Jens Gallenbacher (Hrsg.): INFOS 2015 – Informatik allgemeinbildend begreifen,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2015 125

Big Data im Informatikunterricht: Motivation und Umsetzung

Andreas Grillenberger und Ralf Romeike[1]

Abstract: Das Sammeln und Auswerten von Daten ist heute allgegenwärtig: In vielen Bereichen des täglichen Lebens nimmt die Bedeutung von Daten und datenbezogenen Anwendungen immer mehr zu, z. B. bei der Nutzung sozialer Medien oder bei der Verwaltung großer Mengen an eigenen Daten. Während Daten früher hauptsächlich konsumiert wurden, wird heute auch jeder zum Produzenten immer umfangreicherer Datenmengen. Dabei werden immer größere und vielfältigere Datenmengen verwaltet und verarbeitet. Im Informatikunterricht wird Big Data jedoch bisher kaum thematisiert: Die fachlichen Grundlagen dafür scheinen auf den ersten Blick zu komplex und kaum auf schulischem Niveau verständlich zu sein. In diesem Beitrag werden daher zuerst die wesentlichen Entwicklungen vorgestellt, die sich derzeit im Datenmanagement ereignen, sowie die sich dadurch ergebenden fachlichen Herausforderungen. Um zu demonstrieren, dass solche fachliche Innovationen oft grundlegende Konzepte enthalten, die im Informatikunterricht thematisiert werden können, wird anhand von zwei Unterrichtsszenarien aus diesem Themenbereich exemplarisch vorgestellt, wie Informatik es mittels moderner Ansätze zur Datenverarbeitung ermöglicht, Big Data beherrschbar zu machen.

Keywords: Daten, Datenmanagement, Datenbanken, Kompetenzen, Big Data, NoSQL, Datenanalyse, Datenstromsysteme, Datenverarbeitung, Datenvisualisierung, Alltag, Datenschutz

1 Einleitung

Big Data gehört derzeit wohl zu den maßgeblichen Erscheinungen der Informatik, die alles und jeden betreffen. Nicht nur in Informatik, Wirtschaft und Gesellschaft spielt dieses Thema, die Verarbeitung von großen und vielfältig strukturierten Datenmengen in hoher Geschwindigkeit, in den letzten Jahren eine immer größere Rolle. Auch die Möglichkeiten und Konsequenzen werden heute in der Gesellschaft ausführlich diskutiert: So bestimmen die negativen Seiten dieser Entwicklungen, z. B. die Möglichkeiten der Geheimdienste zur Sammlung und Analyse umfangreicher Datenmengen, seit geraumer Zeit die Nachrichten. Andererseits eröffnen öffentlich verfügbare Datenquellen willkommene Möglichkeiten, die bis vor kurzem noch undenkbar waren. Beispielsweise analysiert *Google Flu Trends* Suchanfragen und prognostiziert dadurch Grippewellen oft genauer und auch wesentlich schneller als herkömmliche Vorhersagen. Analog nutzen Herausgeber von Kreditkarten zunehmend Datenanalysen um Bezahlvorgänge in Echtzeit auf mögliche Betrugsfälle zu prüfen und betroffene Transaktionen abzulehnen [MC13].

Während in einigen Fällen die Datenerfassung und -auswertung für den Nutzer offensichtlich ist, beispielsweise bei intelligenten Stromzählern oder Smart Watches, die ihre Auswertungen dem Nutzer zur Verfügung stellen, werden jedoch auch große Datenmengen im Verborgenen generiert. Smartphones erfassen kontinuierlich die Bewegungsdaten

[1] Friedrich-Alexander-Universität Erlangen-Nürnberg, Didaktik der Informatik, Martensstr. 3, 91058 Erlangen
andreas.grillenberger@fau.de, ralf.romeike@fau.de

ihrer Nutzer und teilen diese sogar dem Hersteller mit, Webseiten (und zunehmend auch Desktop-Anwendungen) protokollieren genaue Nutzeraktionen, selbst Taxifahrten werden heute detailliert erfasst und ausgewertet [Bu14]. Diese Datensammlungen bieten oft einen direkten Mehrwert für den Nutzer: Das Smartphone kann bei Verlust oder Diebstahl geortet werden, die Nutzererfahrung von Anwendungen wird verbessert und Taxis können schon vorab zu stark frequentierten Orten gesendet werden, um die Wartezeiten zu verkürzen. Die Bereitschaft, an datenbasierten Anwendungen zu partizipieren, ergibt sich daher aus Faktoren wie dem Vertrauen in den Anbieter, dem persönlich erfahrenen Nutzen oder dem Umfang der Datenerfassung [Wo14]. Um diese Aspekte, insbesondere das Vertrauen in einen Anbieter und die Möglichkeiten, die sich durch die Datenerfassung ergeben, einschätzen zu können, ist jedoch ein Verständnis der Grundlagen, Möglichkeiten und Risiken moderner Datenverarbeitung nötig.

