Fraunhofer
SIT

FRAUNHOFER-INSTITUT FÜR SICHERE INFORMATIONSTECHNOLOGIE

BEGLEITPAPIER BÜRGERDIALOG

CHANCEN DURCH BIG DATA UND DIE FRAGE DES PRIVATSPHÄRENSCHUTZES

EC SPRIDE

INHALT

IMPRESSUM

Kontaktadresse

Fraunhofer-Institut für Sichere Informationstechnologie SIT
Rheinstraße 75, 64295 Darmstadt
Telefon 06151 869-213
Telefax 06151 869-224
E-Mail info@sit.fraunhofer.de
URL https://www.sit.fraunhofer.de/

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Herausgeber: Michael Waidner
SIT Technical Reports
SIT-TR-2015-06: Big Data und Privatheit
Dr.-Ing. Martin Steinebach, Christian Winter, Oren Halvani, Marcel Schäfer und York Yannikos
ISSN: 2192-8169

Fraunhofer-Informationszentrum Raum und Bau IRB
Postfach 800469, 70504 Stuttgart
Nobelstraße 12, 70569 Stuttgart
Telefon 0711 970-2500
Telefax 0711 970-2508
E-Mail verlag@fraunhofer.de
URL http://verlag.fraunhofer.de

Satz und Layout: Marion Mayer, Darmstadt

BIG DATA UND PRIVATHEIT

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

Fraunhofer-Institut für Sichere Informationstechnologie SIT – März 2015

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.

Dem ökonomischen Mehrwert steht jedoch ein erhebliches Risiko für die Privatsphäre der Bürger gegenüber. Ein Großteil der Daten beinhaltet private Informationen oder lässt sich zumindest in Verbindung zu Personen setzen. Big Data stellt solche Verbindungen her, verknüpft weitere Informationen und erlaubt somit detaillierte Rückschlüsse über individuelle Personen. Die Folge ist ein tiefer Einblick in die Privatsphäre des Einzelnen durch Unternehmen und Behörden.

Dieses Dokument gliedert sich in zwei Teile. Der erste Teil ist als Informationsbroschüre gedacht mit dem Ziel, eine bürgernahe Einführung in das Thema Big Data und in die damit einhergehenden Chancen für die Gesellschaft und Risiken für die Privatsphäre zu geben. Dieser Teil ist eine aktualisierte und geringfügig erweiterte Fassung der Informationsbroschüre, die im November 2014 mit der Ankündigung des Bürgerdialogs veröffentlicht wurde.

Der zweite Teil umfasst Auswertungen und Ergebnisse der Diskussionsveranstaltung »Bürgerdialog Big Data« vom 4. Dezember 2014, zu der eingeladen wurde, um über Chancen und Risiken für die Privatsphäre zu diskutieren und die Meinung der Bürger zu erfassen. Ebenfalls dargestellt wird die Auswertung eines entsprechenden Onlinefragebogens. Schließlich wird auch die allgemeine Meinung zu Big Data untersucht, die in Deutschland über das Internet kommuniziert wird.

*Telefon: 06151 869-349; Telefax: 06151 869-224; E-Mail: martin.steinebach@sit.fraunhofer.de

TEIL I. INFORMATIONSBROSCHÜRE

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.

Durch Big Data ergeben sich eine Reihe neuer IT-Lösungen in unterschiedlichen Bereichen der Gesellschaft. Beispielanwendungen (Kapitel 1) umfassen die Medizin (Prognose von Grippewellen durch Google und die Unterstützung bei der Krebsdiagnose durch IBM Watson), die Polizeiarbeit (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).

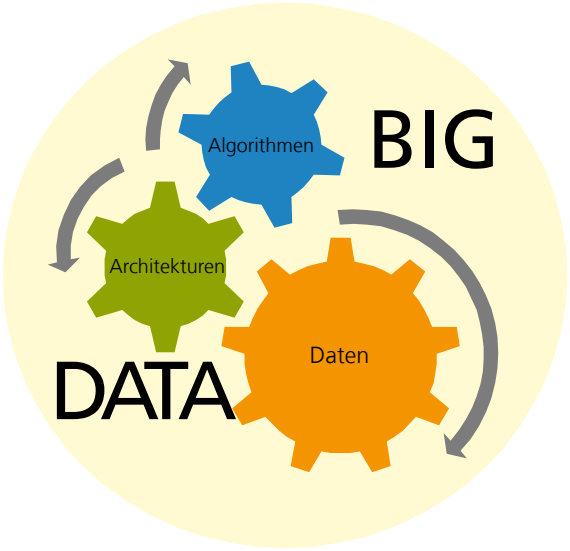

Abbildung 1: Big Data kombiniert Daten, Algorithmen und Systeme.

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.

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 (Kapitel 3). Ein erster Schritt für den Bürger ist die Kenntnis der Gesetzeslage hinsichtlich der Verwendung von personenbezogenen Daten (Abschnitt 3.2). 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 Dokument wird ein Schwerpunkt auf die Thematik von Profiling und Scoring gelegt (Kapitel 4). 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 4.2). 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.

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

Neben den gesellschaftlichen Fragestellungen widmet sich dieses Dokument ebenfalls den grundlegenden Technologien. Big Data bedeutet immer ein Zusammenspiel von leistungsfähigen Rechnerarchitekturen und geeigneter Software. Kapitel 2 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.

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.

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

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 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.

1. BEISPIELANWENDUNGEN

Big Data kann in den unterschiedlichsten Domänen eingesetzt werden. So helfen Big-Data-Anwendungen in der medizinischen Forschung und im Gesundheitswesen. Ebenso werden Big-Data-Technologien für Wettervorhersagen und Klimamodelle verwendet, um dynamische und möglichst echtzeitfähige Modelle zu erstellen. Auch in der Weltraumforschung und bei Teilchenbeschleunigern wird Big Data genutzt. Weitere Anwendungsfelder ergeben sich bei der Sicherheits- und Polizeiarbeit sowie bei der Infrastruktur von Mobilfunknetzen, Internet und intelligenten Stromnetzen (sog. Smart Grids). Auch für die Meinungs- und Trendforschung mittels Daten aus sozialen Medien verspricht Big Data enormes Potenzial. Offensichtliche Anwendungsmöglichkeiten für Big Data bestehen vor allem in den Bereichen Wirtschaft und Konsum. 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.

Eine Auswahl der Beispielanwendungen wird in Kapitel 6 aufgegriffen. Die in den Szenarien eingesetzten Big-Data-Lösungen werden dort hinsichtlich ihrer Chancen und Risiken aus Sicht von Bürgern diskutiert.

1.1. Google 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 Früherkennung, Prävention und Bekämpfung von Grippe leisten. Während die Daten von institutionellen Beobachtungsprogrammen mit ein bis zwei Wochen Verzögerung vorliegen, sind die Grippe-Trends von Google tagesaktuell.

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 [24]. Entsprechende Modelle wurden auch für andere Länder gebildet, u. a. für Deutschland (siehe Abbildung 3).

