Fraunhofer
SIT

FRAUNHOFER-INSTITUT FÜR SICHERE INFORMATIONSTECHNOLOGIE

BEGLEITPAPIER BÜRGERDIALOG

CHANCEN DURCH BIG DATA UND DIE FRAGE DES PRIVATSPHÄRENSCHUTZES

EC SPRIDE

Big Data und Privatheit

Dr.-Ing. Martin Steinebach, Oren Halvani, Marcel Schäfer, Christian Winter und York Yannikos

Fraunhofer-Institut für Sichere Informationstechnologie SIT – November 2014

Ziel dieses Berichts ist es, eine bürgernahe Einführung in das Thema Big Data und in die damit einhergehenden Chancen für die Gesellschaft und Risiken für die Privatsphäre zu geben. Unter Big Data wird das Erheben, Speichern, Zugreifen und Analysieren von großen und teilweise heterogenen, strukturierten und unstrukturierten Datenmengen verstanden.

Big Data stellt eine neue Herangehensweise an den Umgang mit großen Datenmengen dar. Durch neue Algorithmen, die selbstständig Muster und Zusammenhänge in Daten erkennen können, und durch neue Hardware-Lösungen, die in der Lage sind, eine große Datenmenge zeitnah zu verarbeiten, werden die Möglichkeiten für Datenanalysen erheblich vervielfältigt. Das volle Potenzial entfaltet Big Data dann, wenn Analysten in Echtzeit Zusammenhänge in Daten herstellen und prüfen können, um neue Erkenntnisse aus den Daten zu gewinnen. Auch die Datenquellen, die als Basis für die Analysen dienen, sollten möglichst aktuell sein und als kontinuierlicher Fluss von Informationen dem System zugeführt werden.

In den letzten Jahren ergab sich durch Big Data eine Reihe neuer IT-Lösungen in unterschiedlichen Bereichen der Gesellschaft. Beispielanwendungen – siehe Abschnitt 2 – umfassen die Medizin (Prognose von Grippewellen durch Google und die Unterstützung bei der Krebsdiagnose durch IBM Watson), die Polizeiarbeit

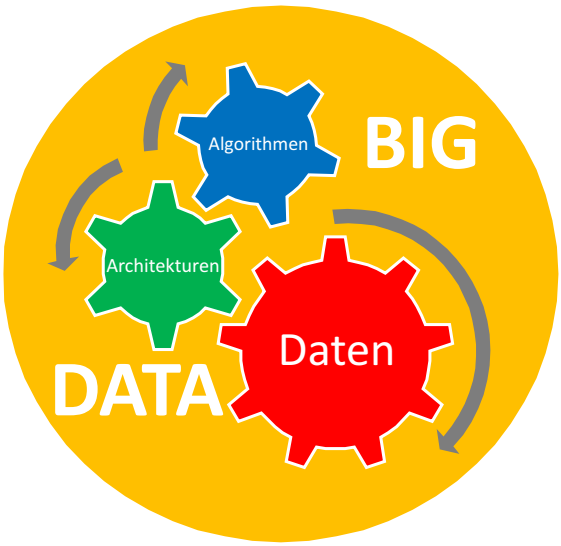

Abbildung 1: *Big Data kombiniert*

(Senken der Kriminalitätsrate durch PredPol und die Festnahme des sogenannten Autobahnschützen durch Ermittlungen des BKA), die Geheimdienste (am Beispiel der Werkzeuge der NSA), die Wirtschaft (Optimierung von Geschäftsprozessen durch Business Intelligence) oder auch die Finanzbranche (Bemessen der Kreditwürdigkeit durch Scoring).

Es wird deutlich, dass Big Data gleichzeitig sowohl eine Chance als auch ein Risiko für die Gesellschaft ist: Die mit der Technologie gewonnenen Erkenntnisse helfen, die Gesundheit und Sicherheit der Bevölkerung zu verbessern und schaffen neue Geschäftsmodelle. Gleichzeitig schaffen sie ein noch nie erreichtes Überwachungspotenzial und verleiten dazu, Individuen auf Zahlen und statistische Faktoren zu reduzieren.

Fraunhofer SIT, Rheinstraße 75, 64295 Darmstadt,
`https://www.sit.fraunhofer.de`
Ansprechpartner: Martin Steinebach, Tel. 06151 869-349,
E-Mail `martin.steinebach@sit.fraunhofer.de`
Dieses Dokument wurde mit LaTeX erzeugt. Das Layout basiert auf einer Vorlage von Frits Wenneker (`http://www.howtotex.com`) und Velimir Gayevskiy (`http://latextemplates.com`).

Big Data kann dazu verleiten, Individuen auf Zahlen zu reduzieren.

Dementsprechend wichtig ist es, dass die Gesellschaft sich damit auseinandersetzt, in welchem Ausmaß und unter welchen Bedingungen Big Data auf personenbezogene Daten angewandt werden soll. Während Datenschützer hier fehlende Transparenz bemängeln, sehen Teile der Industrie im Datenschutz eine Hürde für das Ausschöpfen der Möglichkeiten von Big Data (Abschnitt 4). Ein erster Schritt für den Bürger ist die Kenntnis über die Gesetzeslage hinsichtlich der Verwendung von personenbezogenen Daten (Abschnitt 4.2). Das Bundesdatenschutzgesetz gebietet einen zurückhaltenden Umgang mit diesen und gewährt das Recht auf Einsicht und Korrektur. Anzumerken ist, dass eine praktische Umsetzung dieser Rechte oft mit Hürden versehen ist.

In diesem Bericht wird ein Schwerpunkt auf (die Thematik von) Profiling und Scoring gelegt (Abschnitt 5). Hier handelt es sich um Ausprägungen von Big Data, die anhand vielfältiger Erfahrungswerte eine Bewertung einer einzelnen Person automatisiert durchführen. Persönliche Daten wie Alter, Geschlecht, Wohnort und Beruf, aber auch Informationen aus sozialen Netzwerken oder dem Zahlungsverhalten bei Schulden werden zusammengeführt, um beispielsweise zu entscheiden, ob eine Bestellung per Vorkasse bezahlt werden muss oder per Rechnung bezahlt werden kann. Entsprechende Vorgehensweisen bergen die Gefahr in sich, dass Personen aufgrund ihres Umfeldes ungerecht eingestuft werden oder dass sich ein verhältnismäßig geringfügiges Fehlverhalten in der Vergangenheit lange auf die individuellen Chancen im weiteren Leben auswirkt (Abschnitt 5.2). Außerdem erfahren Personen in der Regel nicht, welche Profile über sie angelegt werden und nach welchen Kriterien sie behandelt werden, wenngleich ein solches Vorgehen in Europa als illegal angesehen wird.

Neben den gesellschaftlichen Fragestellungen widmet sich dieser Bericht ebenfalls den grundlegenden Technologien. Big Data bedeutet immer das Zusammenspiel von leistungsfähigen Rechnerarchitekturen und geeigneter Software. Abschnitt 3 führt in für Big Data notwendige Aspekte wie verteiltes Rechnen oder In-Memory-Datenbanken ein. Weiterhin werden auch Konzepte wie maschinelles Lernen und Data-Mining erläutert. So wird deutlich, wie durch eine Reihe von technischen Innovationen in der jüngeren Vergangenheit eine neue Herangehensweise an die Datenverarbeitung möglich wurde.

1 Einführung

In der öffentlichen Wahrnehmung tritt Big Data oft als eine Revolution im Umgang mit Informationen in Erscheinung. Tatsächlich handelt es sich aber um eine Evolution der Werkzeuge, welche die Datenmengen verarbeiten. Diese erreichen inzwischen eine Qualität und Komplexität, welche die Möglichkeiten herkömmlicher Datenverarbeitung im Sinne von Datenbankabfragen oder statistischen Übersichten weit übertreffen. Durch Big Data werden komplexe Zusammenhänge zwischen unterschiedlichen Daten sichtbar und handhabbar. In Kombination mit stetig wachsenden Datenvolumina und Computerressourcen können so Erkenntnisse gewonnen werden, die von hohem Wert für Behörden, Wirtschaft und Wissenschaft sind.

Fiktives Beispiel „Speiseeis“

Um die Idee von Big Data zu verdeutlichen, nutzen wir ein fiktives und einfaches Beispiel. Wir betrachten, welche Methoden beim Verkauf von Speiseeis herangezogen werden könnten. Dazu gehen wir von folgendem Szenario aus: Eine Eisdiele verkauft das ganze Jahr über Speiseeis. Jeden Morgen werden verschiedene Sorten in verschiedenen Mengen produziert. Im Laufe des Tages werden die Sorten verkauft. Teilweise bleiben Bestände übrig; manchmal geht eine Sorte vor Geschäftsschluss aus und Kunden verzichten auf ihr Eis.

Um die Mengen besser abschätzen zu können, wäre es möglich, Buch zu führen, wie gut sich welche Sorte wann verkauft hat. Wahrscheinlich käme man zu dem Schluss, dass im Sommer etwas mehr Fruchteis und im Winter mehr Milcheis verkauft wird. Entsprechend wird

Abbildung 2: *Der Konsum von Speiseeis kann von vielen Faktoren abhängen. Big Data kann helfen, diese zu entdecken.*

produziert. An einem sonnigen Wintertag kann dies dazu führen, dass nicht ausreichend Fruchteis vorhanden ist. Ein einfaches Modell auf Basis des Kalenders reicht also nicht aus, um den Bedarf wirklich vorherzusagen. Dieses Modell wäre ein Beispiel für eine einfache statistische Lösung, die einen Bezug zwischen dem Datum und dem Verbrauch darstellt.

Eine Big-Data-Lösung würde hinsichtlich der Quellen für die Verbrauchsprognose weiter gehen. Neben den Erfahrungen in Abhängigkeit mit dem Datum wäre ein Bezug zum Wetterbericht interessant: Wie waren die Verkäufe an einem sonnigen Dezembersonntag, wie an einem regnerischen Septembermittwoch. Werden diese Erkenntnisse mit der Wettervorhersage kombiniert, dann lässt sich der Verkauf genauer prognostizieren, falls die Wettervorhersage zutrifft. Hat man ausreichend Datenquellen zur Verfügung, werden eventuell auch unerwartete Zusammenhänge deutlich: Findet ein Bundesligaspiel der regionalen Mannschaft statt, wird weniger Nusseis verkauft. Warum das der Fall ist, kann Big Data nicht beantworten. Eine Verknüpfung der Prognose mit einem Spielplan der Bundesliga hilft aber trotzdem, bessere Vorhersagen zu treffen.

Big Data hat häufig zum Ziel, Zusammenhänge zu erkennen und so bei Entscheidungen zu helfen.

Tatsächlich ist das die Motivation für den Einsatz von Big Data: Zusammenhänge erkennen und dann für Entscheidungen nutzen – hier zur Prognose des zu erwartenden lokalen Eiskonsums. Dabei müssen nicht die Ursachen für die Zusammenhänge aufgespürt und verstanden werden, sondern nur die Zusammenhänge selbst geschickt genutzt werden. Ob der Einbruch im Verkauf von Nusseis an Spieltagen möglicherweise daran liegt, dass Männer die Hauptkonsumenten von Nusseis sind und durch das Spiel weniger Männer Eis essen gehen, ist für die Produktion völlig unerheblich. Es wird erkannt, dass ein Spiel ansteht, eine Prognose über einen reduzierten Bedarf von Nusseis erstellt und die Produktion angepasst.

Natürlich ist eine einzelne Eisdiele noch kein Big Data und die Zusammenhänge kann ein erfahrener Eisverkäufer vielleicht schon ohne die Hilfe eines Computers erkennen. Wenn wir die Eisdiele mit der Filiale einer deutschlandweiten Kette für Speiseeis ersetzen, welche zentral entscheiden muss, welches Eis in welchen Mengen wohin geliefert werden muss und welche Zutaten dafür eingekauft werden müssen, kommen wir allerdings schon in entsprechende Bereiche. Die einzelnen Filialen können den Verbrauch an die Zentrale melden, dort werden die neuesten Wetterprognosen verfolgt und so wird anhand von Verbrauch und Prognose eine optimale Versorgung der Filialen sichergestellt. Ab einer gewissen Komplexität, wenn neben Fernsehprogramm und Wetter beispielsweise auch die lokalen Nachrichten analysiert werden (Vielleicht führen Reiseberichte zu einem hohen Verkauf von exotischen Eissorten?) oder soziale Netzwerke beobachtet werden (Wie wirkt sich eine positive oder negative Erwähnung einer Eissorte der Filiale auf Facebook auf den Verkauf aus?), wird der Betreiber des Systems die Zusammenhänge nicht mehr wirklich durchschauen, sondern vergleichsweise blind den Prognosen vertrauen. Und er wird so in den meisten Fällen den Bedarf gut abschätzen.