Trotz sichtbarer Bemühungen, Themen wie Datenschutz im Unterricht zu berücksichtigen, konzentriert sich Informatikunterricht zum Thema *Daten* bisher vor allem auf den Kontext *Datenbanken* [GR14]. Themen die heute allgegenwärtig sind, z. B. Echtzeit-Datenanalysen, Datenqualität, Metadaten oder die Datennutzung im Alltag, werden bisher kaum betrachtet, wie eine Analyse verschiedener Lehrpläne und Bildungsstandards zeigt [GR14]. Auch in der Fachdidaktik werden diese Entwicklungen bisher kaum thematisiert. Obwohl diese Themen auf komplexen Grundlagen wie Statistik beruhen, zeigen sich die Auswirkungen und Vorgehensweisen aus dem Bereich Big Data jedoch in einer Vielzahl von Gebieten und werden auch außerhalb der Informatik, für hochkomplexe aber auch relativ einfache Zwecke, eingesetzt. Selbst Datenwissenschaftler wie Halevy et. al. [HNP09] schließen, dass einfache mathematische Modelle in Kombination mit großen Datenmengen oft bessere Ergebnisse liefern als komplexe Modelle mit wenigen Daten. Auch im Kontext von Big Data werden viele der bereits seit Jahren eingesetzten Methoden weiterhin verwendet, beispielsweise die Datenanalysemethoden „Klassifikation“, „Clusteranalyse“ und „Assoziation“. Insbesondere die Assoziationsanalyse gewinnt heute stark an Bedeutung. Aufgrund der Verbreitung dieses Themas, dem Einsatz auf verschiedenen Komplexitätsstufen und der zeitlichen Beständigkeit vieler der Grundlagen, kann angenommen werden, dass die weitgreifenden Innovationen im Datenmanagement informatische Grundlagen im Sinne der fundamentalen Ideen [Sc93] beinhalten, die den Informatikunterricht auch über das Thema „Datenbanken“ hinaus wesentlich bereichern können. Damit zeigt sich unmittelbar, dass es eine wichtige Herausforderung für den Informatikunterricht ist, schülergerechte Zugänge zu diesen Themen zu finden. Im Folgenden skizzieren wir daher die grundlegenden Entwicklungen im Datenmanagement und zeigen anhand von zwei Beispielen exemplarisch, wie solche Innovationen aus der Informatik für den Informatikunterricht didaktisch reduziert und somit zugänglich gemacht werden können.

2 Big Data: Grundlagen und aktuelle Entwicklungen

Big Data ermöglicht vielfältige Innovationen im Bereich Datenverwaltung und -verarbeitung, die Voraussetzung für Informatiksysteme sind, wie wir sie heute täglich benutzen: Soziale Netzwerke nutzen *nicht-relationale Datenbanken*, z. B. zur Speicherung von Freundschaftsbeziehungen oder zur Verbesserung der Suchmöglichkeiten [ER13], Daten werden zu verschiedensten Zwecken in *Echtzeit analysiert*, immer häufiger in der *Cloud* gespeichert

und dabei auf eine große Anzahl von *Systemen verteilt*. Diese technischen Innovationen führen zu neuartigen Auswertungsmöglichkeiten von Daten, die immer öfter eine Grundlage für den wirtschaftlichen Erfolg von Unternehmen darstellen: Beispielsweise prognostiziert Amazon mittlerweile gut genug welche Artikel ein Kunde möglicherweise kaufen wird, um diese schon vorher in ein naheliegendes Versandzentrum zu verlegen [He14]. Der Umgang mit Daten hat sich dabei in den letzten Jahren stark gewandelt: Während die Verarbeitung und Speicherung von Daten bisher insbesondere auf der Nutzung von relationalen Datenbanken basierte, ist heute eine Ausweitung dieses Fachgebiets erkennbar. Neben Datenbanken bzw. der Verwaltung strukturierter Daten gewinnen im nun oft als *„Datenmanagement"* bezeichneten Fachgebiet auch weitere Aspekte an Bedeutung, z. B. die Verwaltung un- bzw. wenig strukturierter Daten, Datensicherheit und Datenschutz, gemeinsame Datennutzung, Metadaten und auch Datenqualität (vgl. [DA09]).

Eine grundlegende Innovation stellt dabei die Möglichkeit zur Verarbeitung von *Big Data* dar. Während dieser Begriff auf den ersten Blick insbesondere auf große Datenmengen (*volume*) hindeutet, wird er nach Laney [La01] noch durch zwei weitere „V"s charakterisiert (vgl. Abb. 1): Die unterschiedliche Strukturierung von Daten (*variety*) sowie deren immer schnellere Erzeugung und Verarbeitung (*velocity*). Diese Entwicklung stellt die Verarbeitung und Speicherung von Daten in Datenbanken vor neue Herausforderungen: Neben traditionellen Aspekten, wie Konsistenz, gewinnt heute die verteilte Datenspeicherung stark an Bedeutung. Steigende Datenmengen und Verarbeitungsgeschwindigkeit verhindern die Verarbeitung auf einem einzelnen Datenbankserver (selbst wenn dieser hoch performant ist), eine *vertikale Skalierung* reicht daher für den Umgang mit Big Data nicht aus. Stattdessen muss *horizontale Skalierung* eingesetzt werden: Die Daten werden auf viele (im Einzelnen weniger performante) Server verteilt und parallel verarbeitet [FU14].