Google erstellt mithilfe seiner Schätzmodelle tagesaktuelle Grippe-Trends für mehr als 25 Länder und veröffentlicht diese unter http://www.google.org/ flutrends/intl/de. Dadurch soll

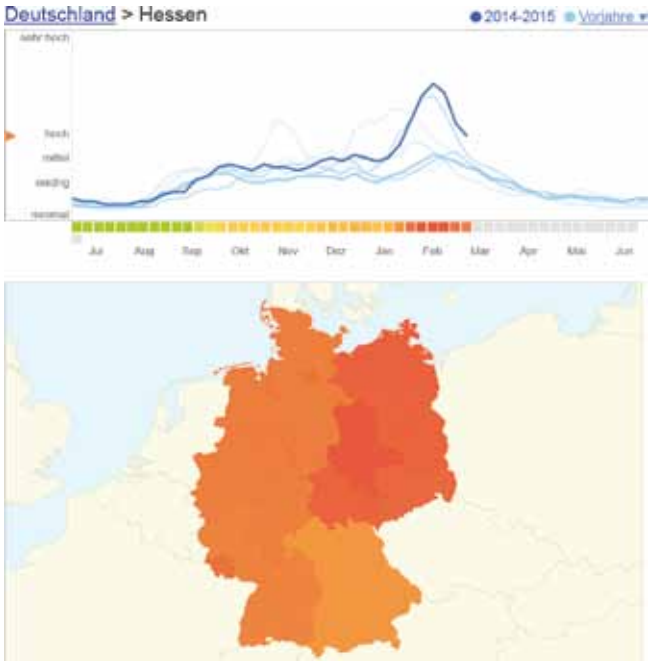

Abbildung 3: Google Grippe-Trends für Hessen vom März 2015. Wie zu erwarten steigen die Suchanfragen zum Thema Grippe im Herbst. Quelle: Google, http://www.google. org/flutrends/intl/de/de/#DE-HE

ermöglicht werden, früher und effizienter auf Grippewellen zu reagieren. Beispielsweise soll die Produktion und Verteilung von Impfstoffen und Medizin optimiert werden. Analog zu den Grippe-Trends veröffentlicht Google Schätzungen für das aktuelle Auftreten von Denguefieber in einigen tropischen Ländern.

Die Grippe- und Denguefieber-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 [24].

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 3.1 detaillierter aufgegriffen.

1.2. 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 langjährigen 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.

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

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 menschlichen 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 passende Frage formulieren zu können. Zu den Technologien zählen u. a. maschinelle Sprachverarbeitung, maschinelles Lernen (siehe Abschnitt 2.3), 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.

Von dem Einsatz der Watson-Technologie für Jeopardy geht keine Gefahr für die Privatsphäre aus, da in diesem Szenario ausschließlich Allgemeinwissen verarbeitet wird. Bei anderen Anwendungen werden aber durchaus personenbezogene Daten verarbeitet. Insbesondere im Gesundheitswesen treten sehr sensible Daten auf.

1.3. Predictive Policing

Eine wachsende Anzahl von Polizeibehörden nutzt Technologien, die als Predictive Policing bezeichnet werden. Solche Technologien berechnen Wahrscheinlichkeiten für zukünftige Straftaten. Verbreitete Varianten von Predictive Policing sind Weiterentwicklungen von Hotspot Mapping – Methoden, die auf einem Stadtplan Bereiche hervorheben, in denen Straftaten besonders wahrscheinlich sind. Predictive Policing legt dabei den Fokus auf die Zukunft, indem es für konkrete zukünftige Zeitpunkte konkrete Wahrscheinlichkeiten berechnet und Handlungsempfehlungen erstellt.

Die Polizei von Santa Cruz in Kalifornien setzt seit Mitte 2011 Predictive Policing ein [46]. Seitdem erhalten die dortigen Polizeibeamten zu Beginn jeder Schicht einen Stadtplan, auf dem einige Quadrate mit einer Kantenlänge von 500 Fuß (ca. 150 Meter) markiert sind (siehe Abbildung 5). Die Beamten sollen sich bei ihren Patrouillen möglichst oft in den markierten Boxen aufhalten, um potentzielle Straftäter abzuschrecken und um nach Auffälligkeiten zu suchen. Die zugrunde liegende Technologie entstammt einem Forschungsprojekt und wird seit 2012 von

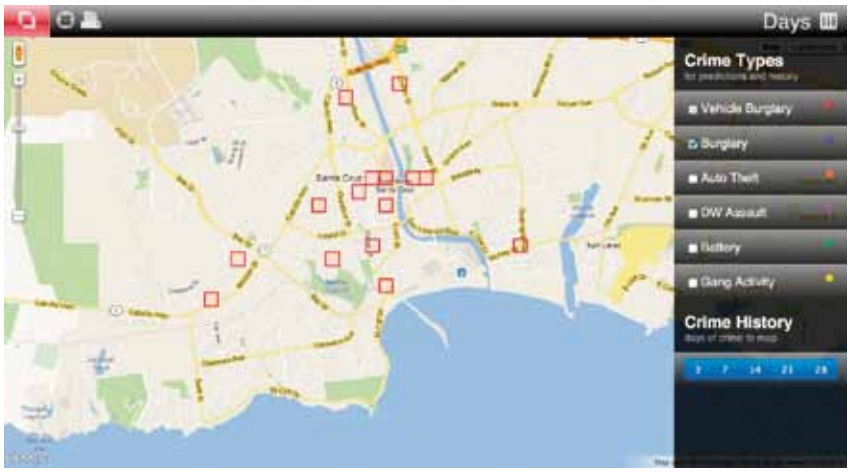

Abbildung 5: Predictive Policing: PredPol markiert Bereiche mit erhöhter Wahrscheinlichkeit für Straftaten auf einer Karte von Google Maps. Quelle: [1]

der eigens dafür gegründeten Firma *PredPol* vertrieben.

Eine ähnliche Lösung namens *Precobs* wird von der bayerischen Polizei seit September 2014 in München und Nürnberg getestet, nachdem sie in Zürich bereits Ende 2013 eingeführt worden war [11]. Die Polizei in Nordrhein-Westfalen möchte die Software ab Oktober 2015 in Köln und Duisburg einsetzen [30].

Die Vorhersagen von PredPol und ähnlichen Verfahren beruhen auf den zuvor erfassten Straftaten. Dabei wird lediglich der Ort, die Zeit und die Art der Straftat (z. B. Einbruch, Diebstahl von/aus einem Fahrzeug oder Gewaltdelikt) verwendet. Die Entwickler des Pred-Pol-Verfahrens haben dieses anhand von 13 Millionen Delikten aus den vergangenen 80 Jahren erstellt und validiert [42].

PredPol nutzt ein statistisches Verteilungsmodell, mit dem auch Häufigkeiten von Erdbeben beschrieben werden. So wie bei einem Erdbeben die Wahrscheinlichkeit für ein weiteres Beben in zeitlicher und räumlicher Nähe erhöht ist, so gibt es ebenfalls eine gewisse Wahrscheinlichkeit, dass beispielsweise ein Einbrecher in eine Nachbarschaft zurückkehrt, in der er einmal Erfolg hatte. Das verwendete Modell liefert auf Basis der vergangenen Ereignisse eine Schätzung der zukünftigen Verteilung der Ereignisse. Bei jedem neuen Ereignis ändert sich die Schätzung. Für eine übersichtliche Darstellung der Prognose werden die Blöcke mit der höchsten Straftatwahrscheinlichkeit auf einer Karte angezeigt.

PredPol trifft mit seinen Vorhersagen doppelt so viele Straftaten wie Analysten.

Bei einer Studie in Los Angeles – dem Ursprungsort des Verfahrens – wurde festgestellt, dass in den von PredPol markierten Blöcken doppelt so viele Straftaten stattfanden wie in den von Analysten ausgewählten Blöcken [22]. Für zahlreiche Städte, die die Software einsetzen, wird in der Presse von einem Rückgang der Kriminalität berichtet.

Die hier dargestellte Form von Predictive Policing vermeidet personenbezogene Daten weitestgehend. Lediglich die erfassten Orte von Straftaten könnten zum Teil einer Person zugeordnet werden – etwa wenn es sich um den Ort eines Wohnungseinbruchs handelt.

