Grundsätze und Standards für die Informatik in der Schule: Bildungsstandards Informatik für die Sekundarstufe I. LOG IN, 28(150/151).

AK Bildungsstandards SII. (2016). Bildungsstandards Informatik für die Sekundarstufe II. LOG IN, 36(183/184).

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature, 457(7232).

Golab, L., & Özsu, M. T. (2003). Issues in Data Stream Management. SIGMOD Rec., 32(2).

Hill, K. (2012). How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did.

Kantardzic, M. (2011). Data Mining (2nd ed.). John Wiley & Sons.

Kemper, A., & Eickler, A. (2015). Datenbanksysteme: Eine Einführung. De Gruyter Oldenbourg.

Kolany-Reiser, B. & Radtke, T. (2018). Ich sammele, also bin ich (Social Credit). Das Szenario eines allumfassenden Bonitätssystems am Beispiel Chinas. ABIDA-Dossier.

Klein, D., Tran-Gia, P., & Hartmann, M. (2013). Big Data. Informatik-Spektrum, 36(3).

Koch, R. (1893). Wasserfiltration und Cholera. Zeitschrift Für Hygiene Und Infektionskrankheiten, 14(1).

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Maury, M. F. (1855). The Physical Geography of the Sea. Harper & Brothers.

Mayer-Schönberger, V., & Cukier, K. N. (2013). Big Data - Die Revolution, die unser Leben verändern wird. FinanzBuch Verlag.

McBurney, V. (2013). The Origin and Growth of Big Data Buzz. toolbox.Com.

Reinsel, D., Gantz, J., & Rydning, J. (2017). Data Age 2025: The Evolution of Data to Life-Critical. International Data Corporation (IDC).

Riley, J. (2017). Understanding Metadata: What is Metadata, and What is it For?: A Primer. NISO.

Schwill, A. (1993). Fundamentale Ideen der Informatik. Zentralblatt Für Didaktik Der Mathematik, 25(1).

Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27(3).

Strava. (2017). Strava Global Heatmap.

Strozzi, C. (n.d.). NoSQL: a non-SQL RDBMS.

Truong, K. (2013). INFOGRAPHIC: Users weighed down by multiple gadgets - survey reveals the most carried devices.

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). Amtsblatt Der Europäischen Union, 56(L 119).

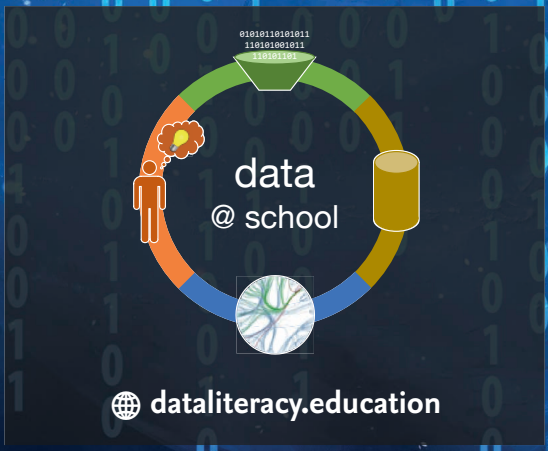
Valdivia, A., Lopez-Alcalde, J., Vicente, M., Pichiule, M., Ruiz, M., Ordobas, M., & others. (2010). Monitoring influenza activity in Europe with Google Flu Trends: comparison with the findings of sentinel physician networks-results for 2009-10. Euro Surveill, 15(29).

The Verge. (2018). Strava's fitness tracker heat map reveals the location of military bases.

Zeller, T. J. (n.d.). Beware Your Trail of Digital Fingerprints.

In dieser Broschüre wird, basierend auf den Ergebnissen des 2013 an der Professur für Didaktik der Informatik der Friedrich-Alexander-Universität Erlangen-Nürnberg begonnenen und seit 2018 an der Freien Universität Berlin weitergeführten Projekts “Datenmanagement im Informatikunterricht”, ein Überblick über die für Informatiklehrkräfte zentralen Aspekte dieses Fachgebiets gegeben.

Dazu wird nicht nur das Fachgebiet Datenmanagement an sich charakterisiert und ein Einblick in verschiedene diesem zugehörige Themen gegeben, sondern auch Alltagsbeispiele betrachtet, die als Kontext und Orientierung für den Informatikunterricht hilfreich sind.

Um eine fachliche Fundierung des Informatikunterrichts zum betrachteten Themenspektrum zu ermöglichen, werden außerdem die im Rahmen einer empirischen Forschungsarbeit ermittelten Schlüsselkonzepte des Datenmanagements sowie darauf aufbauend ein für den allgemeinbildenden konzipiertes und an die Empfehlungen für Bildungsstandards in Informatik der Gesellschaft für Informatik angelehntes Datenkompetenzmodell beschrieben.

Zuletzt wird anhand verschiedener Ideen für den Unterricht eine Grundlage für die Konzeption eigener Unterrichtseinheiten vorgestellt.