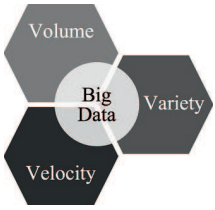

Abb. 1: Drei-V-Modell für Big Data

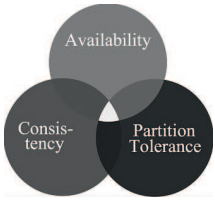

Abb. 2: CAP-Theorem

Bei der verteilten Speicherung von Daten auf mehreren Datenbankservern muss jedoch, wann immer eine Anfrage einen konsistenten Zustand benötigt, dieser erst durch Synchronisation und Prüfung verschiedener Bedingungen (z. B. constraints und das definierte Datenschema) sichergestellt werden. Konsistenz sorgt daher bei der parallelen Datenspeicherung für eine wesentliche Verlangsamung. Dieser inhärente Widerspruch wird durch das *CAP-Theorem* [Br12] beschrieben: Nur zwei der Eigenschaften Konsistenz (***C****onsistency*), hohe und schnelle Verfügbarkeit (***A****vailability*) sowie Partitionstoleranz (***P****artition tolerance*, Möglichkeit zur verteilten Speicherung von Daten) können gleichzeitig sichergestellt werden (vgl. Abb. 2). Bei vielen modernen Datenbankanwendungen steht insbesondere die Verfügbarkeit und verteilte Speicherung im Vordergrund, sodass die bisher meistgenutzten relationalen Datenbanken nicht mehr ausreichen: Diese stellen insbesondere Konsistenz und Verfügbarkeit, nicht aber Partitionstoleranz, sicher. Daher entstehen derzeit vielfältige neue Datenbankmodelle wie Graphdatenbanken (z. B. *Neo4J*) oder dokumentenorientierte Datenbanken (z. B. *CouchDB*), die unter dem Begriff *NoSQL*[2] zusammengefasst

[2] *NoSQL* wird heute als „Not only SQL" interpretiert [Ed11] und als Sammelbegriff für nicht-relationale Datenbanksysteme verwendet.

werden. Diese Datenbanken heben üblicherweise eine wesentliche Einschränkung des relationalen Modells auf: Es wird seitens der Datenbank kein Schema erzwungen, sodass die Konsistenz der Daten nicht durch die Datenbank sichergestellt werden kann, sondern in den Verantwortungsbereich der Anwendung verlagert wird. Dadurch wird jedoch eine wesentlich schnellere Datenverarbeitung ermöglicht. Im Gegensatz zum *ACID*-Paradigma[3] bei relationalen Datenbanken, liegt den NoSQL-Datenbanken daher das *BASE*-Paradigma zugrunde: **B**asically **A**vailable, **S**oft-state, **E**ventually consistent [Ed11].

3 Big Data als Unterrichtsgegenstand

Die Betrachtung der fachlichen Innovationen im Datenmanagement zeigt sich eine Vielzahl neuartiger Konzepte und Vorgehensweisen, die grundlegend für das Verständnis von Phänomenen wie Cloudcomputing und Big Data sind. Es kann vermutet werden, dass diese Konzepte prototypisch für fundamentale Ideen stehen, die zum Erschließen der im Kontext von Big Data entstehenden Phänomene wichtig sind. Im Folgenden stellen wir daher anhand zweier Unterrichtsszenarien die Möglichkeit dar, grundlegende Aspekte von Big Data auf schulischem Niveau zu vermitteln.

3.1 Schürfen nach Daten: Data Mining in Big Data

Um Entscheidungen darüber treffen zu können, ob man Informatiksysteme nutzt, die Daten in großem Umfang sammeln und analysieren, ist es nötig, das Potential und die Gefahren der Big-Data-Verarbeitung einschätzen zu können. Eine der größten Herausforderungen bei der Verarbeitung von Big Data stellt heute die Analyse der Daten dar, wobei sich die Herangehensweise oft deutlich von klassischen Datenanalysen unterscheidet: Statt Daten für eine spezielle Analyse gezielt zu sammeln, wird heute versucht, aus auf Vorrat gesammelten Daten wertvolle Informationen zu gewinnen [DN13]. Zum Zeitpunkt der Datensammlung ist dabei typischerweise weder die Existenz der zu gewinnenden Informationen, noch der Analysezweck bekannt. Diese Analysen werden daher oft, in Analogie zum Goldbergbau, als *Data Mining* bezeichnet: Das Schürfen nach wertvollen Informationen im großen Datenberg[4]. In diesem Zusammenhang werden Daten auch als das neue Öl des 21. Jahrhunderts beschrieben [Wo11]. Die zur Datenanalyse eingesetzten Methoden setzen dabei andere Schwerpunkte als bisher: Durch eine andere Sichtweise auf Daten und die primäre Suche nach Korrelationen statt Kausalitäten können vielfältige Ergebnisse gewonnen werden, die bisher kaum möglich waren.