Es lassen sich aber auch Szenarien erdenken, die tief in die Privatsphäre eindringen. Beispielsweise könnten Daten aus sozialen Netzwerken oder von Überwachungskameras an öffentlichen Plätzen genutzt werden, um Vorhersagen über Personen zu generieren. In Romanen und Filmen werden fiktive Methoden von Predictive Policing zur Schaffung von Dystopien wie *Minority Report* genutzt (siehe Abbildung 6).

Abbildung 6: Die Idee der Erkennung von zukünftigen Straftaten 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

Aber selbst bei dem Verfahren von PredPol gibt es Bedenken bezüglich negativer Auswirkungen auf die Bürger. So wird befürchtet, dass in den markierten Blöcken diskriminierendes Polizeiverhalten zunimmt, Personen ohne klare Indizien polizeilich durchsucht werden und unbescholtene Bürger sich zwingen, jede potenzielle Auffälligkeit zu vermeiden und ihre Freiheit damit unnötigerweise einschränken [2, 46]. Die Öffentlichkeit sieht neben den Chancen von Predictive Policing also auch Risiken, die durch Transparenz und maßvollen sowie aufgeklärten Umgang mit der Technologie vermieden werden können.

1.4. 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 weder durch Fahndungsfahrten mit Autotransportern, öffentliche Aufrufe an Berufskraftfahrer und Bürger noch durch eine ausgeschriebene Belohnung von 100.000 Euro für sachdienliche Hinweise sowie einen Bericht bei *Aktenzeichen XY* 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). Die Gerichtsverhandlungen begannen im August 2014. Die Anklage umfasste 171 Fälle, wobei dem Täter Verstoß gegen das Waffengesetz, Sachbeschädigung, gefährlicher Eingriff in den Straßenverkehr, gefährliche Körperverletzung und in fünf Fällen versuchter Mord vorgeworfen wurde. Ende Oktober 2014 wurde der Schütze zu zehneinhalb Jahren Haft verurteilt.

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

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.

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 durch den großen Aufwand bisher einmalig. Theoretisch könnten auch Daten oder Infrastruktur des deutschen Lkw-Mautsystems genutzt werden, was jedoch gesetzlich nicht zulässig ist. Politische Diskussionen konnten bisher nichts daran ändern. 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 [12].

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 [17].

Nach der Festnahme des Autobahnschützen wurde mehrmals die Zulässigkeit der vorangegangenen Kennzeichenerfassung öffentlich diskutiert. Das BKA weist den Vorwurf der unverhältnismäßigen Datensammelei zurück. Die Bundesregierung hat im September 2013 eine »Kleine Anfrage« der Linken zu Umfang und Zulässigkeit der Datenerhebung beantwortet [13], woraus die meisten hier genannten Zahlen stammen. Der rheinland-pfälzische Datenschutzbeauftragte Edgar Wagner sieht keine ausreichende Rechtsgrundlage für die Kennzeichenerfassung und fordert gesetzliche Neuregelungen [52]. Die Anwälte des Schützen forderten ein Beweisverwertungsverbot.

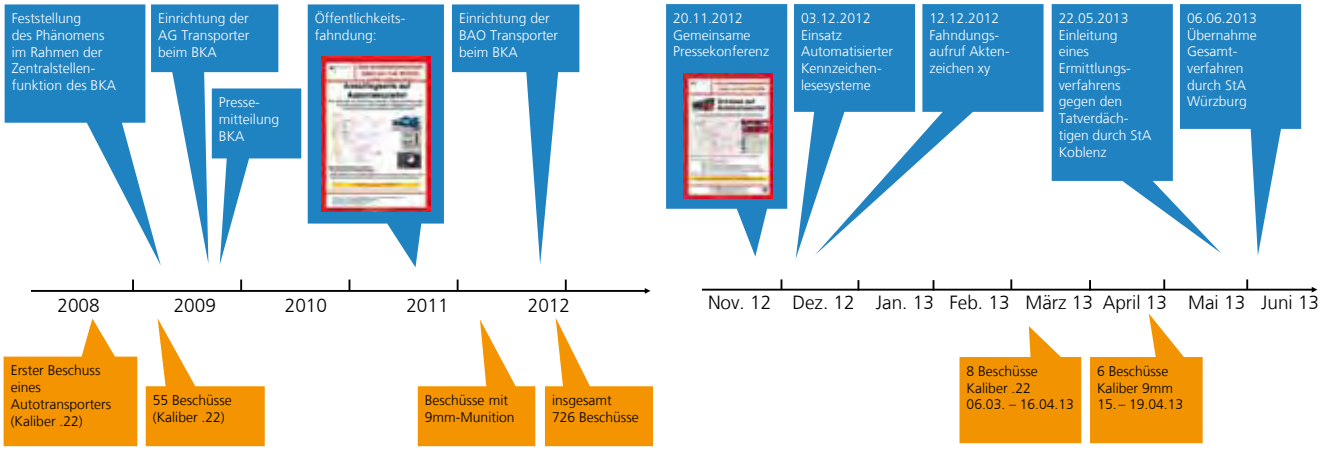

Abbildung 7: Chronologie der Fahndung zum Autobahnschützen. Quelle: BKA, https://www.bka.de/nn_196810/DE/ Presse/Pressemitteilungen/Presse2013/130625__BAOTransporterPressekonferenz.html

1.5. Überwachung durch NSA und GCHQ

Im Juni 2013 veröffentlichte die britische Zeitung *The Guardian* geheime Dokumente, die ihr 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 nach bisherigen Erkenntnissen abgehört wurden oder werden, ist auf *Zeit Online* zu finden [3].

Die Daten werden hauptsächlich direkt auf der Infrastruktur von Telekommunikationsprovidern oder Dienstanbietern mit großer Benutzerbasis erhoben. Ein Beispiel ist die *Operation Eikonal*, in deren Rahmen der deutsche Geheimdienst BND zusammen mit der NSA über mehrere Jahre Kommunikationsdaten auf dem vom Durchsatz her weltweit größten Internet-Knoten DECIX in Frankfurt ausgeleitet 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 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 [23]. 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 2014 unter massiver Kritik verworfen [34]. Im März 2014 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 [5].

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 Friedensnobelpreis 2014 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 [48]. Das Interesse an technischen Schutzmaßnahmen vor umfassender Überwachung ist in der Öffentlichkeit seit den letzten Jahren 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.

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/

1.6. 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 [39]. BI existiert als Begriff schon seit dem 19. Jahrhundert und wird heute verbreitet als Oberbegriff eingesetzt, der verschiedene Ansätze und Vorgänge einschließt. So erläutert Thomas Davenport, dass Business Analytics (BA) als Teilgebiet von BI gesehen werden kann und dass BA einen Schwerpunkt auf statistische Analysen und abgeleitete Vorhersagen setzt [27].

Big Data liefert für BI bzw. BA neue Möglichkeiten, spezifische Muster und Zusammenhänge in (vorwiegend großen) Datenbeständen zu identifizieren und mögliche Trends vorherzusagen. Hierbei können sowohl strukturierte Daten, z. B. Datenbanktabellen, als auch unstrukturierte Daten, beispielsweise Texte aus sozialen Netzwerken, in die Analyse einbezogen werden [32]. 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

Abbildung 9: Business Intelligence (BI)

verstehen.

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 Trendprognose. Andere Formen sind ebenfalls möglich, etwa Zusammenfassungen mit z. B. negativer oder positiver Bewertung eines Produkts.

BI fokussiert in erster Linie auf das Unternehmen und 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.