2 Beispielanwendungen

Big Data kann in den unterschiedlichsten Domänen eingesetzt werden. So helfen Big Data Anwendungen im Gesundheitswesen und der medizinischen Forschung, z. B. mittels datengestützter Diagnose und Behandlung. Bereits jetzt werden für Wettervorhersagen und Klimamodelle Big-Data-Technologien verwendet, um dynamische und möglichst echtzeitfähige Modelle zu erstellen. Ähnliche Verwendungsmöglichkeiten bestehen u. a. in der Weltraumforschung und für Teilchenbeschleuniger. Weitere Anwendungsfelder ergeben sich bei Sicherheits- und Polizeiarbeit sowie bei der Infrastruktur von Mobilfunknetzen, Internet und intelligenten Stromnetzen (sog. Smart Grids). Auch für Meinungs- und Trendforschung mittels Daten aus sozialen Medien verspricht Big Data enormes Potenzial. Der offensichtlichste Vorteil bzw. Nutzen wird wohl den Bereichen Wirtschaft und Konsum zugesprochen. Ob bei Werbung, Kundenbindung und -analyse oder im Kreditwesen, in der Finanz- und Versicherungsmathematik oder bei der sogenannten Business Intelligence für Unternehmen – Big Data findet hier vielfältige Einsatz- und Optimierungsmöglichkeiten.

In den folgenden Abschnitten sind konkrete Beispiele für Big-Data-Lösungen aus manchen der genannten Bereiche aufgeführt. Damit einhergehende Risiken für die Privatsphäre werden ebenfalls betrachtet.

2.1 Googles Grippe-Trends

Menschen, die von Grippe betroffen sind, geben bei Google häufig entsprechende Suchbegriffe ein. Dadurch kann Google aus den Suchanfragen die aktuelle Grippeverbreitung schätzen. Diese Informationen sind schneller verfügbar als Daten aus institutionellen Beobachtungsprogrammen wie dem *European Influenza Surveillance Scheme* (EISS) und können so einen Beitrag für die

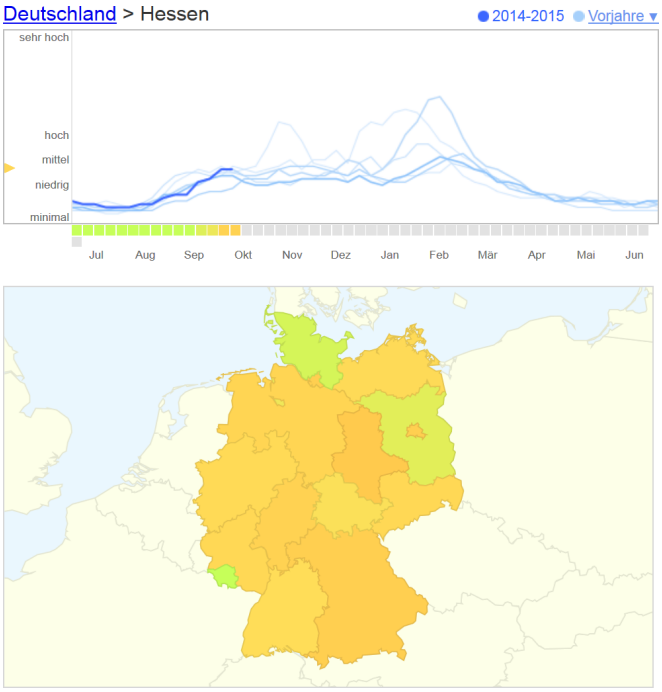

Abbildung 3: *Google Grippe-Trends für Hessen. Wie zu erwarten steigen die Suchanfragen zum Thema Grippe im Herbst. Quelle: Google* `http://www.google.org/flutrends/intl/de/de/#DE-HE`

Früherkennung, Prävention und Bekämpfung von Grippe leisten.

Die erhobenen Daten bestehen aus allen Suchanfragen, die bei Google eingegeben werden. Die Daten beinhalten neben den Suchbegriffen auch den Zeitpunkt der jeweiligen Suchanfrage und den Ort, der durch die IP-Adresse des Nutzers bestimmt wird.

Zunächst hat Google aus den 50 Millionen häufigsten Suchphrasen in einer Historie von fünf Jahren diejenigen Phrasen ermittelt, die am besten mit den Grippedaten der US-amerikanischen *Centers for Disease Control and Prevention* (CDC) zusammenhängen. Daraus ist ein empirisch validiertes Modell entstanden, das aus dem Verhältnis zwischen Suchanfragen zum Thema Grippe und allen übrigen Suchanfragen die Häufigkeit aktueller Grippefälle schätzt [11]. Entsprechende Modelle wurden auch für andere Länder gebildet, u. a. auch für Deutschland (siehe Abbildung 3).

Google veröffentlicht tagesaktuelle „Grippe-Trends“ für mehr als 25 Länder unter `http://www.google.org/flutrends/intl/de`. Dadurch soll ermöglicht werden, früher und effizienter auf Grippewellen zu reagieren. Beispielsweise soll die Produktion und Verteilung von Impfstoffen und Medizin optimiert werden.

Die Grippe-Trends beruhen auf räumlich und zeitlich stark aggregierten Daten, sodass die Privatsphäre von Personen davon nicht betroffen ist. Jedoch besitzt und nutzt Google auch die Rohdaten der Suchanfragen, die über IP-Adressen und Cookies zu umfangreichen Profilen zusammengefügt werden. Google gab 2009 an, Suchanfragen nach neun Monaten zu anonymisieren [11]. Durch weitere Dienste von Google liegen zusätzlich viele Daten und somit viel Wissen über den Internetnutzer bei einem einzigen Konzern. Die Auswirkungen und datenschutzrechtlichen Bedenken werden in Abschnitt 4.1 detaillierter aufgegriffen.

2.2 Watson gewinnt bei Jeopardy

Watson ist ein sogenanntes kognitives Computersystem von IBM, welches Informationen in natürlicher Sprache verarbeitet und basierend hierauf Fragen in natürlicher Sprache beantworten kann. Watson ist benannt nach Thomas J. Watson, einem der Gründungsväter und langjährigem Leiter des IBM-Konzerns. Im Jahr 2011 gewann Watson gegen die zwei menschlichen Champions Ken Jennings und Brad Rutter in der US-amerikanischen Quizshow „Jeopardy!“ (siehe Abbildung 4). In dieser Show bestehen die Rätsel aus einer Aussage (engl. *clue*), zu der die Teilnehmer die Lösung als Fragesatz formulieren müssen. Dabei gilt es die Lösungen in Sekundenschnelle zu finden, um den Mitspielern zuvorzukommen.

Watson war in der Sendung wie die menschlichen Teilnehmer auf sein mitgebrachtes Wissen angewiesen, d. h. er hatte keinen Internetzugang. Sein Gedächtnis bestand aus dem Wissen von umgerechnet 200 Millionen Buchseiten, u. a. aus Wikipedia, der Bibel und allen Ausgaben der New York Times der vorherigen zehn Jahre.

Watson bekam in der Quizshow die Rätsel in dem Moment als Text zugespielt, in dem sie den Teilnehmern angezeigt und vorgelesen wurden. So konnte Watson beginnen, die Rätsel zu verarbeiten und in seinem Wissen nach Assoziationen zu suchen, sobald auch die mensch-

Abbildung 4: *Watson beim Wettstreit mit Jennings und Rutter in der Jeopardy-Show. Quelle: YouTube* `http://www.youtube.com/watch?v=lI-M7O_bRNg`

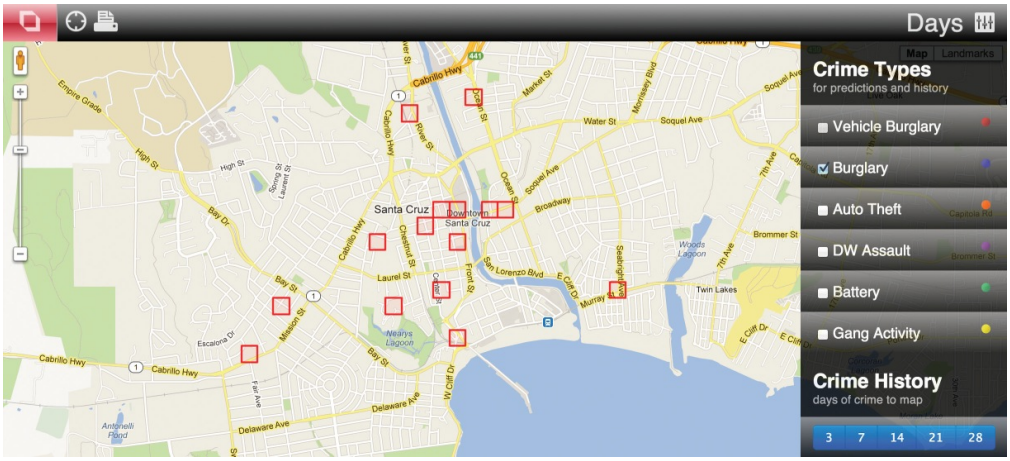

Abbildung 5: *Predictive Policing: PredPol markiert Bereiche mit erhöhter Wahrscheinlichkeit für Straftaten auf einer Google-Maps-Karte. Quelle: [1]*

lichen Teilnehmer darüber nachdachten. Watson lief bei seinem Jeopardy-Auftritt auf einem Rechnerverbund mit 2880 logischen Prozessorkernen und 15 Terabyte Arbeitsspeicher.

Watson benutzt zahlreiche Technologien, um zu den gestellten Rätseln die am wahrscheinlichsten passendste Frage formulieren zu können. Zu den Technologien zählen u. a. maschinelle Sprachverarbeitung, maschinelles Lernen (siehe Abschnitt 3.3), Logik, Suchmaschinenverfahren (Volltextsuche, semantische Abfragen, etc.) sowie diverse Heuristiken und Kategorisierungsmechanismen, um Querbezüge herzustellen. Darüber hinaus benötigt Watson Zugriff auf entsprechendes Hintergrundwissen in Form von Datenbanken und unstrukturierten Texten, die dabei unterschiedlich annotiert sind (z. B. Redewendungen/Phrasen, Wortsynonyme und andere semantische Relationen).

IBM macht Watson-Technologie sukzessive für verschiedene Bereiche anwendbar, unter anderem im Kundenservice, Gesundheitswesen (insbesondere für die Krebsbehandlung) und in der Finanzbranche. Nach den Vorstellungen von IBM werden kognitive Anwendungen in der Zukunft allgegenwärtig sein.

2.3 Predictive Policing

Das *Los Angeles Police Department* nutzt eine Big-Data-Anwendung, um wahrscheinliche Schauplätze zukünftiger Verbrechen vorherzusagen und hier die Polizeipräsenz zu verstärken (siehe Abbildung 5). Es gab 33 Prozent weniger Einbrüche, 21 Prozent weniger Gewaltverbrechen und 12 Prozent weniger Eigentumsdelikte in dem Gebiet, in welchem die Vorhersage-Algorithmen eingesetzt wurden [18].

PredPol führte zu 33% weniger Einbrüchen, 21% weniger Gewaltverbrechen und 12% weniger Eigentumsdelikten.

Hierfür wurden zunächst die Daten von 13 Millionen Delikten aus den vergangenen acht Jahrzehnten in das zugrunde liegende mathematische Modell eingespeist. Solche Daten sind beispielsweise die Geodaten vergangener Hauseinbrüche, Autodiebstähle und Überfälle, welche die Polizei erfasst hat. Der Datenbestand wird ständig durch neue, aktuelle Straftaten ergänzt.