Unterrichtsgegenstand Data-Mining In diesem Beispiel stellen wir im Kontext der Gewinnung von Informationen aus verschiedenen *Open-Data*[5]-Datensätzen die Thematisierung einiger grundlegender Data-Mining-Methoden im Informatikunterricht beispielhaft dar. Obwohl Datenbanken dabei als Werkzeug eingesetzt werden und daher auch Kompetenzen

[3] ACID: **A**tomicity, **C**onsistency, **I**solation, **D**urability [KE13]

[4] Der Begriff *Data Mining* ist in Bezug auf die Analogie jedoch missverständlich, da diese eher den Begriff *Information Mining* nahelegt, da es um den Abbau von Informationen geht.

[5] Unter *Open Data* werden Datensätze verstanden, die frei zugänglich zur Verfügung gestellt werden.

in diesem Bereich erworben werden können, wird der Schwerpunkt hier auf die Grundlagen und Möglichkeiten von Datenanalysen gelegt und Schrittweise an diese herangeführt[6]. Als Datenquelle für dieses Beispiel wählen wir zwei Open-Data-Sätze, die alle Vorfälle, die der Polizei von San Francisco im Jahr 2014 gemeldet wurden[7], sowie alle Anrufe, die dort bei der Servicenummer „311“ seit 2008 eingingen[8], beinhalten. Moderne Datenanalyse-Tools (z. B. „Tableau Public“, „Microsoft PowerQuery“ oder „CartoDB“) ermöglichen einen direkten Zugriff auf solche Datensätze. Eine wichtige Grundlage aller Data-Mining-Prozesse sind die Datenanalysemethoden *Klassifikation*, *Assoziation* und *Clusterbildung*. Während die Klassifikation von Daten auch im traditionellen Datenbankunterricht in Form von Gruppierungen vorkommt, stellen die Clusterbildung (Zusammenfassen von Datensätzen, die bezüglich der zu analysierenden Eigenschaft gleich oder ähnlich sind) und das Prinzip der Assoziation (d. h. von einem Attributwert oder einer Menge von Attributwerten auf weitere Eigenschaften zu schließen, die nicht explizit im Datensatz genannt sind) neue Vorgehensweisen dar.

Für viele Analyseziele reicht es dabei heute aus, Korrelationen zwischen Daten zu erkennen anstatt nach Kausalzusammenhängen zu suchen: *Google Flu Trends* sucht beispielsweise nach Suchbegriffen, die als Indikator für eine Grippeerkrankung dienen können. Für die Prognose der Ausbreitung von Grippewellen ist es nebensächlich, ob ein Sinn-Zusammenhang zwischen einem Suchbegriff und einer Grippeerkrankung erkennbar ist, eine hohe Korrelation tatsächlicher Grippeerkrankungen und der Suche nach einem Begriff aus. Dies wird möglich, da heute (idealerweise) alle verfügbaren Informationen zum Gegenstand der Untersuchung gespeichert und somit analysiert werden können, sodass wesentlich geringere Stichprobenfehler vorliegen als bei klassischen Analysen, die sich auf kleine Datenmengen oder beschränkte Stichproben konzentrieren. Durch die meist ungeprüfte Speicherung der verfügbaren Daten besteht jedoch die Gefahr, dass durch ungenaue oder fehlerbehaftete Daten auch fehlerhafte Ergebnisse entstehen. Bei ausreichend großen Datenmengen kann jedoch davon ausgegangen werden, dass die Fehler nur einen geringen Anteil ausmachen und somit das Ergebnis, wenn überhaupt, nur geringfügig negativ beeinflussen.

Ziel des Unterrichts ist es, dass die Schülerinnen und Schüler …

- einfache Datenanalysen an vorgegebenen Datensätzen durchführen
- die Datenanalysemethoden Klassifikation und Assoziation am Beispiel nachvollziehen und erklären
- die Möglichkeiten und Gefahren von Big-Data-Analysen erkennen
- verstehen, dass die Qualität der gewonnenen Information nicht nur von der Analyse der Daten sondern insbesondere auch von deren Interpretation abhängt
- erkennen, dass der Einfluss der Datenqualität mit ansteigender Datenmenge abnimmt
- den Unterschied zwischen Kausalität und Korrelation erkennen und am Beispiel erklären

[6] Je nach organisatorischen Gegebenheiten (z. B. Lehrplan) kann der Schwerpunkt jedoch auch stärker in Richtung relationaler Datenbanken verschoben werden.