1.7. 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.

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

Da dieses Scoring standardisierte Bewertungen auf Grundlage von Statistiken liefert, wird trotz persönlicher Daten und individuellem Score die Individualität 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 Kapitel 4 noch einmal in allgemeinerer Form behandelt.

2. 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 [33].

Abbildung 11: Die grundlegenden Herausforderungen werden als die drei »Vs« beschrieben. Je nach Anwendung kommen weitere »Vs« hinzu.

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, Onlinekunden 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.

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.

2.1. 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. Dadurch wird eine deutliche Geschwindigkeitssteigerung von Schreib- und Lesevorgängen erreicht. 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.

In-Memory Analytics benutzt In-Memory-Datenbanken als Basistechnologie, welche die Daten im Hauptspeicher eines oder mehrerer Computer vorhalten und Schnittstellen für die Datenverarbeitung anbieten. 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 Familiennamen, für den Vornamen, für die Postleitzahl etc. In vielen Big-Data-Anwendungsfällen kann jedoch nicht von solch strukturierten Daten ausgegangen werden. Tatsächlich ist es eine der Stärken von Big Data, auch auf unstrukturierten Daten arbeiten zu können. Daher sind inzwischen immer mehr Datenbankkonzepte 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. Dabei fasst der Begriff NoSQL viele unterschiedliche Architekturkonzepte zusammen, die je nach Art der Daten und Anwendungsszenario zum Einsatz kommen. NoSQL-Systemen wird oft eine höhere Performanz und einfachere Skalierbarkeit als herkömmlichen Datenbanklösungen zugesprochen.

NoSQL-Datenbanken sind im Medienumfeld und bei sozialen Netzwerken sehr verbreitet, wo große Mengen unstrukturierter Daten anfallen. Bekannte Beispiele sind die Open-Source-Lösung MongoDB oder auch IBM Notes, welches vor 2013 unter dem Namen Lotus Notes vertrieben wurde und schon lange eine verteilte und dokumentorientierte Architektur aufweist.

2.2. 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. MapReduce ist ein Ansatz, der auf viele datenintensive Aufgaben angewendet werden kann. Dabei muss der Anwender im Wesentlichen eine

Map-Funktion definieren, die zum Lösen der Teilaufgaben verwendet wird, und eine Reduce-Funktion, die für das Zusammenführen von Teilergebnissen genutzt wird. Beide Funktionen haben eine festgelegte Form von Ein- und Ausgabe. Ein MapReduce-Framework, z. B. Apache Hadoop, sorgt dafür, dass die Map- und Reduce-Aufgaben in einem Rechnernetz verteilt ausgeführt werden.

2.3. 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). 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. Spam und 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 Nachrichtenkontext 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 Kombination 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 (ähnlich einer Anwendung von regulären Ausdrücken innerhalb einer Textsuche) und entsprechend zu bewerten bzw. zu klassifizieren. Dabei wird zwischen expliziten und impliziten Mustern 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.

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 gilt es bei DM typischerweise, neue Muster, also implizites Wissen, in Daten ausfindig zu machen. Bei ML dagegen werden Muster vorausgesetzt, um überhaupt erst Modelle zu konstruieren, mit denen Daten anschließend automatisch klassifiziert bzw. gruppiert werden können. 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.

3. 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 [6] 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 z. B. auch Facebook, die NSA oder die Schufa gesehen. Quelle: PC Magazin, http://www.pcmagazin.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 begünstigt, dass die Technologie immer leistungsfähiger wird.

3.1. Profilbildung anhand der Verschmelzung von Google-Diensten

Google hat im März 2012 viele seiner Dienste (Gmail, YouTube, Google+ etc.) zusammengelegt [36], 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 (Download-Service für Android-Anwendungen) schreiben zu können, benötigt ein Nutzer ein verknüpftes Google+ Konto [21]. Auch für Android, das von Google vertriebene Betriebssystem 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.

3.2. 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.

Eine klare Einteilung in personenbezogene und nicht personenbezogene Daten erweist sich jedoch oftmals als schwierig, da 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).

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 häufig Aussagen über Personengruppen gemacht werden sollen, wofür aber personenbezogene Daten herangezogen werden müssen. 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.

3.2.1. Datenschutzprinzipien

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** (§ 3a BDSG) 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** (§ 39 BDSG): 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** (§ 4a BDSG) 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** (§ 34 BDSG) 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 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** (§ 35 BDSG) geben ihm das Recht, falsche Daten berichtigen und bestimmte Daten löschen bzw. sperren zu lassen. 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.

3.2.2. Datenschutz in Europa und den USA

Neben dem BDSG wird auch die Umsetzung der aktuell zur Diskussion stehenden Datenschutz-Grundverordnung der Europäischen Union [16] 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, wird die Grundverordnung rechtsverbindlich und das BDSG wird davon abgelöst.

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.

In den USA fehlt eine unabhängige Datenschutzaufsicht.

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 *Patriot Act* zu. Dieses US-amerikanische Bundesgesetz sichert den Behörden, insbesondere FBI, CIA und NSA, weitreichende Rechte zu und setzt kollidierende Gesetze einzelner Bundesstaaten außer Kraft. Als Reaktion auf die Anschläge des 11. September 2001 verabschiedet, hat es zum Ziel, den internationalen Terrorismus zu bekämpfen. Die weitreichenden Auswirkungen auf die Privatsphäre wurden bereits in Abschnitt 1.5 behandelt.

3.3. Technische Schutzmaßnahmen

Um bei personenbezogenen Daten datenschutzrechtliche Bestimmungen umzusetzen, existiert eine Reihe von technischen Lösungen. Dabei ist zu unterscheiden 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 voneinander **abgeschottet** sind. So können Anwender nicht gegenseitig ihre Daten in einem laufenden Verarbeitungsprozess einsehen.

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

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) [47], 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 [10]. 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 [31].

4. 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.

4.1. 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.

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 1.7 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 geliefert 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.

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

Terrorbekämpfung: Insbesondere durch das groß angelegte Sammeln und Auswerten von

Daten durch Geheimdienste (siehe Abschnitt 1.5) 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. Hier wird zunächst ein Täterprofil des unbekannten Täters basierend auf den Informationen zur Tat erstellt. Die Ermittler vergleichen dieses Profil mit geeigneten Datenbeständen, um darin einen Kreis verdächtiger Personen zu identifizieren. Die Datenbestände werden durch die Verknüpfung verschiedener Quellen, z. B. behördliche Register, gewonnen und betreffen einen bestimmten Teil der Bevölkerung. Diese Ermittlungsmethode bezeichnet man als Rasterfahndung.

4.2. 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 vor den Risiken und stellen auch die rechtliche Grundlage des Profilings infrage. So stellte bereits im Jahr 2005 der damalige Bundesbeauftragte für den Datenschutz und die Informationsfreiheit, Peter Schaar, die Risiken von Profiling und Scoring heraus [43]. 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 bekommt und selbst der Zahnersatz nur gegen Vorkasse gewährt wird [...].« [43, Seite 5]

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

Abstrakter formuliert Thilo Weichert, Datenschutzbeauftragter des Landes Schleswig-Holstein, 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.« [50, Abschnitt 1]

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 [51].

5. LITERATUREMPFEHLUNGEN

An dieser Stelle möchten wir interessierte Bürger auf weiterführende Literatur zu den Themen Big Data und Privatsphäre hinweisen.