Das Verfahren basiert auf einem Algorithmus zur Vorhersage der Nachbeben von Erdbeben. Es wurde festgestellt, dass diese Nachbeben Mustern gehorchen und mit einer ausreichend großen Datenbasis mit einer

Abbildung 6: *PredPol ist ein Beispiel dafür, wie Science-Fiction Wirklichkeit wird. Die Idee der Erkennung von Straftaten vor ihrer Ausführung wurde bereits 1956 von Philip K. Dick in seiner Kurzgeschichte* The Minority Report *thematisiert und 2002 im Hollywood-Blockbuster* Minority Report *verfilmt. Quelle: 20th Century FOX*

hohen Zuverlässigkeit prognostiziert werden können. Wie beim Ursprungsverfahren zeigen sich auch nach Straftaten entsprechende „Nachbeben“ in Form von kriminellen Aktivitäten, die in der Zukunft mit einer gewissen Wahrscheinlichkeit eintreten werden. Die Software, die in Zusammenarbeit mit der University of California und dem Startup PredPol entstand, wurde vielmals optimiert und ist mittlerweile in der Lage, Daten in Echtzeit zu verarbeiten. Das bedeutet, sobald eine Straftat erfasst wird und die entsprechenden Daten, z. B. der exakte Ort des Geschehens, in das Modell eingespeist werden, liefert das Modell darauf aufbauend bereits Vorhersagen für die nächsten Straftaten.

In der hier dargestellten Form von Predictive Policing wird die Privatsphäre von Bürgern wenig berührt. Neben Geodaten mit persönlichem Bezug, z. B. Wohnungen und Häuser, in die eingebrochen wurde, könnten aber auch weitere personenbezogene Daten erhoben werden. So lassen sich Szenarien erdenken, die tief in die Privatsphäre eindringen. Beispielsweise lässt sich aus Daten von Jugendämtern über besonders auffällige Familien oder Daten von Schulen über Schulschwänzer und Schulhofdelikte für einzelne Personen oder Familien errechnen, wie wahrscheinlich eine Straftat von oder an den betroffenen Personen begangen wird. In Romanen und Filmen werden fiktive Methoden von Predictive Policing zur Schaffung von Dystopien wie *Minority Report* genutzt (siehe Abbildung 6).

2.4 BKA klärt Autobahnschüsse

Von 2008 bis 2013 versuchte die Polizei einen Autobahnschützen zu finden. Auf deutschen Autobahnen wurde mehr als 750 mal willkürlich auf Fahrzeuge geschossen – vorzugsweise Autotransporter, meist auf der Gegenfahrbahn. Eine Autofahrerin wurde im November 2009 am Hals getroffen. Durch die Einschusswinkel an den Fahrzeugen war davon auszugehen, dass der Täter in einem Lkw saß. Weiterführende Erkenntnisse konnten durch Fahndungsfahrten mit Autotransportern, öffentlichen Aufrufen an Berufskraftfahrer und Bürger, einer ausgeschriebenen Belohnung von 100.000 Euro sowie eines Berichts bei *Aktenzeichen XY* nicht gewonnen werden. Erschwerend für die Ermittlungen war, dass die Einschüsse meist erst am Fahrtziel bemerkt wurden. Erst durch eine groß angelegte Datenerfassung und -auswertung konnte der Täter ermittelt und im Juni 2013 verhaftet werden. Eine chronologische Übersicht der Ereignisse bis zur Verhaftung wurde vom BKA veröffentlicht (siehe Abbildung 7). Ende Oktober 2014 wurde der Schütze zu zehneinhalb Jahren Haft verurteilt.

Zum Ermittlungserfolg führte die Nutzung von 13 Kennzeichenerfassungssystemen an sieben Standorten auf Autobahnen in fünf Bundesländern von Dezember 2012 bis Juni 2013. Wenn dem BKA Schüsse gemeldet wurden, dann rekonstruierten die Ermittler die Zeiträume, in denen der Täter Kontrollpunkte passiert haben musste, und sicherten die jeweils relevanten Kennzeichendaten. Die übrigen Daten wurden jeweils automatisch zehn Tage nach deren Erhebung gelöscht. Die gesicherten Daten umfassten 3,8 Millionen Kennzeichen. Zu 50 Kennzeichen wurden die Halter ermittelt.

Während der Ermittlung wurden u. a. 3,8 Millionen Autokennzeichen erfasst.

Außerdem wurden durch Funkzellenabfragen im November 2009 rund 15.000 Datensätze sowie im Jahr 2012 rund 579.000 Datensätze erhoben. Zu 312 Rufnummern wurden die Anschlussinhaber ermittelt.

In den erfassten Daten wurden Kreuztreffer gesucht, d. h. Fahrzeuge, die in mehreren Fällen zu passender Zeit Kontrollpunkte passiert haben. Im April 2013 konnten so die Ermittlungen auf den Lkw des Täters eingegrenzt werden.

Kennzeichenerfassung und Funkzellenabfragen liefern Daten über viele unbeteiligte Bürger. Aus diesen Daten können Aufenthalts- und Bewegungsprofile abgeleitet werden. Auch wenn bei dieser Ermittlung die zugrunde liegende Datenmenge im Kontext von Big Data relativ klein ist, ist die Anzahl der betroffenen Personen groß.

Die systematische Erfassung von Kennzeichen zur Strafermittlung war bisher einmalig. Theoretisch könnten auch Daten oder Infrastruktur des deutschen Lkw-Mautsystems genutzt werden, was jedoch gesetzlich nicht zulässig ist. Kürzlich plädierte Hans Peter Bull, ehemaliger Bundesbeauftragter für den Datenschutz (1978–1983), in der aktuellen Diskussion um die Pkw-Maut für eine gesetzliche Erlaubnis zur Nutzung von Mautdaten zur Bekämpfung von Straftaten [5].

Funkzellenabfragen gehören zur gängigen Ermittlungspraxis. Hierbei besteht die Kritik, dass Funkzellenabfragen zu häufig eingesetzt werden, oft nicht im Verhältnis zur Straftat stehen und zu selten nötig oder nützlich für die Ermittlungen sind. Der Berliner Datenschutzbeauftragte Alexander Dix bemängelt, dass Löschfristen oft nicht beachtet werden oder die Löschung nicht dokumentiert wird, und dass die gesetzlich geregelte Benachrichtigung von Betroffen oft versäumt wird [8].

Nach der Festnahme des Autobahnschützen wurde mehrmals die Zulässigkeit der vorangegangenen Kennzeichenerfassung öffentlich diskutiert. Das BKA weist den Vorwurf der unverhältnismäßigen Datensammelei

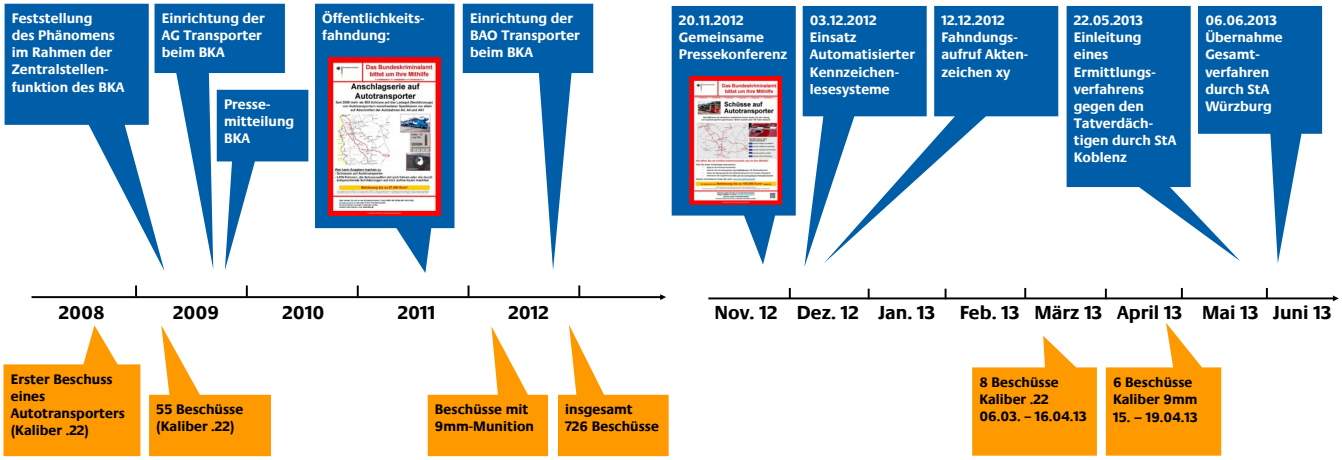

Abbildung 7: *Chronologie der Fahndung zum Autobahnschützen. Quelle: BKA, Pressemitteilung vom 25.06.2013* `https://www.bka.de/nn_196810/DE/Presse/Pressemitteilungen/Presse2013/130625_BAOTransporterPressekonferenz.html`

zurück. Die Bundesregierung hat im September 2013 eine „Kleine Anfrage“ der Linken zu der Datenerhebung beantwortet [6], woraus die meisten hier genannten Zahlen stammen. Der rheinland-pfälzische Datenschutzbeauftragte Edgar Wagner sieht keine ausreichende Rechtsgrundlage für die Kennzeichenerfassung und fordert gesetzliche Neuregelungen [25]. Die Anwälte des Schützen forderten ein Beweisverwertungsverbot.

2.5 Überwachung durch NSA und GCHQ

Im Juni 2013 veröffentlichte der britische *Guardian* geheime Dokumente, die ihm durch den früheren NSA-Mitarbeiter Edward Snowden übermittelt worden waren. Aus diesen Dokumenten geht hervor, dass der US-amerikanische Geheimdienst NSA zusammen mit dem britischen Pendant GCHQ seit spätestens 2007 einen großflächigen Überwachungsapparat installiert hat, um möglichst umfassend und global Kommunikationsdaten verdachtsunabhängig und unbemerkt mitzuschneiden, langfristig zu speichern und auszuwerten.

Ziel ist laut der überwachenden Parteien die rechtzeitige Erkennung von Bedrohungsszenarien und Verhinderung von geplanten terroristischen Anschlägen durch eine frühzeitige Identifikation involvierter, aber bisher unbekannter Personen. Aufgrund der Tatsache, dass auch sensible Daten einzelner Unternehmen erhoben und ausgewertet werden, ist anzunehmen, dass neben der Bekämpfung von Terrorismus auch wirtschaftliche Interessen bei der Überwachung eine Rolle spielen.

Aus den bisher veröffentlichten Dokumenten geht hervor, dass unter anderem folgende Daten erhoben werden:

- Sämtliche Verbindungsdaten aus E-Mail-Verkehr und Telefongesprächen in den USA, vollständige Telefongespräche von 122 Regierungschefs weltweit.
- Kommunikationsdaten zahlreicher Botschaften.
- Standortdaten von Mobiltelefonen.
- Benutzerdaten von Firmen wie Google, Yahoo, Microsoft oder Facebook.
- Kommunikations- und Benutzerdaten unterschiedlicher Personengruppen, z. B. Benutzer der Anonymisierungs-Software Tor, Mitglieder von Gruppierungen wie Anonymous oder Anhänger bestimmter Religionen wie beispielsweise des Islams.

Eine umfangreiche und ständig aktualisierte Liste der Daten, die erhoben werden, ist auf *Zeit Online* zu finden [2].

Die Daten werden hauptsächlich direkt auf der Infrastruktur von Telekommunikationsprovidern oder Dienstanbietern mit großer Benutzerbasis erhoben. Ein Beispiel ist die kürzlich bekannt gewordene Operation *Eikonal*, in deren Rahmen der deutsche Geheimdienst BND zusammen mit der NSA über mehrere Jahre Kommunikationsdaten des (vom Durchsatz her weltweit größten) Internet-Knotens DE-CIX in Frankfurt überwacht hat. Des Weiteren werden trotz verwendeter Sicherheitsmechanismen Daten aus Firmennetzwerken und mittels versteckter Hintertüren aus privater und unternehmensspezifischer Hardware erhoben. Die Durchdringung durch die Geheimdienste ist enorm. Es ist davon auszugehen, dass aus praktisch jedem Haushalt und jedem Unternehmen (teilweise erhebliche Mengen an) Daten erhoben werden.