[7] https://data.sfgov.org/Public-Safety/SFPD-Incidents-from-1-January-2014/tmnf-yvry (abgerufen: 20.04.2015)

[8] https://data.sfgov.org/City-Infrastructure/Case-Data-from-San-Francisco-311-SF311-/vw6y-z8j6 (abgerufen: 20.04.2015)

Ein möglicher **Unterrichtsablauf** gliedert sich in folgende Phasen:

1. **Kennenlernen von Big Data:** Durch Untersuchung eines großen Datensatzes erkennen die Lernenden den Aufbau von Datensätzen sowie die Bedeutung der verschiedenen Attribute und können Ideen zu den darin implizit enthaltenen Informationen sammeln.
2. **Klassifizierung von Daten:** Beim Filtern des Kriminalfälle-Datensatzes nach Stadtteilen und Bestimmung der Anzahl gemeldeter Ereignisse in diesem lernen die Schülerinnen und Schüler die Klassifizierung von Daten nach gegebenen Merkmalen kennen. Nebenbei kann die Aussagekraft von Analyseergebnissen am Beispiel diskutiert werden: Eine mögliche Fehlinterpretation der gewonnenen Information wäre, dass ein Stadtteil aufgrund einer höheren Zahl von Kriminalfällen in diesem gefährlicher ist.
3. **Assoziationen zwischen Daten:** Es kann festgestellt werden, dass die Adresse auch den Stadtteil festlegt. Während diese Assoziation offensichtlich ist, sind es andere weniger: Auch die Beschreibung eines Vorfalls (anscheinend ein vordefinierter Text) legt die Kategorie fest. Im Unterricht kann anhand dieser beiden Beispiele das Weglassen der redundanten Attribute Stadtteil bzw. Kategorie diskutiert werden. Dabei ist es wichtig, den Lernenden bewusst zu machen, dass anhand des Datensatzes nur eine Korrelation zwischen den beiden Attributen gefolgert werden kann, aber keine Kausalität. Durch das Weglassen können daher Fehler entstehen. Weitere Informationen, wie zum Beispiel beim Zusammenhang zwischen Adresse und Stadtteil vorhanden, können jedoch einen Kausalzusammenhang untermauern, sodass derartige Fehler ausgeschlossen werden.
4. **Verknüpfung von Datensätzen:** Aus dem Datensatz können verschiedene Informationen auf einfache Weise gewonnen werden, beispielsweise die Abhängigkeit der Anzahl an Delikten vom Stadtteil: Die Daten wurden in geeigneter Weise gesammelt, um solche Auswertungen durchführen zu können. Falls jedoch weitergehende Informationen gewonnen werden sollen, müssen weitere Daten herangezogen werden. Mit dem Datensatz ist es beispielsweise nicht direkt möglich zu analysieren, ob eine Korrelation zwischen Verkehrsunfällen und Ausfällen der Straßenbeleuchtung vorliegt. Es existiert jedoch auch kein Datensatz, in dem diese Information direkt enthalten ist. Durch Annahme einer Assoziation, nämlich dass im Fall einer ausgefallenen Beleuchtung eine Meldung bei der zuständigen Behörde eingeht, kann diese Information jedoch gewonnen werden: Die Anrufe bei der Servicenummer 311 der Stadt San Francisco liegen als Datensatz vor und können unter der Annahme dieser Assoziation in einen neuen Datensatz, der die benötigten Informationen beinhaltet, überführt werden. Indem diese Informationen und der ursprüngliche Datensatz im Rahmen eines Mash-Up[9] zusammengefasst werden, kann untersucht werden, ob die gesuchte Korrelation vorliegt.
5. **Rückblick:** Im Rückblick kann erkannt werden, dass selbst wenige zusätzliche Daten dazu beitragen können, wesentlich umfangreichere Informationen zu gewinnen. Dabei zeigt sich, dass Fehler in den Daten, wie sie beispielsweise auch durch die ungenaue Assoziation „*keine Beschwerde → Beleuchtung funktionsfähig*“ produziert werden, sich zwar im Einzelfall negativ auswirken, bei großen Datenmengen jedoch nur noch einen geringen Einfluss haben.

[9] Unter einem *Mash-Up* wird allgemein die Verknüpfung verschiedener Datensätze verstanden. In relationalen Datenbanken kann dies mit einen Join verglichen werden.

Ein wichtiges Merkmal von Data Mining, das durch dieses Beispiel verdeutlicht wird, ist der geänderte Umgang mit Daten: Statt für bestimmte Zwecke gesammelte Daten auszuwerten, können heute auch große Datenmengen auf unbekannte oder bei der Sammlung der Daten noch nicht angestrebte bzw. vermutete Zusammenhänge untersucht und dadurch wertvolle Informationen gewonnen werden.