Die Publikationen des BITKOM richten sich hauptsächlich an Fach- und Führungspersonen in Unternehmen, sind aber auch für den Bürger lesenswert und unter http://www.bitkom.org/de/publikationen/1357.aspx zu finden. Mehrere Leitfäden behandeln das Thema Big Data [7, 8, 9]. An dieser Stelle soll insbesondere auf die Sammlung von Beispielanwendungen im Leitfaden »Big Data im Praxiseinsatz« hingewiesen werden [7, Kapitel 10].

Das *Forum Privatheit* ist eine interdisziplinäre, vom BMBF geförderte Plattform verschiedener Universitäten, Forschungsinstitute und öffentlicher Datenschutzeinrichtungen. Auf der Homepage https://www. forum-privatheit.de sind aktuelle Veranstaltungen, Literaturhinweise und weiteres Wissenswertes rund um das Thema Privatheit im Internet zu finden.

Die Informationsbroschüren in der »Blauen Reihe« (https://www.datenschutzzentrum.de/blauereihe/) des *Unabhängigen Landeszentrums für Datenschutz Schleswig-Holstein* (ULD) liefern praxisnahe und leicht verständliche Übersichten zu unterschiedlichen Themen rund um den Daten- und Privatsphärenschutz, u. a. zu Verbraucher-Scoring.

Risiken von Anwendungen zur automatischen Generierung von Schlussfolgerungen mittels Big Data werden in dem Artikel »Denkverbote für Star-Trek-Computer?« herausgearbeitet [49]. Dabei wird die aktuelle Gesetzeslage zu Datenschutz und Scoring kritisch untersucht.

Die Studie »Kommerzielle digitale Überwachung im Alltag« gibt einen vertieften Einblick in die alltäglich stattfindende Datenerfassung, -aggregation und -auswertung sowie den damit verbunden Datenhandel [14]. Insbesondere sind viele Beispiele zu Profiling aufgeführt.

Das Buchkapitel »Seven Types of Privacy« untersucht die Wechselwirkung von Technologie und Privatheit [19]. Dafür werden sieben Aspekte von Privatheit beschrieben, und es wird für Beispielanwendungen aufgezeigt, welche Aspekte der Privatheit davon berührt werden. Diese Arbeit basiert unter anderem auf Roger Clarkes Abhandlungen zu Privatsphäre und »Dataveillance« [15].

Die TED-Playlist »The dark side of data« auf http: //www.ted.com/playlists/130/the_dark_side_of_data liefert eine Videoauswahl an interessanten kritischen Vorträgen z. B. zur Überwachung der Bürger durch Regierungsapparate oder kommerzielle Unternehmen, aber auch zu generellen IT-sicherheitsrelevanten Themen wie bspw. sichere Passwörter.

Der Enthüllungsjournalist Glenn Greenwald, den Edward Snowden für die Veröffentlichung des NSAÜberwachungsskandals ausgewählt hatte, beschreibt in dem Buch »No Place to Hide« seine Erlebnisse und Einsichten zum NSA-Skandal [26].

Die Überwachung in Deutschland durch deutsche und ausländische Geheimdienste wird von Josef Foschepoth in dem Buch »Überwachtes Deutschland« als historischer Entwicklungsprozess dargelegt [20]. Dieser Prozess wirft einen neuen Blick auf die Entwicklung der Bundesrepublik und ist für die Erklärung der aktuellen Überwachungspraxis wichtig.

TEIL II.
AUSWERTUNGEN ZUM BÜRGERDIALOG

Die Chancen und Risiken verschiedener Big-Data-Anwendungen können Auswirkungen für jeden einzelnen Bürger haben. Deshalb ist es wichtig zu erfahren, welche Haltung die Bürger hierzu haben. Möglicherweise sehen sie einen persönlichen oder gesellschaftlichen Gewinn, z. B. in Form von mehr Komfort oder Sicherheit. Damit verbundene Eingriffe in die Privatsphäre können unterschiedliche Reaktionen hervorrufen. Der Bürger kann die Datensammlung zugunsten der Chancen dulden, im Sinne einer Post-Privacy-Haltung als belanglos betrachten oder als Überwachung und Freiheitsverlust empfinden. Die Stimmen der Bürger werden hier auf verschiedene Weisen erfasst und wiedergegeben:

Um ein Stimmungsbild der Bürger bzgl. Big Data und Privatsphärenschutz zu erhalten, wurde am 4. Dezember 2014 ein Informations- und Diskussionsabend am Fraunhofer SIT als Bürgerdialog mit Open-Space-Diskussionsforum veranstaltet. Impulse der einzelnen Themengruppen sind in Kapitel 6 aufgeführt.

Daneben wurde im November und Dezember 2014 eine Onlinebefragung durchgeführt, um die Einstellung der Bürger zu Big Data und Privatsphärenschutz zu erheben. Dazu wurden auch Kenntnisse und Meinungen bzgl. Datenschutz sowie Angaben zur eigenen Internetnutzung abgefragt. Eine Auswertung der Onlinebefragung befindet sich in Kapitel 7.

Weiterhin wurden Onlineartikel zu Big Data und die zugehörigen Nutzerkommentare sowie Tweets mit dem Hashtag #bigdata gesammelt und analysiert (siehe Kapitel 8), um auch hier die Haltung des Bürgers zum Thema Big Data und Privatsphärenschutz zu erfassen.

6. IMPULSE VOM BÜRGERDIALOG

Am 4. Dezember 2014 wurde am Fraunhofer SIT in Darmstadt ein Bürgerdialog zum Thema »Big Data – Chancen und Risiken« durchgeführt. In einem ersten Teil wurde ins Thema eingeführt und ein Vortrag über »Big Data in der Meteorologie« von Dr. Jürgen Seib vom Deutschen Wetterdienst gehalten. Der zweite Teil bestand aus einem Open-Space-Dialog mit den Besuchern: Es wurden Thementische angeboten, an denen die Autoren des vorliegenden Dokuments zur Diskussion der Chancen und Risiken bezüglich der Szenarien »NSA: Big-Data-Werkzeuge bei Geheimdiensten«, »Google Grippe-Trends«, »Scoring bei Banken und Versicherungen«, »IBM Watson gewinnt bei Jeopardy« sowie »BKA ermittelt Autobahn-Schützen« einluden. Die Auswahl der Szenarien stellt dabei einen Querschnitt aus den in diesem Dokument behandelten Szenarien dar (siehe Kapitel 1). Die Auswahl geschah auf Basis der Griffigkeit der Szenarien; solche mit einem nachvollziehbaren Bezug zum täglichen Leben wurden bevorzugt. Im Folgenden werden die Eindrücke und Impulse der Thementische zusammengefasst.

Abbildung 16: Zum Bürgerdialog erschienen Besucher verschiedener Altersstufen.

6.1. Google Grippe-Trends

Am Thementisch »Google Grippe-Trends« diskutierten die Teilnehmer die Chancen und Risiken der gleichnamigen Big-Data-Anwendung. In den Grippe-Trends veröffentlicht Google tagesaktuelle Grippe-Schätzungen, die auf den Suchanfragen der Google-Nutzer beruhen (siehe Abschnitt 1.1).

Bei der Erörterung von Chancen durch die Grippe-Trends nannten die Teilnehmer die rechtzeitige Deckung des Bedarfs an Medikamenten ebenso wie räumlich und zeitlich passend platzierte Werbung für Arzneimittel. Darüber hinaus wurde die Möglichkeit erkannt, Ausbreitungswege der Grippe besser nachzuvollziehen und daraus Empfehlungen zur Vorbeugung abzuleiten. Die Teilnehmer erwogen auch den zusätzlichen gesellschaftlichen Nutzen von Schätzungen für weitere Krankheiten. Ein klares Geschäftsmodell für Google hinter der Erstellung und Veröffentlichung der Grippe-Trends konnte in der Diskussion jedoch nicht ermittelt werden.