Ein für die überwachenden Parteien weiterer wichtiger Aspekt für umfangreiches und störungsfreies Datensammeln ist das Aushebeln von gängigen Verschlüsselungsverfahren, die zur Übertragung oder zur Speicherung von Daten verwendet werden. Hier werden gezielt Fehler in verbreiteter Verschlüsselungs-Software gesucht und ausgenutzt oder die Herausgabe privater Schlüssel zur Entschlüsselung erzwungen. Weiterhin ist bekannt, dass die NSA durch Mitwirkung bei der Standardisierung von Verschlüsselungsverfahren (teilweise erfolgreich) versucht, Hintertüren in diese einzubauen.

Zur Analyse des massiv hohen Datenaufkommens werden offensichtlich zahlreiche verschiedene Verfahren aus den Bereichen Information-Retrieval und Data-Mining eingesetzt. Offiziell sind diese nicht dokumentiert. Bekannt ist, dass Software wie das vielfach in den Medien aufgegriffene *XKeyscore* verwendet wird, um gezielt Inhalte aus dem Gesamtdatenbestand zu filtern und anzuzeigen (Beispiel: „Zeige alle VPN-Verbindungen vom Iran ins Ausland“).

Durch Art und Umfang der erhobenen Daten werden Analysen ermöglicht, die sich zur Bildung umfangreicher und detaillierter Profile einzelner Personen und Personenkreise eignen. Aus diesen Profilen lässt sich wiederum ableiten, von welchen Personen aus potenziell terroristische oder anderweitig als relevant definierte Handlungen ausgehen. Weiterhin lassen sich aus Daten, die in der Infrastruktur von Unternehmen abgegriffen werden, wirtschaftlich relevante Informationen extrahieren und nutzen.

Durch die Überwachung zentraler Telekommunikationsknotenpunkte ist praktisch jeder Bürger betroffen, der das Internet oder Handy-/Telefonverbindungen benutzt. Obwohl nur ein äußerst kleiner Teil der erhobenen Daten relevant für Ermittlungen mit terroristischem oder anderem strafrechtlich relevanten Hintergrund ist, werden nach aktueller Sachlage alle einmal erhobenen Daten über längere Zeiträume gespeichert und vorgehalten. Beispiele mit besonderer Beeinträchtigung der Privatsphäre sind mitgeschnittene Webcam-Aufnahmen von Yahoo-Benutzern, Gespräche über Skype oder E-Mails inklusive aller Anhänge. Mit dem aktuell bekannten Ausmaß der Überwachung kann prinzipiell jegliche Kommunikationsform über Internet oder (Mobil-)Telefon, die nicht mit sicherer Ende-zu-Ende-Verschlüsselung durchgeführt wird, als durch die NSA und das GCHQ überwacht und gespeichert angesehen werden.

Die öffentliche Reaktion auf das Bekanntwerden der Telekommunikationsüberwachung durch die NSA und den GCHQ fiel bisher unterschiedlich aus. Von politischer Seite wurden die Überwachungspraktiken scharf kritisiert, jedoch sind bisher keine signifikanten Konsequenzen gezogen worden. Im August und September 2013 wurden zahlreiche „Kleine Anfragen“, überwiegend initiiert durch die Opposition, an die Bundesregierung gestellt, die zur Klärung des Ausmaßes an Überwachung explizit in Deutschland beitragen sollten. Die Antworten darauf wurden von der Bundesregierung überwiegend als Verschlusssache erklärt und liegen der Öffentlichkeit nicht vor. Ein sogenanntes No-Spy-Abkommen, das den gegenseitigen Verzicht auf Spionage zwischen Deutschland und den USA beinhalten sollte, wurde 2013 von der Bundesregierung thematisiert, jedoch von den USA Ende Februar 2014 abgelehnt [10]. Unmittelbar nach den ersten Snowden-Veröffentlichungen wurde Ende Juni 2013 vom Generalbundesanwalt beim Bundesgerichtshof ein Beobachtungsverfahren und anschließend Vorermittlungen bezüglich der bekannt gewordenen Überwachungsmaßnahmen eingeleitet, ein Ermittlungsverfahren wurde jedoch im Mai dieses Jahres unter massiver Kritik verworfen [15]. Im März dieses Jahres wurde der NSA-Untersuchungsausschuss eingerichtet mit dem Ziel, das Ausmaß der Überwachung durch ausländische Geheimdienste in Deutschland zu klären. Im September kritisierte der Untersuchungsausschuss Behinderung bei der Aufklärungsarbeit durch die Bundesregierung [3].

Abbildung 8: *Protest gegen staatliche Überwachung. Quelle: Gamezone* `http://www.gamezone.de/Politik-Thema-237122/News/NSA-und-GCHQ-spionieren-Smartphone-Nutzer-per-Angry-Birds-und-Google-Maps-aus-1106933/`

Nicht nur von der deutschen Bevölkerung wurde Snowdens Engagement mit sehr großer Zustimmung aufgenommen. Snowden wurde unter anderem der *Right Livelihood Award* („Alternativer Nobelpreis“) verliehen, weiterhin wurde er für den diesjährigen Friedensnobelpreis nominiert. Innerhalb des vergangenen Jahres fanden global zahlreiche Demonstrationen gegen die Überwachungspraktiken statt, teilweise mit mehreren Zehntausend Demonstranten. In Deutschland wurden

weiterhin zahlreiche offene Briefe formuliert, Petitionen eingerichtet, gegen Überwachung demonstriert (siehe Abbildung 8) und Strafanzeigen erstattet, unter anderem gegen die Bundesregierung wegen Ausübung illegaler Agententätigkeit und diesbezüglich Kooperation mit britischen und US-amerikanischen Geheimdiensten [22]. Das Interesse an technischen Schutzmaßnahmen vor umfassender Überwachung ist in der Öffentlichkeit seit dem letzten Jahr gestiegen, beispielsweise erfreuen sich Crypto-Partys, bei denen man sich über Themen wie Datenschutz und Verschlüsselung informieren kann, nicht zuletzt aufgrund der zugenommenen Medienpräsenz des Themas „Überwachung durch die Geheimdienste“ größerer Beliebtheit.

2.6 Business Intelligence

Business Intelligence (BI) hat das Ziel, ökonomisches Wissen über das eigene Unternehmen und das kommerzielle Umfeld zu generieren. Dabei muss das Wissen den Entscheidungsträgern auf den unterschiedlichen Ebenen zum richtigen Zeitpunkt in entsprechender Form zur Verfügung stehen [17].

Mithilfe von BI lassen sich spezifische Muster und Zusammenhänge in (vorwiegend großen) Datenbeständen identifizieren und so mögliche Trends vorhersagen. Hierbei können sowohl strukturierte Daten, z. B. Datenbank-Tabellen, als auch unstrukturierte Daten, beispielsweise Texte aus sozialen Netzwerken, in die Analyse einbezogen werden [13]. Die erhobenen Daten werden auf internen oder externen Servern gespeichert und analysiert. Der Anwender interagiert mithilfe einer (Web-)Schnittstelle, welche oftmals vielfältige Visualisierungsmöglichkeiten anbietet. Die unterschiedlichen Sichtweisen auf die Datenmengen sollen helfen, spezifische Muster aus den Daten hervorzuheben und/oder deren Zusammenhänge zu verstehen.

Die zugrunde liegenden Technologien von BI sind maschinelles Lernen und Data-Mining (siehe Abschnitt 3.3).

Die Vorteile von BI sind u. a. besseres Verstehen und Optimieren von Unternehmensprozessen. Als Ergebnis von BI dient beispielsweise ein ausführlicher Report, Statistiken oder eine kurze Trend-Prognose. Andere Formen sind ebenfalls möglich, etwa Zusammenfassungen mit z. B. negativer oder positiver Bewertung eines Produkts.

Abbildung 9: *Business Intelligence (BI)*

BI fokussiert in erster Linie das Unternehmen bzw. dessen Produkte. Der Mensch als Individuum steht daher nicht im Vordergrund. Allerdings wird der Mensch in der Masse betrachtet, wenn z. B. Zielgruppen analysiert werden. Hierbei fallen in der Regel abstrakte, nicht auf einzelne Personen beziehbare Daten an, wie das Geschlecht oder das (ungefähre) Alter der Person.

2.7 Scoring und Kreditvergabe

Dass über die Vergabe eines Kredits unter anderem in Abhängigkeit des Wohnorts entschieden wird, ist lange bekannt. Eine Bank kann im Vorfeld Erkenntnisse über typische Kreditausfallraten in der Umgebung sammeln und so Kreditwürdigkeit und Wohnort miteinander in Verbindung setzen.

Big Data führt dieses Konzept weiter: Neben Einkommen, Vermögen und Wohnort können Bildungsstand, beruflicher Werdegang, Branche des Arbeitgebers, Familienstatus, Kfz-Besitz und viele weitere Faktoren in internen Datenbanken zusammengeführt werden. Erweitert werden diese Datenbanken um Einträge aus sozialen Netzen, dem WWW und Bewertungen von Auskunfteien wie Schufa, Creditreform, Arvato Infoscore und Bürgel Wirtschaftsinformationen. Die Abhängigkeiten zwischen diesen gesammelten Informationen und dem Tilgungsverhalten werden anhand von früheren Vorgängen hergestellt, aus denen Muster abgeleitet werden, die mit dem vorliegenden Fall verglichen werden. Auf dieser Grundlage wird für jeden Antragsteller ein individueller Score ermittelt. Dieser Score dient als Entscheidungsgrundlage für die Kreditvergabe und/oder zu welchen Konditionen der Kredit gewährt wird.

Da dieses Scoring standardisierte Bewertungen auf Grundlage von Statistiken liefert, wird trotz persönlicher Daten und individuellem Score die Individualität

Abbildung 10: *Die Kreditwürdigkeit einer Person wird mittels Scoring über komplexe Zusammenhänge berechnet.*

des Einzelnen selten berücksichtigt. So kann es vorkommen, dass man einen relativ niedrigen Score nur aufgrund des gewählten Wohnorts und dessen schlechtere bisherige Bewertung erhält. Das Thema Scoring wird in Abschnitt 5 noch einmal in allgemeinerer Form behandelt.

3 Technische Grundlagen

Der Begriff *Big Data* wird im Kontext der Informationsgewinnung aus „großen“ Datenbeständen verwendet. Dabei ist nicht der bloße Umfang eines Datenbestandes entscheidend, sondern die Kombination verschiedener technischer Herausforderungen für die Datenverarbeitung im Kontext einer immensen „Datenflut“. Big Data wird oft durch die Eigenschaften *Volume, Velocity, Variety* (kurz „3V“) und die damit verbundenen Herausforderungen charakterisiert, was auf Doug Laney zurückgeht [14].

Volume steht für eine große Datenmenge, die mit herkömmlichen Ansätzen der Datenverarbeitung kaum erschließbar ist. Eine einheitliche Grenze, ab der von Big Data gesprochen wird, existiert nicht. Üblicherweise werden bei Big Data Datenmengen verarbeitet, die mindestens im Terabyte-Bereich liegen.

Velocity steht für die hohe Datenentstehungsrate und Notwendigkeit schneller Ergebnisgenerierung, oft sogar in Echtzeit. Nur so können Anwendungen realisiert werden, die beispielsweise Kreditkartenbetrug unterbinden, Online-Kunden passende Empfehlungen geben oder Analysten eine interaktive Erkundung von Zusammenhängen in den Daten ermöglichen.

Variety bedeutet, dass unterschiedliche Datenquellen und Datenformate, die teilweise keine einfach zu verarbeitende Struktur aufweisen, gemeinsam betrachtet werden. So werden herkömmliche Datenbanken, in denen beispielsweise Personen mit Namen, Vornamen und Alter standardisiert organisiert sind, ebenso betrachtet wie Texte, aus denen Namen und andere Elemente erst ermittelt werden müssen, und Bilder, aus denen Inhalte mittels Bildanalysen erkannt werden müssen.

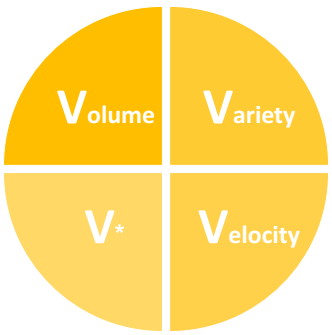

Abbildung 11: *Die grundlegenden Herausforderungen werden als die drei „Vs“ beschrieben. Je nach Anwendung kommen weitere „Vs“ hinzu.*

Teilweise werden weitere Aspekte hinzugefügt und mit einem „V“ beschrieben, beispielsweise die folgenden: *Veracity* steht für Vertrauenswürdigkeit der Daten oder der gezogenen Schlüsse. *Value* betont, dass Big Data letztendlich immer eine Wertschöpfung der Daten beabsichtigt. *Visualization* stellt die intuitive Darstellung der Ergebnisse heraus.