3.2 Fischen im Unendlichen: Datenstromsysteme

Eine grundlegende Herausforderung stellt heute die Verarbeitung von Big Data (nahezu) in Echtzeit dar: Viele Anwendungen, wie die Live-Erkennung von Kreditkartenbetrug oder die Auswertung von Sensordaten (u. a. bei der Tsunami-Erkennung), benötigen eine schnelle Datenauswertung, müssen aber zugleich umfangreiche Datenmengen verarbeiten.

Unterrichtsgegenstand Datenströme Obwohl es heute möglich ist, immer größere Datenmengen zu speichern und zu analysieren, ist dieses Vorgehen nicht in allen Fällen zielführend: Bei der Betrachtung von kontinuierlichen Datenströmen, die u. a. durch soziale Medien oder Sensoren erzeugt werden, enthält in der Regel nur ein Teil der verfügbaren Daten relevante Informationen (beispielsweise, dass ein definierter Grenzwert überschritten wurde). Gleichzeitig wird die schnelle Verfügbarkeit der Analyseergebnisse, idealerweise in Echtzeit, immer grundlegender (z. B. im automatisierten Börsenhandel). Der bisherige Ansatz, alle erzeugten Daten in einer Datenbank zu speichern, ist daher in diesen Fällen weniger geeignet. Stattdessen reicht es aus, nur die für den konkreten Zweck benötigten Daten auszuwählen und direkt zu verarbeiten, sodass eine weitere Speicherung nicht nötig ist. Im Vergleich zu einer Datenbank, bei der die Daten permanent gespeichert und verschiedene Abfragen auf diese angewendet werden (vgl. Abb. 3), verarbeiten daher sog. *Datenstromsysteme* den Datenstrom direkt, indem vorher definierte Abfragen auf den Datenstrom angewendet werden, ohne dass eine dauerhafte Datenspeicherung nötig ist (vgl. Abb. 4). Anschaulich kann das Prinzip einer Datenbank mit einem Hamster verglichen werden, der sein Futter auf Vorrat sammelt, während ein Datenstromsystem dagegen einem Bären ähnelt, der nur Fische aus einem Fluss fischt um diese direkt zu fressen. In einem Datenbanksystem können daher jederzeit weitere Anfragen auf bereits analysierte Daten angewendet werden, während in einem Datenstromsystem die Daten nach der Verarbeitung verloren sind[10]. In Zusammenhang mit Big Data gewinnen diese Systeme immer stärker an Bedeutung: Während die schnelle Verarbeitung sehr großer Datenmengen mit bewährten Systemen kaum möglich ist[11], stellen Datenstromsysteme eine

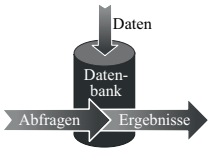

Abb. 3: Funktionsweise eines DBMS

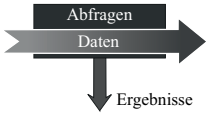

Abb. 4: Funktionsweise eines DSS

[10] Häufig wird dem entgegengewirkt, indem alle oder ein Ausschnitt der Daten neben der Echtzeitverarbeitung im Datenstromsystem in einem anderen System (z. B. Datenbank) für längere Zeit gespeichert werden.

[11] Um den Twitter-Nachrichtenstroms zu analysieren müssten ca. 6000 Tweets pro Sekunde abgespeichert und verarbeitet werden [Kr13]. Nimmt man als Größe eines Tweets inklusive aller Metadaten, wie GPS-Position oder Hashtags, durchschnittlich 500 Byte an, wächst die Datenmenge sekündlich um 3 MB an, täglich daher um über 250 GB. Die Speicherplatzgrenzen typischer Datenbanksysteme wären damit nach wenigen Tagen erreicht.

effiziente Lösung dar und ermöglichen durch direkte Verarbeitung schnelle Analyseergebnisse, aber auch eine hohe Speicherplatzeffizienz. Im Kontext von Big Data können daher die Grenzen und Funktionsweise moderner und im Alltag oft präsenter Echtzeitsysteme verdeutlicht werden.

Ziel des Unterrichts ist es, dass die Schülerinnen und Schüler ...