Als mögliches Risiko sahen die Teilnehmer, dass die Grippe-Trends oder denkbare ähnliche Anwendungen unnötige Panik auslösen könnten. Es wurde auch überlegt, dass Fehlprognosen durch die mediale Präsenz der Grippe-Erkrankung prominenter Personen oder durch Angriffe aus Botnetzen verursacht werden könnten. In der Tat wurde in den USA die Grippewelle 2012/13 von Google stark überschätzt, da die Medien in jener Saison dem Thema Grippe große Aufmerksamkeit schenkten (https://drive.google.com/file/ d/0B1Ul69AUsTn1WWdJUnJFYnNDbkk/view).

Eine Einschränkung der Privatsphäre sahen die Teilnehmer nicht unmittelbar bei den veröffentlichten Grippe-Trends. Es wurden jedoch große Bedenken bzgl. der starken Präsenz verschie-

dener Google-Dienste im Internet und der aggregierten Profilerstellung aus den Nutzerdaten dieser Dienste (siehe Abschnitt 3.1) geäußert. Konkret wurde befürchtet, dass Nutzerprofile in die falschen Hände geraten könnten oder Erkenntnisse über Nutzer z. B. an Versicherungen verkauft werden könnten.

6.2. IBM Watson

Die Teilnehmer des Thementischs »IBM Watson« wurden gefragt, wie sie zu automatisierten Analysen im Big-Data-Umfeld am Beispiel von IBM Watson stehen (siehe Abschnitt 1.2). Dafür wurden die Anwendungsszenarien Kriminalitätsbekämpfung (Strafverfolgung, Fahndung und Ermittlungsarbeit), Marketing (individuelle Werbung), Medizin (Unterstützung von Diagnosen) sowie Wetter und Umwelt (insbesondere Vorhersagen von Naturkatastrophen) diskutiert. Es stellte sich heraus, dass die Mehrheit der Teilnehmer dem Thema eher kritisch gegenüber steht. Die Gründe hierfür siedelten jedoch weniger im Bereich des Privatsphärenschutzes, als in der Furcht vor Fehlern in der Technik (Identifikation falscher Verdächtiger, fehlgeleitete Interpretation bei der Erstellung von Diagnosen, falscher Alarm vor Wirbelstürmen und unberechtigte »Panikmache«).

Des Weiteren wurde das Thema der Anonymität auf Webplattformen besprochen. Die Teilnehmer diskutierten, inwiefern für sie der Einsatz von Anonymisierungs- und insbesondere De-Anonymisierungstechniken gerechtfertigt werden könne. Die Mehrheit der Teilnehmer sprach sich dafür aus, dass der Einsatz von De-Anonymisierung legitim sei, sofern es um die Identifikation von Personen geht, die im direkten Zusammenhang mit kriminellen Handlungen stehen (bspw. bei der Rekrutierung von Dschihadisten). Dem entgegen stand eine Minderheit, die sich für ein generelles Verbot von De-Anonymisierungstechniken aussprach. Begründet wurde dies mit dem hohen Missbrauchspotenzial, das De-Anonymisierungstechniken bergen, indem sie kriminelle Handlungen erst ermöglichen (bspw. Erpressung oder Mobbing von Personen, die sich kritisch zu einem bestimmten Sachverhalt äußerten und erst durch De-Anonymisierung identifiziert werden können).

Neben der Vielzahl an kritischen Äußerungen und Bedenken bzgl. Big-Data-Analysemethoden ist jedoch auch erwähnenswert, dass es durchaus als positiv angesehen wurde, wenn Analysemethoden als Unterstützung von Expertenmeinungen herangezogen werden (bspw. wenn eine automatisierte Analyse die Erkenntnis eines Experten unterstützt).

6.3. Autobahnschütze

Am Thementisch »Autobahnschütze« diskutierten die Teilnehmer den Einsatz von Big Data mittels eingescannter Autokennzeichen und Mobilfunkeinwahlen zur Identifikation des sogenannten Autobahnschützen (siehe Abschnitt 1.4).

Durchgängig wurden die Vorteile flächendeckender Verkehrsüberwachung mithilfe von Mautsystemen oder – wie im Falle des Autobahnschützen – von separaten Kennzeichenlesegeräten gegenüber den Nachteilen abgewogen, aber mit unterschiedlicher Gewichtung. Als

positiv wurde der Schutz gegen Kriminalität, besonders vor schwer gesellschaftsschädigender Kriminalität, hervorgehoben. Negativ empfanden die Teilnehmer jedoch die generelle Erhebung von Bewegungsdaten unbescholtener Bürger. Begründet wurde dies durch die Erweckung von Begehrlichkeiten zu zweckfremder Nutzung.

Als weiteres Beispiel für eine solche doppelbödige Anwendung diskutierten die Teilnehmer den illegalen Ankauf von CDs aus der Schweiz, mithilfe derer Steuersünder identifiziert und überführt werden konnten.

Als Staaten, die bereits eine solche Überwachung vornehmen, wurden die Schweiz sowie England (Videoüberwachung in London) aufgeführt. Das scheine die Lebendigkeit der jeweiligen Demokratie nicht zu gefährden, wurde von den Diskussionsteilnehmern aber dennoch als Vorbild für Deutschland abgelehnt.

Die Diskussion ging schließlich dazu über, Werkzeuge zum Selbstdatenschutz vorzustellen, darunter die Browser-Plugins NoScript und Ghostery. Das beruhte auf dem allgemeinen Unbehagen vor der Gefahr, dass normale Browsing-Metadaten zweckentfremdet verwertet werden können. Ein vergleichbarer Selbstschutz wäre im Straßenverkehr nicht vorhanden.

6.4. NSA

Zu Beginn der Diskussion wurden die Teilnehmer des Thementischs »NSA: Big-Data-Werkzeuge bei Geheimdiensten« gefragt, inwieweit sich ihr Empfinden über die Sicherheit ihrer Privatsphäre im Netz aufgrund der bekannt gewordenen Überwachungsmaßnahmen (siehe Abschnitt 1.5) geändert habe. Die Mehrheit der Teilnehmer sagte dazu, dass ihr Vertrauen in Hard- und Software, die mit schützenswerten Daten umgeht, maßgeblich gesunken sei. Weitgehend Konsens war, dass der Bürger sich mittlerweile damit abfinden müsse, dass seine Telekommunikationsdaten überwacht werden und dass Eigeninitiative bzgl. Schutzmaßnahmen gegen Überwachung gezeigt werden müsse. Hinsichtlich der Sicherheit von Verschlüsselungsverfahren wurde im Dialog mit den Bürgern Unsicherheit deutlich, wobei unter anderem unklare Medienberichte als Ursache genannt wurden. Konkret wurden Zweifel hinsichtlich der Sicherheit der heutigen Verfahren zur Transportund Ende-zu-Ende-Verschlüsselung (bspw. TLS/SSL, PGP/GPG) gegen die technischen Möglichkeiten der überwachenden Institutionen geäußert.