Je nach Anwendung treffen die oben genannten Eigenschaften mehr oder weniger zu. Gemeinsam ist allen Big-Data-Lösungen letztendlich, dass durch eine Verarbeitung von Daten neue Zusammenhänge erkannt und so neue Erkenntnisse gewonnen werden sollen. Dies ist zwar schon lange ein Ziel der Informatik, durch die im Folgenden kurz vorgestellten technischen Fortschritte ist eine Umsetzung allerdings deutlich einfacher geworden. Wir unterscheiden zwischen Technologien für die Datenhaltung, die verteilte Berechnung und die analytische Verarbeitung.

3.1 Datenhaltung

Technologien zur Datenhaltung orientieren sich an ihren Anforderungen, dem Zweck der Haltung und den Formaten, in denen die Daten vorliegen. Die folgenden zwei Technologien sind klassische Beispiele für die Datenhaltung im Big-Data-Kontext.

In-Memory-Technologien: In-Memory Analytics verfolgt die Idee, die gesamte Datenbasis während der Verarbeitung im Hauptspeicher (RAM) vorzuhalten, um nicht auf langsame Speichermedien wie Festplatten zugreifen zu müssen. Dies wurde in der jüngeren Vergangenheit durch sinkende Kosten bei Hauptspeichermodulen und durch die Verbreitung von 64-Bit-Systemen vorangetrieben, die zum Adressieren entsprechend großer Speicher nötig sind.

Auch sehr große Datenbanken werden heute vollständig im Hauptspeicher gehalten.

Eine Basistechnologie stellen In-Memory-Datenbanken dar. Diese halten die Daten im Hauptspeicher eines oder mehrerer Computer vor, was im Gegensatz zur bisher üblichen Speicherung auf Festplattenlaufwerken steht. Dadurch wird eine deutliche Geschwindigkeitssteigerung von Schreib- und Lesevorgängen erreicht. Bekannte In-Memory-Lösungen sind beispielsweise SAP Hana oder Terracotta von der Software AG.

NoSQL-Datenbanken: Datenbanken sind in der Regel stark strukturiert. Sie enthalten Datensätze, die

jeweils identisch aufgebaut sind. Eine Datenbank, die Adressen verwaltet, hat beispielsweise für jeden Datensatz ein Feld für den Namen, für den Vornamen, für die Postleitzahl und weitere ähnliche Felder. In vielen Big-Data-Anwendungsfällen kann jedoch nicht von solchen strukturierten Daten ausgegangen werden. Tatsächlich ist es eine der Stärken von Big Data, nicht nur auf strukturierte Daten angewiesen zu sein. Daher sind inzwischen immer mehr Datenbank-Konzepte entstanden, die Daten unabhängig von einer fest vorgegebenen Struktur speichern können. Diese werden als NoSQL-Datenbanken bezeichnet, um sie von den herkömmlichen strukturierten SQL-Datenbanken abzugrenzen.

NoSQL-Datenbanken stellen somit einen Mittelweg dar, zwischen einem Datensystem, auf welchem beliebige Daten gespeichert werden können, und einer SQL-Datenbank, in der nur strukturierte Daten abgelegt werden können. Gleichzeitig bieten sie eine stärkere Strukturierung als ein Dateisystem, was den Zugriff auf die Daten erleichtert und beschleunigt.

3.2 Verteiltes Rechnen

Um Big-Data-Prozesse in Echtzeit realisieren zu können, wird oft eine Rechenleistung benötigt, die ein einzelner Computer nicht zur Verfügung stellen kann. Die Rechenlast kann jedoch auf eine (möglicherweise große) Anzahl von einzelnen Rechnern aufgeteilt werden. Jeder einzelne Rechner nutzt seine Ressourcen, um seine zugewiesene Teilaufgabe zu lösen. Sind alle Teilaufgaben gelöst, werden diese anschließend zusammengeführt, um die Gesamtaufgabe abzuschließen. Dies stellt die allgemeine Idee beim verteilten Rechnen dar.

Komplexe Aufgaben werden bei Big Data automatisch auf mehrere Rechner verteilt.

Ein bekanntes Beispiel für verteiltes Rechnen ist MapReduce, welches von Google im Jahr 2004 eingeführt wurde. Hierbei handelt es sich um ein äußerst effizientes Programmiermodell der verteilten Berechnung. Heutzutage setzten viele führende Firmen MapReduce-Technologien (z. B. Apache Hadoop) ein, um Big-Data-Lösungen umzusetzen. Dadurch werden keine speziellen und teuren High-End-Computer benötigt, da handelsübliche Standard-Rechner für die Bearbeitung der kleineren Teilaufgaben meist ausreichen. Dies wiederum erspart zusätzliche Hardware-Kosten. Firmen können oftmals auf ausgemusterte Rechner zurückgreifen, um die für MapReduce benötigten Rechnernetze aufzubauen.

3.3 Analytische Verarbeitung

Die Methoden zur Gewinnung von Erkenntnissen aus Daten lassen sich als maschinelles Lernen und Data-Mining zusammenfassen.

Unter **maschinellem Lernen** (ML) werden verschiedene algorithmische Verfahren verstanden, welche unter anderem Zusammenhänge in Daten herausarbeiten, mittels derer anschließend weitere Daten bearbeitet werden können. Erst werden Zusammenhänge und Regeln aus bekannten Daten „gelernt“, um anschließend neue (unbekannte) Daten mit den Verfahren zu verarbeiten. Wird beispielsweise ein ML-Verfahren mit Texten trainiert, die als deutsch oder englisch gekennzeichnet sind, kann das Verfahren später anhand einfacher Buchstabenfolgen selbstständig einen neuen Text als deutsch oder englisch erkennen.

Maschinelles Lernen stellt das künstliche Generieren von Wissen aus Daten dar.

ML ist ein integraler Bestandteil zahlreicher Anwendungen wie beispielsweise Spracherkennung, (Hand-) Schrifterkennung, Kundensegmentierung, Stimmungsanalyse oder Betrugserkennung (z. B. Kreditkartenmissbrauch). Weiterhin dienen ML-Verfahren in der Industrie auch oftmals zur Prozessoptimierung. Aus technischer Sicht ist ML eine Zusammensetzung zahlreicher Lernverfahren, welche sich in folgende Kernbereiche einteilen lassen:

Die **Klassifikation** hat das Ziel (ähnliche) Objekte zu einem Oberbegriff zusammenzufassen. Ein einfaches Beispiel sind E-Mails. Sie können z. B. in relevant/irrelevant, bzw. Spam/Nicht-Spam klassifiziert werden. Als Ergebnis stehen oftmals miteinander verknüpfte Wahrscheinlichkeiten oder andere statistische Maße, für deren Interpretation zumeist weitere Verfahren nötig sind.

Unter **Clustering** werden Lernverfahren verstanden, die ohne Hintergrundwissen versuchen, Daten auf unterschiedliche Art und Weise zu gruppieren. Eine solche Gruppierung basiert in der Regel auf strukturellen Ähnlichkeiten zwischen den Daten. So werden beispielsweise Texte im Nachrichten-Kontext zu unterschiedlichen Gruppen (z. B. „Politik“, „Wirtschaft“, „Kultur“) zusammengefasst.

Beim **Regellernen** geht es darum, aus expliziten Merkmalen implizite Regeln bzw. Zusammenhänge zu „erlernen“. Diese wiederum können z. B. dafür genutzt werden, unbekannte Daten zu klassifizieren oder auch automatisierte Schlussfolgerungen zu ermöglichen. So kann beispielsweise aus Einkäufen abgeleitet werden, welche Ware X sich zu einer bestimmten Zeit in Kombi-

nation mit Ware Y besonders gut verkaufen lässt. Diese Information kann dann einem Ladenbetreiber helfen diese Waren entsprechend nebeneinander zu platzieren.

Bei der **Mustererkennung** werden Muster aus bekannten Daten extrahiert, um diese anschließend auf neue (ungesehene) Daten anzuwenden und entsprechend zu bewerten bzw. zu klassifizieren. Dabei wird zwischen expliziten und impliziten Muster unterschieden. Explizite Muster in Texten können beispielsweise Füllwörter, Satzzeichen oder Wortfragmente sein, mit deren Hilfe Autoren anonymen Texten zugeordnet werden können. Implizite Muster dagegen sind solche Muster, für deren Erkennung und Herleitung erst die zugrunde liegenden Daten in eine Zwischenstufe überführt werden müssen. Letztere sind insbesondere im Rahmen von Big Data wichtig, um etwa Querbezüge zwischen heterogenen Daten herzustellen. Beispielweise lassen sich, je nachdem welche Art der Visualisierung gewählt wurde, unterschiedliche Erkenntnisse gewinnen und Schlüsse ziehen und folglich unterschiedliche Muster aufdecken. Für das Beispiel von Textdaten bedeutet dies, dass aus unterschiedlichen Visualisierungen, z. B. die Auftrittswahrscheinlichkeiten unterschiedlicher Wörter oder Phrasen, unterschiedliche implizite Muster hergeleitet werden können.

Data-Mining („Daten-Bergbau“) bezeichnet das intelligente, größtenteils automatisierte Finden und Erkennen von relevanten Mustern in großen Datenmengen. Data-Mining ist eng mit ML verwandt; oft dienen Data-Mining-Verfahren als Voranalyse für maschinelles Lernen. Als gemeinsamer Nenner verstehen sich hierbei Konzepte und Verfahren, mit denen Datensätze unter anderem selbstständig klassifiziert oder hinsichtlich ihrer Ähnlichkeit gruppiert („geclustert“) werden. Im Gegensatz zu ML, wo ein Prozess ohne die Interaktion eines Menschen verläuft, ist bei Data-Mining oftmals der Mensch in den Prozess involviert, insbesondere wenn die Erkenntnisse visualisiert und ausgewertet werden sollen. Ein weiterer Unterschied zu ML ist die allgemein ergebnisoffene Zielsetzung bei Data-Mining, wohingegen bei ML meist die Art der Problemlösung im Fokus steht.

4 Implikationen für die Privatheit

Big Data führt zu einer neuen Qualität der Datenverarbeitung und somit zu neuen Chancen und Möglichkeiten in unterschiedlichen Bereichen. An dieser Stelle soll nun betrachtet werden, welche Auswirkung diese Technologie auf die Privatheit hat.

Neue Technologien führen oft zu einem Spannungsfeld zwischen dem technisch Möglichen und dem ethisch Vertretbaren. Die Gesellschaft muss sich erst über die Konsequenzen der Technologie im Klaren werden und dann Regeln für den Umgang mit ihr finden. Ein Beispiel dafür aus der Vergangenheit ist die Situation des Urheberrechts im Internet. Als um das Jahr 2000 herum Dienste wie Napster Musik in Form von MP3-Dateien plötzlich frei verteilbar und so kostenfrei verfügbar machten, begann eine noch heute andauernde Diskussion um eine gerechte Wahrung verschiedener Interessen sowie deren technische und rechtliche Konsequenzen.

Auch Big Data führt zu neuen Herausforderungen im Umgang mit Daten. Konzepte, die ursprünglich als ausreichend zum Schutz der Privatsphäre betrachtet wurden, weichen auf, weil immer mehr Daten miteinander verknüpft werden können. Große Mengen unterschiedlicher Daten werden zusammengefügt, um neue Methoden der Wertschöpfung zu realisieren, ohne dabei von Anfang an auch Aspekte des Datenschutzes zu berücksichtigen.

Die Hälfte der Unternehmen sieht Datenschutz als Hindernis.