- erkennen, dass selbst große Datenmengen mit geeigneten Informatiksystemen in nahezu Echtzeit verarbeitet werden können
- die Grenzen der Verarbeitung großer Datenmengen mit Datenbanken und Datenstromsystemen erkennen und erklären
- Datenstromsysteme als Lösung für die Herausforderungen bei der Verarbeitung von Big Data erläutern
- mögliche Probleme und Fehlerquellen moderner Echtzeit-Datenanalysen erkennen
- die Alltagsrelevanz von Echtzeitanalysen erkennen und Beispiele für deren Nutzung im Alltag nennen

Ein möglicher **Unterrichtsablauf** gliedert sich in folgende Phasen:

1. **Echtzeit-Datenverarbeitung im Alltag:** Als Kontext werden verschiedene Beispiele diskutiert, bei denen Echtzeitdatenverarbeitung benötigt wird, z. B. die Analyse von Sensordaten (z. B. Tsunamierkennung), von Kreditkartentransaktionen (z. B. Ablehnung verdächtiger Kreditkartentransaktionen) oder von Social-Media-Daten (z. B. Beurteilung des Erfolgs eines Produkts).
2. **Grenzen der Datenverarbeitung mit Datenbanksystemen:** Am Beispiel der Analyse von Beiträgen in sozialen Medien anhand bestimmter Stichwörter (z. B. Produktnamen, Marken) werden die Grenzen beim Einsatz einer (relationalen) Datenbank, insbesondere hinsichtlich der Komplexität des Datenmodells und des Speicherplatzbedarfs, diskutiert.
3. **Skalierung und Datenstromsysteme:** Vor diesem Hintergrund kann ausgelotet werden, wie diese Grenzen überwunden werden können: Neben den Grenzen der vertikalen Skalierung kann horizontale Skalierung als Möglichkeit – die aber die Echtzeitfähigkeit einschränkt – diskutiert werden. Als innovativer Ansatz wird die direkte Verarbeitung des Datenstroms mit Datenstromsystemen vorgestellt.
4. **Kennenlernen von Datenstromsystemen:** Hierzu kann ein einfaches Datenstromsystem herangezogen werden. Für diesen Zweck entwickeln wir derzeit eine Anwendung, die es erlaubt, den Datenstrom von Twitter in Echtzeit mit durch den Nutzer (in gewissem Maße) anpassbaren Abfragen zu analysieren. Bei der eigenen Implementierung einer solchen Anwendung können die Lernenden die Prinzipien und Konzepte von Datenstromsystemen selbst erkennen und umsetzen.
5. **Grenzen von Datenstromsystemen:** Beim „Versuch“ nachträglich Abfragen zu bearbeiten, erkennen die Schüler die Grenzen eines Datenstromsystems. Sie erkennen, dass die Nutzung von Datenstromsystemen kaum möglich ist, wenn die Kriterien zur

Formulierung einer Abfrage vorher nicht genau bekannt sind. Somit sind mit Datenstromsystemen z. B. Ausreißeranalysen nur möglich, wenn ein üblicher Wertebereich definiert werden kann. In Datenbanken hingegen kann der übliche Bereich und Ausreißer aus diesem im Nachhinein statistisch analysiert werden. Je nach Anwendungszweck (z. B. Tsunamivorhersage) ist jedoch die Analyseverzögerung von Nachteil.

Weder DBMS noch DSS können in allen Anwendungsfällen gleichermaßen eingesetzt werden. Sie stellen unterschiedliche Herangehensweisen an die Datenverarbeitung dar, an denen gut erkannt werden kann, dass für verschiedene Anwendungszwecke verschiedene Lösungen nötig sind. Gerade im Zusammenhang mit immer größeren Hauptspeichermengen gewinnen Datenstromsysteme an Bedeutung: Die flüchtige Speicherung und schnelle Verarbeitung erlaubt zeitnah zur Verfügung stehende Resultate, die jedoch durch genauere (da auf einem größeren Datenbestand basierende), aber langsamer erzeugte Ergebnisse, aus datenbankbasierten Analysen unterstützt werden können.

4 Diskussion

Die beiden beschriebenen Beispiele greifen zwei wesentliche Aspekte von Big Data auf, die auch im Alltag immer stärker an Bedeutung gewinnen, und zeigen an ihnen innovative Vorgehensweisen der Informatik auf: Durch die immer größer werdenden Datenmengen werden einerseits immer umfangreichere Analysen der Daten möglich, andererseits werden Daten auch immer schneller verarbeitet, idealerweise in Echtzeit. Während heute die Gefahren von und kritische Sichtweisen auf die Erhebung großer Datenmengen ausführlich diskutiert werden, u. a. auch im Zusammenhang mit dem NSA-Skandal, nutzt gleichzeitig jeder die vielfältigen Möglichkeiten moderner Datenverwaltung und generiert dabei umfangreiche Datenmengen. Dabei kommt auch dem dargestellten Data Mining und der Echtzeitanalyse, möglicherweise unter Nutzung von Datenstromsystemen, eine wesentliche Bedeutung zu. Ein grundlegendes Verständnis der Grundlagen, Herangehensweisen und Ideen dieser Systeme wird dabei immer wichtiger, da sie heute allgegenwärtig sind. Insbesondere ist auch die Teilnahme am gesellschaftlichen Diskurs und die Einschätzung solcher Anwendungen und Dienste nur möglich, wenn eine entsprechende Grundlage durch den Informatikunterricht gelegt wurde. Durch die zuvor dargestellten Beispiele konnte exemplarisch gezeigt werden, dass Big Data Aspekte beinhaltet, die für den Informatikunterricht auf einem geeigneten Niveau didaktisch reduziert unterrichtet werden können, und somit entgegen des ersten Eindrucks nicht grundsätzlich zu komplex für den Schulunterricht ist. In beiden Beispielen können dabei Aspekte erkannt werden, die als potentielle fundamentale Ideen [Sc93] gehandelt werden können: Neben den Datenanalysemethoden sind unter anderem auch das Selektieren geeigneter Daten, der Umgang mit Schnittstellen/APIs, das Erkennen von Möglichkeiten und Grenzen von Informatiksystemen, Parallelisierung bzw. Aufsplittung von Daten zur effizienteren Verarbeitung, die Betrachtung von Stichproben (Sampling), u. v. m. in diesen Beispielen enthalten. Diese Ideen können nur in einem Informatikunterricht vermittelt werden, der sich zwar auf die Grundlagen konzentriert, dabei aber moderne Entwicklungen und Beispiele miteinbezieht. Durch die Beschäftigung mit Themen wie moderner Datenverarbeitung und Datenmangement bzw. Big Data, kann der Informatikunterricht wesentlich zu einer überlegten Positionierung im Spannungsfeld, das sich durch diese neuen Möglichkeiten ergibt, beitragen: Jeder muss heute zwischen dem