Als positiver Aspekt der Snowden-Enthüllungen wurde genannt, dass der Bürger sich selbst seit dem letzten Jahr stärker mit dem Schutz der eigenen Privatsphäre auseinandergesetzt habe und dass mittlerweile viele Möglichkeiten zur Verfügung stünden, um sich umfassend zu informieren. Ebenfalls als positiv wurde hervorgehoben, dass einige Entwickler von Hard- und Software explizit in Datenschutz und -sicherheit investiert hätten und nun entsprechend besser geschützte Produkte anbieten würden (bspw. sichere Chat-Software wie Text-Secure). Kontrovers wurde die Meinung einer Minderheit diskutiert, die es befürwortete, den Schutz der Privatsphäre einem höchstmöglichen Maß an Sicherheit für den Bürger unterzuordnen und somit eine vollständige verdachtsunabhängige Überwachung durch legitimierte Behörden zu akzeptieren. Die Teilnehmer waren sich jedoch einig, dass aktuell keine Belege existierten, die sowohl eine Notwendigkeit

des Einsatzes eines solchen Überwachungsapparats als auch eine dadurch signifikant verbesserte Effektivität hinsichtlich der Kriminalitäts-/Terrorbekämpung bestätigten.

6.5. Scoring

Am Thementisch »Scoring bei Banken und Versicherungen« wurde das Schwerpunktthema dieses Dokuments behandelt (siehe Abschnitt 1.7 und Kapitel 4). Hier waren die Teilnehmer sowohl bei den als kritisch angesehenen Aspekten als auch bei den Chancen durch Scoring einer Meinung: Generelles Bedenken wurde hinsichtlich der Intransparenz von Kredit-Scoring geäußert. So sei für den Bürger nicht nachvollziehbar, wie und auf welcher Grundlage der erzielte Score zustande kommt. Ebenfalls ausdrücklich unzufrieden waren die Teilnehmer mit der Undurchsichtigkeit und der daraus empfundenen Machtlosigkeit darüber, wer persönliche Daten erhebt bzw. wer dazu befugt ist.

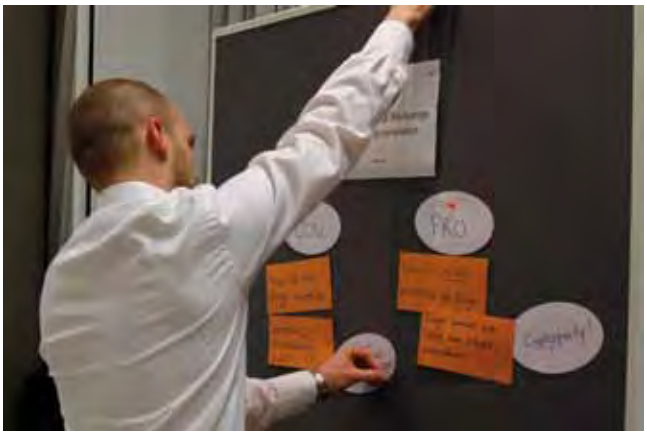

Abbildung 17: Während der Tischgespräche wurden Chancen und Risiken der Szenarien herausgearbeitet und diskutiert.

Sowohl ethisch als auch sozial bedenklich sei, dass von Daten vieler Beobachtungen pauschal auf alle Menschen geschlossen werde, womit die Gefahr bestünde, Minderheiten zu benachteiligen. Dem entgegen stehe jedoch, dass Bankberater Kredite basierend auf ermittelten Scores bewilligen können, anstatt subjektive Entscheidungen z. B. aufgrund des äußeren Erscheinungsbilds der Antragsteller zu fällen. Als positive Folge sahen die Teilnehmer die Möglichkeit, durch Zahlungsunfähigkeit der Bürger hervorgerufene Krisen im Voraus zu verhindern oder zumindest abzuschwächen. Ein weiterer positiver Aspekt von Scoring sei, dass präzisere Einschätzungen von Risiken zu günstigeren Angebote führen können. Es wurde diskutiert, dass je mehr Daten für die Erstellung des Angebots hinzugerufen würden, desto fairer das individuell angepasste Angebot ausfiele.

Die Teilnehmer waren sich bewusst, dass die herausgestellten positiven Aspekte in direktem Widerspruch zu den aufgeführten negativen Aspekten stehen. Als Vorschlag für eine höhere Akzeptanz des Scorings wurden Kompromisse vorgeschlagen. Bspw. könnten solche Daten, die vom einzelnen nicht zu beeinflussen sind (z. B. Erbkrankheiten und generell Daten aus dem Verwandtenkreis), beim Scoring nicht berücksichtigt werden.

6.6. Zusammenfassung

Die grundsätzliche Einstellung der Besucher zu Big Data war vorherrschend kritisch, über alle Szenarien hinweg wurden immer wieder Befürchtungen hinsichtlich des Datenschutzes geäußert. Aber es war auch zu beobachten, dass in Diskussionen durchaus Kompromissbereitschaft

vorhanden war. So wurde im Szenario des »Autobahnschützen« zwar erst die Überwachung von Autofahrern kritisiert, dann aber schnell eingeräumt, dass entsprechende Methoden zu Aufdeckung von Verbrechen durchaus erwünscht sind.

Akzeptanz für Big-Data-Lösungen erfordert eine größere Transparenz hinsichtlich ihrer Verwendung.

Dementsprechend ist als Ergebnis des Bürgerdialogs festzuhalten, dass eine allgemeine Zunahme von Datensammlungen und Analysen erst einmal abgelehnt wird. Wenn aber nachvollziehbare Gründe für solche Schritte vorhanden sind, und diese für den Bürger transparent dargestellt werden, dann existiert durchaus eine Mehrheit für eine Nutzung von Big Data. Vorherrschende Beispiele sind Kriminalitätsbekämpfung vom Aufdecken von Steuerdelikten über Kindesmisshandlung bis hin zur Terrorbekämpfung. Aber auch die Verbesserung medizinischer Behandlung wird als Grund für Big Data akzeptiert.

In den Fällen, in denen Big Data grundlegend abgelehnt wurde, kann davon ausgegangen werden, das diese Ablehnung unter anderem auf einem Missverständnis hinsichtlich des Einsatzes von Big Data basiert. So wurde bei IBM Watson die Unterstützung bei der Diagnose von Krankheiten negativ beurteilt, allerdings mit Argumenten, die auf ein Unbehagen in Anbetracht vollständig vom Computer erstellter Diagnosen schließen lassen. In der Praxis wird Big Data hier aber eher assistierend dem Arzt beiseite stehen und führt schon heute zu signifikant besseren Diagnosen. Hier gilt es, mehr Transparenz über die tatsächliche Nutzung von Big Data zu schaffen, um die Akzeptanz der Technologie zu steigern.

Abbildung 18: Ein verbreitetes Thema an den Tischen war das Misstrauen in diejenigen, die Big Data einsetzen. Auch wenn akzeptierte Anwendungen identifiziert werden konnten, bleibt immer die Befürchtung, dass die Daten auch anders eingesetzt werden.

7. ONLINEBEFRAGUNG

Begleitend zum Bürgerdialog wurde eine Onlineumfrage im November und Dezember 2014 durchgeführt. An der Umfrage nahmen 202 Personen teil. Davon entfielen 29 Prozent auf weibliche und 66 Prozent auf männliche Teilnehmer. Die restlichen Teilnehmer machten hierzu keine Angabe.

Bei der Altersverteilung war der größte Anteil der Teilnehmer 31 bis 50 Jahre alt. Dies waren 42 Prozent. Im Alter von 18 bis 30 Jahren waren 23 Prozent der Teilnehmer und zwischen 51 und 65 Jahren waren 27 Prozent. Immerhin 6 Prozent der Teilnehmer waren älter als 65 Jahre. Es nahm keine Person unter 18 Jahren teil. Berücksichtigt man die unterschiedliche Größe der Altersspannen, ergibt sich eine recht gleichmäßige Verteilung in dem Bereich von 18 bis 65 Jahren, wobei die Teilnehmer im Alter von 31 bis 50 Jahren pro Jahrgang mit kleinem Vorsprung am stärksten vertreten waren.