Interessant ist hier auch die Sichtweise der Industrie: In einer Umfrage des BITKOM vom Februar 2014 [4] wurde festgestellt, dass etwas mehr als die Hälfte aller befragten Unternehmen den Datenschutz als Hindernis für den Einsatz von Big Data sieht. Ähnlich wurden von den Unternehmen auch die Hürden durch die Anforderungen an die IT-Sicherheit gesehen – hier war es knapp die Hälfte der Unternehmen. In immerhin 17 Prozent der Unternehmen sind keine Prozesse für den Umgang mit personenbezogenen Daten festgelegt. Der BITKOM hat für die Studie 507 Unternehmen mit mindestens 50 Mitarbeitern befragt.

Letztendlich kann Big Data als ein typisches Dual-Use-Phänomen gesehen werden: Die Technologie bringt sowohl Chancen als auch Risiken mit sich. Nur wenn eine konkrete Anwendung diskutiert wird, kann hier eine Aussage getroffen werden, wie viel Chance und wie viel Risiko vorliegt. Andere Technologien, für die dies gilt, sind Filtertechnologien, die sowohl zur Spam-Bekämpfung als auch zur Zensur verwendet werden können, oder Überwachungssysteme, die sowohl zur Verbrechensbekämpfung als auch zum Ausspähen von Bürgern eingesetzt werden können. Selbst Kryptografie wird kontrovers diskutiert: Die einen sehen in ihr die einzige Chance auf Privatheit bei der Kommunikation, die anderen eine Möglichkeit für Verbrecher, sich ungestört miteinander auszutauschen.

Abbildung 12: *Die Aktivitäten von Google haben die Firma zu einem beliebten Beispiel für einen Datenkraken gemacht. Ähnlich werden aber auch Facebook, die NSA oder die Schufa gesehen. Quelle: PC Magazin* `http://www.pc-magazin.de/ratgeber/google-und-der-datenschutz-86503.html`

Es ist abzusehen, dass der Nutzen von Big Data eine kontinuierliche Weiterentwicklung der zugrunde liegenden Technologien begünstigen wird. Mit gesteigerter Leistungsfähigkeit der Technologien werden jedoch auch die damit verbundenen Risiken steigen. Um diesen Risiken entgegenzuwirken, sind sowohl rechtliche als auch technische Aspekte bekannt, die in den folgenden Abschnitten aufgezeigt werden. Sie geben dem Bürger das Recht, sich gegen einen vermuteten Missbrauch personenbezogener Daten zu wehren. Und sie erlauben es Betreibern von Big-Data-Lösungen, einen Kompromiss zwischen Chancenoptimierung und Risikominimierung zu finden.

Der Erfolg von Big Data wird dazu führen, dass die Technologie immer leistungsfähiger wird.

4.1 Profilbildung anhand der Verschmelzung von Google-Diensten

Google hat im März 2012 viele seiner Dienste (Gmail, YouTube, Google+, etc.) zusammengelegt [16], um alle erhobenen Daten eines Nutzers zu einem Profil kombinieren zu können. Ein solches Profil ermöglicht Google seine Nutzer genauer als bisher zu beschreiben, da hier Privates (z. B. YouTube-Kommentare) und Geschäftliches (z. B. E-Mail-Verkehr über Gmail) vermischt wird und dadurch bisher technologisch belegte Grenzen überwunden werden. Mehr noch, um Kommentare auf YouTube oder innerhalb von Google Play schreiben zu können, benötigt ein Nutzer ein verknüpftes Google+ Konto [9]. Auch für Android, das von Google vertriebene Betriebsystem für Smartphones und Tablets, ist ein solches Konto für viele Anwendungen und Funktionen erforderlich. Dies führt dazu, dass Google ein sehr viel umfassenderes Bild von seinen Nutzern zusammenstellen kann und von Datenschützern als „Datenkrake“ (siehe Abbildung 12) bezeichnet wird: Bereits die Adressaten von E-Mails ermöglichen das Aufspannen eines sozialen Netzes (Gmail). Google-Maps-Abfragen ergeben ein Bewegungsprofil. YouTube verrät viel über private Interessen, wie beispielsweise Musikgeschmack. Besonders aufschlussreich sind allerdings Suchanfragen. Gelingt es, diese einem Benutzerprofil zuzuordnen, lässt sich viel über aktuelle Themen, die den Nutzer beschäftigen, ableiten. Hinzu kommt das Surf-Verhalten des Nutzers, das großflächig über Dienste wie Google Analytics (Nutzeranalyse, die auf ca. der Hälfte aller populären Webseiten genutzt wird) und Google AdSense (Werbemodul, das auf vielen Webseiten vorhanden ist) erfasst werden kann.

4.2 Rechtliche Grundlagen

Big Data ist ein Ansatz, dessen Umsetzung große Mengen von Daten erfordert. Hinsichtlich des Datenschutzes ist zu unterscheiden, ob diese Daten personenbezogen sind oder nicht. Personenbezogene Daten sind alle Daten, die auf eine bestimmbare Person hinweisen oder ihr zugeordnet sind. Einfache Beispiele sind körperliche Merkmale der Person, aber auch ihre Telefonnummer oder ihr Wohnort. Nicht personenbezogene Daten sind Daten, für die (auch in Zukunft) keine Zuordnung zu handelnden oder betroffenen Personen möglich ist. Das gilt u. a. für Daten, die sich ausschließlich auf Geräte und Produkte, nicht aber auf ihre Nutzer beziehen, z. B. Sensordaten zur Ortung von Transportgütern in der automatisierten Logistik. Die Schwierigkeit einer klaren Einteilung in personenbezogene und nicht personenbezogene Daten lässt sich auch daran erkennen, dass

manche Daten zwar zunächst nicht personenbezogen sind, aber durch das Zusammenfügen mit anderen Daten im Rahmen von Big Data dann als personenbezogen einzustufen sind. Dazu zählen u. a. (Sensor-)Daten, die zwar direkt von Maschinen erzeugt werden, aber einen direkten Personenbezug vorweisen (z. B. Assistenzsysteme im Gesundheitswesen). Es sind allerdings nicht die persönlichen Daten von Interesse, sondern die Daten der zugehörigen Gruppe.

Der Bürger darf bestimmen, welche Informationen über ihn zur Verfügung stehen.

Eine bedeutende Rolle in diesem Konflikt kommt dem Grundrecht auf informationelle Selbstbestimmung zu (BVerfGE 65,1). So ist festgelegt, dass der Bürger selbst bestimmen darf, welche Information über ihn zu welcher Zeit zur Verfügung stehen darf. In der Praxis ist dieses Recht auf Daten mit direktem Personenbezug beschränkt. Ein wesentlicher Aspekt bei Big Data ist jedoch, dass es oftmals nicht um persönliche Daten geht, sondern dass das Interesse stattdessen den Daten der zugehörigen Gruppe gilt. Es gibt hier einen Übergang von einer individuellen Selbstbestimmung („Was passiert mit *meinen* Daten?“) zu einer gesellschaftlichen Selbstbestimmung („Was passiert mit *unseren* Daten?“), und damit auch zu neuen Formen der Wahrnehmung dieser Selbstbestimmung. Die große Herausforderung in diesem Sinne ist, entsprechend zu differenzieren und gesetzliche Verbote bestimmter Verarbeitungen bspw. mithilfe von Ethikkommissionen auszusprechen.

Die folgenden rechtlichen Grundlagen stellen vereinfacht dar, welche Regeln Big Data bei der Verarbeitung von personenbezogenen Daten beachten sollte. Da es sich hier um eine Technologie handelt, die neu und im Wandel begriffen ist, gehen die Meinungen und deren Interpretation, wie weit diese Regeln umgesetzt werden können und müssen, in der Praxis auseinander.

Für personenbezogene Daten gelten Datensparsamkeit, Zweckbindung, Einwilligung, Auskunfts- und Eingriffsrecht.

Zusammengefasst sind es vor allem die folgenden Prinzipien des Bundesdatenschutzgesetzes (BDSG), welche häufig in Zusammenhang mit Big Data diskutiert werden: Datensparsamkeit, Zweckbindung, Einwilligung und Auskunftsrecht sowie Eingriffsrecht.

Die **Datensparsamkeit** erweckt schon dem Namen nach den Eindruck, nur schwer mit Big Data vereinbar zu sein. Das BDSG schreibt vor, dass bei der Verarbeitung personenbezogener Daten so wenige Daten wie möglich gesammelt, gespeichert und genutzt werden sollen. Dies soll nach Möglichkeit auch anonymisiert oder pseudonymisiert geschehen, wenn der Aufwand dazu nicht unverhältnismäßig hoch ist. Das Vorgehen bei Big Data hingegen ist oft, erst einmal eine möglichst große Menge an Daten zu sammeln und dann zu analysieren, welche dieser Daten sich wie in Beziehung setzen lassen, um neue Erkenntnisse zu gewinnen. Zum Zeitpunkt des Sammelns ist folglich der genaue Zweck noch unbestimmt; es kann nicht entschieden werden, welche Daten notwendig sind und welche verworfen werden können.

In diesem Sinne eng verbunden mit der Datensparsamkeit ist die **Zweckbindung**: Personenbezogene Daten, die für einen Zweck erhoben werden, dürfen nicht ohne Weiteres für einen anderen Zweck verwendet werden. Das bedeutet, dass ein Unternehmen, welches personenbezogene Daten völlig korrekt unter Beachtung des Datenschutzes beispielsweise zum Versenden von Verbraucherinformationen erhoben hat, diese nicht ohne Weiteres zur Produktoptimierung einsetzen kann. Es wird entweder eine gesetzliche Erlaubnis benötigt oder aber die Einwilligung des Betroffenen. Die Zweckbindung gebietet auch, dass Daten nur erhoben werden dürfen, wenn ihr Zweck bereits klar definiert ist. Im Falle von Big Data kann dies bedeuten, dass umfangreiche Daten erneut erhoben werden müssen, wenn sie zu einem Zweck ungleich ihrem ursprünglichen verwendet werden sollen.

Um personenbezogene Daten erheben zu dürfen, bedarf es nach dem BDSG entweder einer gesetzlichen Erlaubnis oder einer **Einwilligung** durch den betroffenen Bürger („Verbot mit Erlaubnisvorbehalt“). Diese ist nur wirksam, wenn sie auf der freien Entscheidung des Betroffenen beruht. Er ist auf den vorgesehenen konkreten Zweck der Erhebung, Verarbeitung oder Nutzung sowie, soweit nach den Umständen des Einzelfalles erforderlich oder auf Verlangen, auf die Folgen der Verweigerung der Einwilligung hinzuweisen. Die Einwilligung bedarf der Schriftform, soweit nicht wegen besonderer Umstände eine andere Form angemessen ist. Solche Umstände sind beispielsweise gegeben, wenn eine hohe Dringlichkeit in einem Not- oder Krankheitsfall besteht. Hier genügt eine mündliche Einwilligung. Auch wenn die Daten direkt bei der Erfassung anonymisiert werden, reicht dies aus. Soll die Einwilligung zusammen mit anderen Erklärungen schriftlich erteilt werden, ist sie besonders hervorzuheben.

Auch nachdem personenbezogene Daten erhoben wurden, hat der Bürger Rechte. Das **Auskunftsrecht** besagt, dass der Bürger das Recht hat, zu erfahren, welche Daten über ihn gespeichert werden und wozu. Auf Verlangen muss die verantwortliche Stelle ihm Auskunft

erteilen über Herkunft und Art der gespeicherten Daten, den Empfängern dieser Daten und den Zweck der Speicherung. Eine entsprechende Anfrage kann jährlich und kostenfrei angefordert werden. Allerdings kann eine solche Auskunft eingeschränkt werden, wenn die Wahrung von Geschäftsinteressen der Daten erhebenden Instanz wichtiger als die Auskunftspflicht angesehen wird. An dieser Stelle gehen die Ziele von Big Data und die Privatheit des Einzelnen wieder weit auseinander. Es ist fraglich, ob ein einzelner Bürger hier seine Interessen gegen einen (multinationalen) Konzern durchsetzen kann.

Die Ziele von Big Data und Privatheit gehen oft weit auseinander.