Schutz der eigenen Privatsphäre durch Vermeidung der Weitergabe von Daten und dem Komfort, den viele moderne Systeme bieten, abwägen – eine Herausforderung, die ohne entsprechendes (informatisches) Hintergrundwissen kaum bewältigt werden kann.

Literaturverzeichnis

[Br12] Brewer, Eric: CAP twelve years later: How the "rules" have changed. Computer, 45(2):23–29, 2012.

[Bu14] Burns, Will: Taxi Stockholm Shows Us That Big Data Needs A Big Idea. 2014. http://onforb.es/1shlLLL, abgerufen: 20.04.2015.

[DA09] DAMA International: The Dama Guide to the Data Management Body of Knowledge . Take It With You. Technics Publications Llc, 2009.

[DN13] Dorschel, Joachim; Nauerth, Philipp: Big Data und Datenschutz – ein Überblick über die rechtlichen Herausforderungen. Wirtschaftsinformatik & Management, (2):32–38, 2013.

[Ed11] Edlich, Stefan; Friedland, Achim; Hampe, Jens; Brauer, Benjamin; Brückner, Markus: NoSQL. Hanser, Carl Gmbh + Co., 2011.

[ER13] Eifrem, Emil; Rathle, Philip: Why the most important part of Facebook Graph Search is "Graph". 2013. http://neo4j.com/?p=81, abgerufen: 20.04.2015.

[FU14] FUJITSU Technology Solutions GmbH: White Paper Lösungsansätze für Big Data. 2014. http://globalsp.ts.fujitsu.com/dmsp/Publications/public/wp-bigdata-solution-approaches-de.pdf, abgerufen: 20.04.2015.

[GR14] Grillenberger, Andreas; Romeike, Ralf: A Comparison of the Field Data Management and its Representation in Secondary CS Curricula. In: Proceedings of WiPSCE 2014. ACM, Berlin, 2014.

[He14] Heuzeroth, Thomas: Amazon verschickt Waren schneller, als Sie kaufen. Die Welt, 2014. http://www.welt.de/wirtschaft/webwelt/article123990975, abgerufen: 20.04.2015.

[HNP09] Halevy, Alon; Norvig, Peter; Pereira, Fernando: The unreasonable effectiveness of data. Intelligent Systems, IEEE, 24(2):8–12, 2009.

[KE13] Kemper, Alfons; Eickler, André: Datenbanksysteme. Oldenbourg Wissensch.Vlg, 2013.

[Kr13] Krikorian, Raffi: New Tweets per second record, and how! 2013. https://blog.twitter.com/node/2845, abgerufen: 20.04.2015.

[La01] Laney, Douglas: 3D Data Management: Controlling Data Volume, Velocity, and Variety. Bericht, META Group, 2001.

[MC13] Mayer-Schönberger, Viktor; Cukier, Kenneth: Big Data - Die Revolution, die unser Leben verändern wird. FinanzBuch Verlag, München, 2013.

[Sc93] Schwill, Andreas: Fundamentale Ideen der Informatik. Zentralblatt für Didaktik der Mathematik, 1993.

[Wo11] World Economic Forum: Personal Data: The Emergence of a New Asset Class. 2011. http://www3.weforum.org/docs/WEF_ITTC_PersonalDataNewAsset_Report_2011.pdf, abgerufen: 20.04.2015.

[Wo14] World Economic Forum: Rethink Personal Data: Trust and Context in User-Centred Data Ecosystems. 2014. http://www3.weforum.org/docs/WEF_RethinkingPersonalData_TrustandContext_Report_2014.pdf, abgerufen: 20.04.2015.