Der Bürger kann auch aktiv gegen über ihn gespeicherte Daten vorgehen: Die **Eingriffsrechte** geben ihm das Recht, falsche Daten berichtigen und bestimmte Daten löschen bzw. sperren zu lassen. Dies gilt beispielsweise für „Daten über die rassische oder ethnische Herkunft, politische Meinungen, religiöse oder philosophische Überzeugungen, Gewerkschaftszugehörigkeit, Gesundheit, Sexualleben, strafbare Handlungen oder Ordnungswidrigkeiten“, sofern „ihre Richtigkeit von der verantwortlichen Stelle nicht bewiesen werden kann“ (§ 35 Abs. 2 Nr. 2 BDSG). Ähnlich zum Thema Auskunft bedeutet dies für Big-Data-Anwender, dass sie theoretisch jederzeit über die Daten derart verfügen können müssen, dass eine Korrektur, Sperrung oder Löschung ohne Weiteres möglich ist. Auch hier kann der Sammelnde widersprechen, beispielsweise wenn eine Löschung nur mit unverhältnismäßig hohem Aufwand geschehen kann. Die Verhältnismäßigkeit muss dann wieder individuell geklärt werden.

Neben dem BDSG wird auch die Umsetzung der aktuell zur Diskussion stehenden Datenschutz-Grundverordnung der Europäischen Union [7] Einfluss auf die Ausrichtung von Big Data haben. Diese Verordnung sieht u. a. vor, dass Profiling ausdrücklich unter den Einwilligungsvorbehalt des Betroffenen gestellt werden soll. Erzielen Europäischer Rat, Europäisches Parlament und Europäische Kommission eine Einigung, ist die Grundverordnung rechtsverbindlich und würde das BDSG ablösen.

In den USA fehlt eine unabhängige Datenschutzaufsicht.

In diesem Zusammenhang wird auch interessant werden, wie außereuropäische Staaten die Datenschutz-Grundverordnung auffassen werden. Insbesondere in den USA, wo viele weltweit agierende IT-Unternehmen ansässig sind, gibt es keine umfassende unabhängige Datenschutzaufsicht, die das Recht auf Privatsphäre vertritt. Zwar gibt es „Invasion of Privacy“ als rechtlich definierten Klagegrund oder ein konstitutionell zugesichertes Recht auf Privatheit gegenüber regierungsabhängigen Institutionen, jedoch bezieht sich dieses eher auf z. B. Privatheit im eigenen Haus und weniger auf digitale Daten. Der Zugriff auf private Daten ist in vielen Fällen gesellschaftlich akzeptiert, z. B. eine Bonitätsprüfung vor der Vereinbarung eines Arbeitsverhältnisses oder vor der Anmietung einer Wohnung.

Datenschutzregelungen gibt es nur in einzelnen Teilbereichen wie den *Children's Online Privacy Protection Act* (COPPA) und im Bereich der Krankenversicherungen den *Health Insurance Portability and Accountability Act* (HIPAA). Eine landesweit gültige Regelung für den allgemeinen Umgang mit persönlichen Daten existiert jedoch nicht. Viele Gesetzesentwürfe und -Vorschläge der letzten drei Jahre zum Thema Privatsphäre wurden allesamt abgelehnt. Die einzelnen Bundesstaaten agieren dahingehend weiterhin relativ autark.

Eine besondere Rolle kommt dem *USA Patriot Act* zu. Dieses landesweite Gesetz sichert den US-amerikanischen Behörden, insbesondere FBI, CIA und NSA, weitreichende Rechte zu und setzt kollidierende Gesetze einzelner Bundesstaaten außer Kraft. Als Reaktion auf die Anschläge des 11. Septembers 2001 verabschiedet, hat es zum Ziel, den internationalen Terrorismus zu bekämpfen. Die weitreichenden Auswirkungen auf die Privatsphäre wurden bereits in Abschnitt 2.5 behandelt.

4.3 Technische Schutzmaßnahmen

Um bei personenbezogenen Daten datenschutzrechtliche Bestimmungen umzusetzen, existiert eine Reihe von technischen Lösungen. Dabei ist zu unterscheiden

Abbildung 13: *Die Möglichkeiten zum Schutz der Privatsphäre sind meist sehr technisch*

zwischen Lösungen, die Dienstanbieter als Datensammler zum datenschutzfreundlichen Umgang mit Big Data einsetzen, und Lösungen, die Nutzer einsetzen können, um sich vor der Hergabe zu vieler personenbezogener Daten zu schützen (Selbstdatenschutz).

Zum datenschutzfreundlichen Umgang mit Big Data sollten Anbieter wie auch bei anderen Anwendungen, die sensitive Daten verarbeiten, diese sowohl **verschlüsselt** abspeichern als auch übertragen, um ein Ausspähen der Daten durch Dritte zu erschweren. Da Big-Data-Lösungen oft von mehreren Anwendern parallel genutzt werden, ist es wichtig, dass die Anwender gegenseitig **abgeschottet** sind. So können Anwender nicht gegenseitig ihre Daten in einem laufenden Verarbeitungsprozess einsehen.

Auch zur Verarbeitung der Daten selbst gibt es datenschutzfreundliche Sicherheitsansätze, die unter dem Begriff *Privacy-Preserving Data Mining* zusammengefasst werden. Beim **Anonymisieren** werden identifizierende Merkmale aus den Datensätzen gelöscht. Dieser Vorgang soll nach BDSG § 3 Abs. 6 nicht oder nur mit unverhältnismäßig hohem Aufwand umkehrbar sein. Oft bestehen an diese Anonymität bestimmte Vorgaben, die beschreiben, wie groß eine Gruppe von Personen mindestens sein muss, auf die mittels der vorhandenen Daten eingegrenzt werden kann. Hier spricht man von k-Anonymität (k-Anonymity) [21], wobei k die Größe der nicht unterscheidbaren Personengruppe bestimmt. Nimmt man als Beispiel ein Wohnhaus mit fünf Wohnungen, in denen zusammen 13 Personen wohnen, dann würde eine Adresse mit Straßennamen und Hausnummer nur die Menge der 13 Personen beschreiben, aber keine genauere Eingrenzung ermöglichen. An einem einfachen und unverfänglichen Beispiel erklärt: Möchte ein Lieferservice für Pizza öffentlich darstellen, wohin er in einem Ort welche Pizzen liefert, könnte er die Bestellungen anonymisiert veröffentlichen, indem er aus ihnen Namen, Telefonnummer und Wohnungsnummer löscht, wenn eine Anonymität von 13 Personen ausreicht. Weitere Konzepte für Privacy-Preserving Data Mining sind l-Diversity, t-Closeness und Differential Privacy.

Beim **Pseudonymisieren** werden die Namen oder andere identifizierende Merkmale nicht einfach gelöscht, sondern durch ein Pseudonym ersetzt. Wer dieses Pseudonym kennt, kann den zur Person gehörenden Datensatz weiterhin identifizieren. Ein anderer Weg ist die **Datenaggregation**: Hier werden mehrere Datensätze zusammengefasst. So könnte für das Beispiel oben der durchschnittliche Verbrauch von Trinkwasser für das Gebäude gespeichert werden, statt diesen pro Familie auszuweisen.

Big Data erleichtert potentiell das Aufheben von Anonymität.

Eine wichtige Beobachtung bei den technischen Maßnahmen zur Sicherstellung der Privatheit ist, dass der oben genannte unverhältnismäßige Aufwand, der nach dem BDSG als Grenze der Umkehrbarkeit von Anonymität gilt, durch Big Data relativiert werden könnte. Denn die Verfügbarkeit von Big-Data-Verfahren, die komplexe Zusammenhänge viel effizienter ableiten können, kann in der Praxis zu höheren notwendigen Hürden bei der Umkehrbarkeit führen.

Auf der anderen Seite gibt es Lösungen für den Selbstdatenschutz von Bürgern. Allgemein unterscheidet man zwischen Tools zur Verschlüsselung, Tools zur Durchsetzung von Anonymität und Pseudonymität, Filter-Tools, Policy-Tools und Tools zum Rechtemanagement bei mobilen Apps [19]. Konkrete Beispiele zum Selbstdatenschutz mit Handreichungen zu ihrer Nutzung finden sich etwa auf der Webseite des Landesdatenschutzbeauftragten von Rheinland-Pfalz (`http://www.datenschutz.rlp.de/de/selbstds.php`). Das *Forum Privatheit* vermittelt Hintergrundwissen und praktische Informationen zum Selbstdatenschutz [12].

5 Profiling und Scoring

Werden personenbezogene Daten verarbeitet, um eine Person zu beschreiben, zu bewerten oder Prognosen über sie zu erstellen, spricht man von Profiling. Der Begriff wird schon lange verwendet, beispielsweise in der Kriminalistik, die das Erstellen von Täterprofilen kennt. Im Kontext von Big Data wird er benutzt, um automatisierte Verfahren zu beschreiben, die aus großen Mengen personenbezogener Daten aus oft unterschiedlichen Quellen Profile ableiten. Zwei verbreitete Ausprägungen von Profiling sind Scoring und Personalizing. Beim Scoring wird angestrebt, personenbezogene Daten auf einen Wert (Score) zu projizieren, der einen einfachen Vergleich mit anderen Personen ermöglicht. Beim Personalizing wird auf die Abstraktion durch einen Wert verzichtet. Hier werden anhand der Datenlage die Person betreffende Fragen beantwortet.

5.1 Ausprägungen

Scoring wird heute bereits in einer Vielzahl von Ausprägungen angewandt. Die folgenden Beispiele sollen einen kleinen Überblick über die Durchdringung von Scoring im Alltag aufzeigen.

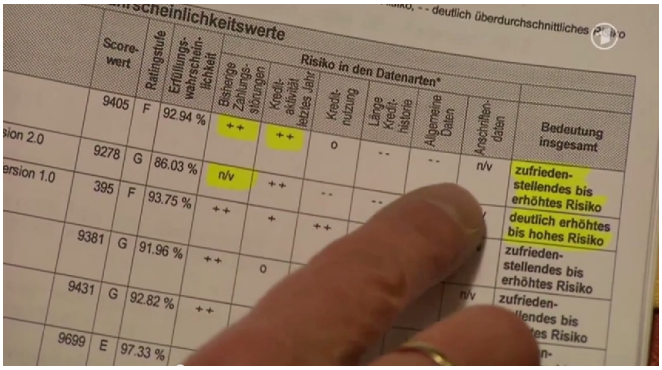

Abbildung 14: *Bonitätsauskunft der Schufa. Auskunfteien wie die Schufa sind intensive Nutzer von Big-Data-Technologien. Quelle: YouTube* `https://www.youtube.com/watch?v=xQHxbSkkpxY`

Kreditwürdigkeit: Ob und zu welchen Konditionen einem Kunden ein Kredit gewährt wird (z. B. Laufzeit, Zinsen) hängt insbesondere davon ab, wie gut sein von Auskunfteien errechneter Score ist (siehe Abbildung 14). Dieses Beispiel wurde bereits in Abschnitt 2.7 vertieft.

Versicherungen: Ähnlich dem Beispiel Kreditwürdigkeit, handeln Versicherungen Verträge anhand von Scores aus. Anhand dieses Wertes entscheidet sich, ob und zu welchen Konditionen ein Kunde eine Versicherung erhält.

Gezielte Werbung: Werbe-Scoring wird verwendet, um zukünftiges Kauf- und Konsumverhalten vorherzusagen. Anhand des Scores kann entschieden werden, welche Werbung welchem Kunden aufgezeigt wird, und auf welche Art und Weise er sie erhält.

Bewerberbewertung: Vor allem große Unternehmen holen im Vorfeld von Personalentscheidungen Daten über Bewerber ein und führen ein Scoring durch. Anhand des Scores wird ein Bewerber bewertet und potenziell aussortiert.

Personalmanagement: Analog zum Beispiel Bewerberbewertung führen manche große Unternehmen ein Scoring ihrer Mitarbeiter durch (z. B. Zigarettenpausen, Toilettenbesuche, Privattelefonie, Smartphonenutzung). Sind tiefgreifende Personalentscheidung wie Kündigungen oder Vertragsverlängerungen zu treffen, unterstützt der entsprechende Score die zu fällende Entscheidung.

Terrorbekämpfung: Insbesondere durch das groß angelegte Sammeln und Auswerten von Daten durch Geheimdienste (siehe Abschnitt 2.5) ist das Beispiel Terrorbekämpfung bekannt. Personen werden als potenzielle Terroristen anhand des aus den erhobenen Daten errechneten Scores klassifiziert und entsprechend behandelt.

Kriminalitätsaufklärung: Profiling kommt auch zum Einsatz, wenn es um Aufklärung von Straftaten geht. Die Polizei speichert potenziell wichtige Daten von Straftätern und erstellt entsprechende (Täter-)Profile. Nach einer Straftat wird anhand der Profile der Kreis potenzieller Täter eingeschränkt und die Aufklärung beschleunigt.

5.2 Kritik

Die Möglichkeit, Personen anhand über sie verfügbarer Daten zu analysieren oder Prognosen über sie zu erstellen, ist weit verbreitet. Insbesondere Datenschutzbeauftragte warnen immer wieder über die Risiken und stellen auch die rechtliche Grundlage des Profilings in Frage. So stellte bereits im Jahr 2005 der damalige Bundesbeauftragte für den Datenschutz und die Informationsfreiheit, Peter Schaar, die Risiken von Profiling und Scoring heraus [20]. Ein angesprochenes Risiko liegt darin, dass Daten über lange Zeit hinweg die Beurteilung von Personen beeinflussen und so Fehler, die in der fernen Vergangenheit liegen, noch lange einen deutlichen Einfluss auf die Chancen dieser Person haben:

„Es darf nicht dazu kommen, dass z. B. ein junger Mensch, der im Alter von zwanzig Jahren auch nach einer Mahnung seine Handyrechnung nicht bezahlen konnte, anschließend kein Konto mehr eröffnen kann, keine Wohnung findet, keinen Versicherungsvertrag

Abbildung 15: *Kritik an Scoring: Eine Person wird durch Analyse von beispielsweise Beziehung, Finanzen, Hobbys, Gesundheit, Medienkonsum und Ernährung auf eine Zahl reduziert.*

bekommt und selbst der Zahnersatz nur gegen Vorkasse gewährt wird [...].“

Abstrakter formuliert der Landesbeauftragte für den Datenschutz Schleswig-Holstein Thilo Weichert in einem DuD-Artikel [23] das Risiko:

„Die Gefahren des Kredit-Scoring für den Konsumenten bestehen darin, dass über die Zuordnung von Erfahrungswerten aus Verträgen mit anderen Konsumenten Schlüsse gezogen werden, die dem jeweiligen gescorten Konsumenten nicht gerecht werden, weil individuelle Umstände nicht oder falsch in die Bewertung einbezogen werden.“

Seine Kritik beschreibt in erster Linie die Gefahr, dass allgemeine Aussagen über bestimmte Zusammenhänge aus den gesammelten Daten getroffen werden. Diese Zusammenhänge führen dann für eine individuelle Person zu Nachteilen, die an Diskriminierung grenzen. So kann eine Entscheidung über eine Kreditwürdigkeit von Wohnort oder abonnierten Zeitungen abhängen, was für den Betroffen zum einen nicht transparent ist, zum anderen aber auch im konkreten Fall durch andere, nicht beachtete Faktoren entkräftet werden könnte.

Oft mangelt es beim Scoring bezüglich der Vorgehensweise an Transparenz.

Ein grundsätzlicher Kritikpunkt am Scoring ist, dass hier Daten über Personen auf eine für den Betroffenen intransparente Weise und in der Regel ohne ihre Kenntnis zusammengeführt werden, um eine Bewertung dieser Person durchzuführen. Auch wenn bekannt ist, um welche Daten es sich handelt, ist oft nicht bekannt und nachvollziehbar, wie diese Daten miteinander in Bezug gesetzt werden, um die Person im Vergleich zu den Referenzdaten anderer Personen zu bewerten. Zum einen gelten die entsprechenden Verfahren als Betriebsgeheimnisse der bewertenden Unternehmen, zum anderen liefern Big-Data-Verfahren wie bereits erwähnt oft auch Antworten ohne einen unmittelbar nachvollziehbaren Weg dahin. Das BDSG fordert hierzu beim Scoring zwar wissenschaftlich nachvollziehbare Vorgehensweisen, eine tatsächliche Prüfung, ob sich an die Datenschutzrichtlinien gehalten wurde, ist allerdings bisher in der Praxis nicht erfolgt [24].

Ausblick

Diesem Dokument folgt nun eine Themenveranstaltung, die aufklären und zu Diskussionen anregen soll. Außerdem wird es eine Online-Befragung zu dieser Thematik geben. Bis Ende 2014 werden alle Anregungen und die als (besonders) wichtig angesehenen Aspekte seitens der Bürger zusammengetragen. Es kann sich hierbei um ein Bestätigen und Bekräftigen der bereits diskutierten Themen handeln. Willkommen sind aber auch neue Sichtweisen, Bedenken und Anregungen. Möglich ist ebenfalls, dass eine Anwendung von Big Data übersehen wurde, die von einem Großteil der Bürger aber als motivierendes Beispiel für die Chancen durch Big Data gesehen wird. Dann wird das Dokument entsprechend erweitert. Oder es kommen neue Hinweise auf Kritikpunkte an Scoring hinzu. Genauso kann ein völlig anderer Typ von risikobehafteten Anwendungen aufgezeigt werden, der ebenfalls diskutiert werden sollte.

Ein abschließendes Dokument wird unsere Vorarbeiten und die Anregungen durch den Bürgerdialog auf eine geeignete Weise zusammenfassen, um neue Impulse für zukünftige Forschungen im Umfeld Big Data und Datenschutz zu liefern. Im Anschluss an den Bürgerdialog werden in der zweiten Phase des Forschungsvorhabens „Big Data und Privatsphärenschutz vom Bürgerdialog bis zur risikobehafteten explorativen Grundlagenforschung“ am Kompetenzzentrum EC SPRIDE wissenschaftliche Arbeiten durchgeführt, die die bereits in Abschnitt 4.3 kurz diskutierten technologischen Aspekte erneut aufgreifen und den Stand der Technik verbessern sollen.

Literatur

[1] Bachner, Jennifer: *Predictive policing: Preventing crime with data and analytics*. Report, IBM Center for The Business of Government, Johns Hopkins University, Juni 2013. http://www.businessofgovernment.org/sites/default/files/Predictive%20Policing.pdf.

[2] Beuth, Patrick: *Snowden-Enthüllungen – Alles Wichtige zum NSA-Skandal*. Zeit Online, Oktober 2013. http://www.zeit.de/digital/datenschutz/2013-10/hintergrund-nsa-skandal/komplettansicht, Inhalt zuletzt aktualisiert am 03.11.2014.

[3] Biermann, Kai: *NSA-Ausschuss sieht nur schwarz*. Zeit Online, September 2014. http://www.zeit.de/politik/deutschland/2014-09/nsa-ausschuss-akten-geschwaerzt.

[4] BITKOM: *Potenziale und Einsatz von Big Data*. Studienbericht, BITKOM, Mai 2014. http://www.bitkom.org/files/documents/Studienbericht_Big_Data_in_deutschen_Unternehmen.pdf.

[5] Bull, Hans Peter: *Es war einmal ein Datenschutz-Märchen*. Süddeutsche.de, November 2014. http://www.sueddeutsche.de/digital/pkw-maut-der-bundesregierung-es-war-einmal-ein-datenschutz-maerchen-1.2200854.

[6] Bundesministerium des Inneren: *Kennzeichenerfassung und Funkzellenabfrage im sogenannten Autotransporter-Fall*. Drucksache 17/14794, Deutscher Bundestag, September 2013. http://dip21.bundestag.de/dip21/btd/17/147/1714794.pdf.

[7] Committee on Civil Liberties, Justice and Home Affairs (LIBE): *General data protection regulation*. Inofficial consolidated version, European Parliament, Oktober 2013. http://www.janalbrecht.eu/fileadmin/material/Dokumente/DPR-Regulation-inofficial-consolidated-LIBE.pdf.

[8] Dix, Alexander: *Abschlussbericht zur rechtlichen Überprüfung von Funkzellenabfragen*. Prüfbericht, Berliner Beauftragter für Datenschutz und Informationsfreiheit, September 2012. http://datenschutz-berlin.de/attachments/896/Pr__fbericht.pdf.

[9] Fröhlich, Christoph: *Googles große Zwangseingemeindung*. Stern, Januar 2013. http://www.stern.de/digital/online/google-pflicht-fuer-youtube-und-co-googles-grosse-zwangseingemeindung-1952778.html.

[10] Ganslmeier, Martin: *Cyber-Dialog statt No-Spy-Abkommen*. Tagesschau, September 2014. http://www.tagesschau.de/ausland/cyberdialog100.html.

[11] Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski und Larry Brilliant: *Detecting influenza epidemics using search engine query data*. Nature, 457:1012–1014, Februar 2009. http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html.

[12] Karaboga, Murat, Philipp Masur, Tobias Matzner, Cornelia Mothes, Maxi Nebel, Carsten Ochs, Philip Schütz und Hervais Simo Fhom: *Selbstdatenschutz*. White Paper, Forum Privatheit, August 2014. https://www.forum-privatheit.de/forum-privatheit-de/texte/veroeffentlichungen-des-forums/themenpapiere-white-paper/Forum_Privatheit_White_Paper_Selbstdatenschutz_Web.pdf.

[13] Koeffer, Sebastian: *Mit Predictive Analytics in die Zukunft blicken*. Computerwoche, Juli 2014. http://www.computerwoche.de/a/mit-predictive-analytics-in-die-zukunft-blicken,2370894.

[14] Laney, Doug: *3D data management: Controlling data volume, velocity and variety*. Research note, META Group, Februar 2001. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

[15] Leyendecker, Hans und Georg Mascolo: *Generalbundesanwalt will nicht in NSA-Affäre ermitteln*. Süddeutsche Zeitung, Mai 2014. http://www.sueddeutsche.de/politik/abgehoertes-merkel-handy-generalbundesanwalt-will-nicht-in-nsa-affaere-ermitteln-1.1977054.

[16] Milian, Mark: *Google to merge user data across its services*. CNN, Januar 2012. http://edition.cnn.com/2012/01/24/tech/web/google-privacy-policy/.

[17] Raaz, Andreas: *Business Intelligence – Anwendung und Historie*. Whitepaper, PST Software & Consulting, Juli 2014. http://www.pst.de/fileadmin/user_upload/_de/pdf/Whitepaper_BI_Historie.pdf.

[18] Rijmenam, Mark van: *The Los Angeles Police Department is predicting and fighting crime with big data*. Use case, BigData Startups, April 2014. http://www.bigdata-startups.com/BigData-startup/los-angeles-police-department-predicts-fights-crime-big-data/.

[19] Roßnagel, Alexander, Eric Bodden, Philipp Richter und Siegfried Rasthofer: *Schutzmaßnahmen gegen datenschutzunfreundliche Smartphone-Apps*. Datenschutz und Datensicherheit, 37(11):720–725, November 2013.

[20] Schaar, Peter: *Verbraucherpolitik in der digitalen Welt – Der gläserne Kunde?* Stellungnahme, Bundesbeauftragter für den Datenschutz, April 2005. http://www.bfdi.bund.de/SharedDocs/Publikationen/VerbraucherpolitikInDerDigitalenWelt-DerGlaeserneKunde.html.

[21] Sweeney, Latanya: *K-anonymity: A model for protecting privacy*. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):557–570, Oktober 2002.

[22] Thoma, Jörg: *CCC stellt Strafanzeige gegen Bundesregierung*. Golem.de, Februar 2014. http://www.golem.de/news/spionageaffaere-ccc-stellt-strafanzeige-gegen-bundesregierung-1402-104324.html.

[23] Weichert, Thilo: *Datenschutzrechtliche Anforderungen an Verbraucher-Kredit-Scoring*. Datenschutz und Datensicherheit, 29(10):582–587, Oktober 2005.

[24] Weichert, Thilo: *Big Data und Datenschutz*. Stellungnahme, Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein, März 2013. https://www.datenschutzzentrum.de/bigdata/20130318-bigdata-und-datenschutz.pdf.

[25] Wienand, Lars: *Autobahnschütze: RLP-Datenschützer fordert Gesetzesänderung für Massen-Kennzeichen-Erfassung*. Interview mit Edgar Wagner, Rhein-Zeitung, August 2014. http://www.rhein-zeitung.de/region_artikel,-Autobahnschuetze-RLP-Datenschuetzer-fordert-Gesetzesaenderung-fuer-Massen-Kennzeichen-Erfassung-_arid,1192889.html.