Der Fragebogen enthielt 27 themenspezifische Fragen, die sich auf drei Themenblöcke verteilten. Der erste Block befasste sich mit dem generellen Nutzen von Big Data (siehe Abschnitt 7.1 und 7.2). Im zweiten Block ging es um den Schutz der Privatsphäre im Kontext von Big Data (siehe Abschnitt 7.3 und 7.4). Der dritte Block umfasste allgemeine Fragen zum Nutzerverhalten der Teilnehmer im Internet (siehe Abschnitt 7.5). Nach der Auswertung der einzelnen Themenblöcke werden am Ende des Kapitels Querbezüge zwischen den Antworten dargestellt (siehe Abschnitt 7.6).

Die meisten folgenden Abbildungen zeigen die Häufigkeiten der in den Antworten genannten Begriffe. Die Teilnehmer haben die Begriffe durch Ankreuzen ausgewählt und konnten bei vielen Fragen weitere Begriffe selbst eingeben. Mehrfachnennungen waren erlaubt – außer bei Antwortmöglichkeiten, die sich offensichtlich gegenseitig ausschließen. Einige der Antwortmöglichkeiten sind in den Abbildungen aus Platzgründen gekürzt.

Manche Teilnehmer haben einzelne Fragen in den Themenblöcken übersprungen, sodass zu den meisten Fragen etwas weniger als 202 Antworten vorliegen. Wo es signifikant erscheint, wird die Menge der fehlenden Antworten genannt. Prozentangaben in den Abbildungen und im Text beziehen sich auf die Teilnehmer, die die jeweilige Frage beantwortet haben, sofern nicht ausdrücklich die Gesamtheit der Umfrageteilnehmer als Bezugsgröße genannt wird, was bei den Angaben zu fehlenden Antworten der Fall ist.

7.1. Beurteilung von Big Data

Die zentrale Frage war, ob die Bürger Big Data eher als Chance oder als Bedrohung wahrnehmen. Beantwortet wurde diese Frage (im Gegensatz zu den meisten anderen Frage) nicht durch Ankreuzen sondern mittels eines Schiebereglers. Eine Einstellung ganz links bedeutete »große Chance«, ganz rechts entsprechend »starke Bedrohung«. Die Antworten werden hier den Werten 0 (ganz links) bis 100 (ganz rechts) zugeordnet. Einen Überblick über die Verteilung der Antworten liefert Abbildung 19.

Interessanterweise fällt das Votum hier ausgeglichener aus, als dies die Betrachtung von Kommentaren im Internet (siehe Kapitel 8) vermuten lässt. Dennoch ist zu erkennen, dass mehr

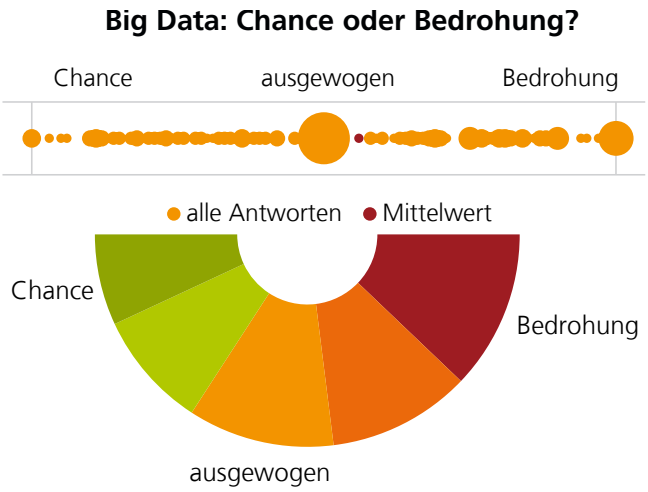

Abbildung 19: Nehmen die Befragten Big Data als Chance oder Bedrohung wahr? Die Punkte in der oberen Grafik stellen die Einstellungen des Schiebereglers dar, wobei die Größe eines Punktes der Häufigkeit der jeweiligen Antwort entspricht. In der unteren Grafik sind die Antwortwerte in gleich große Schritte gruppiert und als Kreissegmente gemäß der Gruppengröße dargestellt.

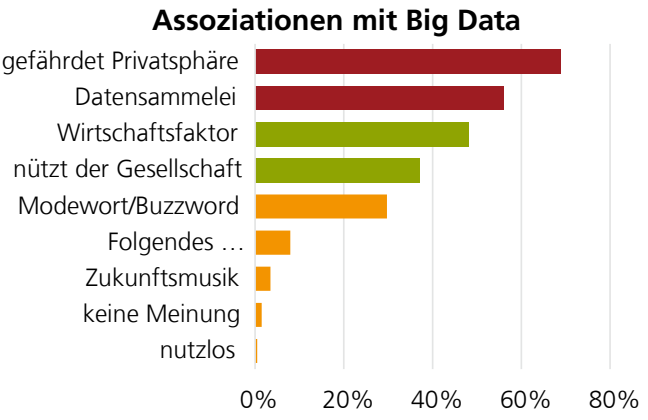

Abbildung 20: Was ist Big Data nach Meinung der Befragten?

Teilnehmer eher eine Bedrohung in Big Data sehen als eine Chance. Zudem wurde auf der Seite der Bedrohung häufiger eine besonders ausgeprägte Bewertung vergeben als auf der Seite der Chance. Einige Teilnehmer (13 %) nahmen keine Einstellung an dem Schieberegler vor, so dass der voreingestellte Wert von 50 als Antwort übernommen wurde. Der Durchschnitt aller Antworten liegt bei 56, was einer leichten Tendenz zur Bedrohung entspricht.

Eng verwandt mit der Frage nach Chance oder Bedrohung ist die Frage, was Big Data subjektiv für den Teilnehmer bedeutet. Abbildung 20 zeigt die Ergebnisse: Gut zwei Drittel haben »Gefahr für die Privatsphäre« angekreuzt, mehr als die Hälfte sieht darin eine »Datensammelei«. Die kritischen Stimmen überwiegen hier also ebenso wie in der vorherigen Frage. Aber auch positive Aspekte erreichten gute Werte. So sehen viele der Teilnehmer in Big Data einen »wichtige[n] Wirtschaftsfaktor« und bezeichnen Big Data als »nützlich für Gesellschaft und Bürger«. Unter »Folgendes … « gaben manche Befragte eine inhaltliche Beschreibung von Big Data und andere übten weitere Kritik.

26 % der Teilnehmer sehen in Big Data gleichzeitig Gefahr und Nutzen.

Erwähnenswert ist, dass 26 Prozent der Teilnehmer sowohl die Gefahr für die Privatsphäre als auch den Nutzen für die Gesellschaft sehen. Damit sind diese zwei Ansichten (annähernd) stochastisch unabhängig, d. h., wenn ein Teilnehmer eine der Ansichten vertritt, hat das keinen Einfluss darauf, ob er die andere Ansicht auch vertritt. Es kann davon ausgegangen werden, dass die Teilnehmer, die beide Ansichten vertreten, einen kritischen und risikobewussten Umgang mit Big Data fordern, den Einsatz aber nicht grundsätzlich ablehnen.

7.2. Vertrauen und Rechtfertigung

Wem die Befragten hinsichtlich der Nutzung von Big Data vertrauen, zeigt Abbildung 21. Dabei ist insgesamt eine ausgeprägte Zurückhaltung zu sehen: Mehr als jeder Dritte vertraut keiner Branche. Forschung und Wissenschaft sind deutlicher Spitzenreiter, wobei hier über die

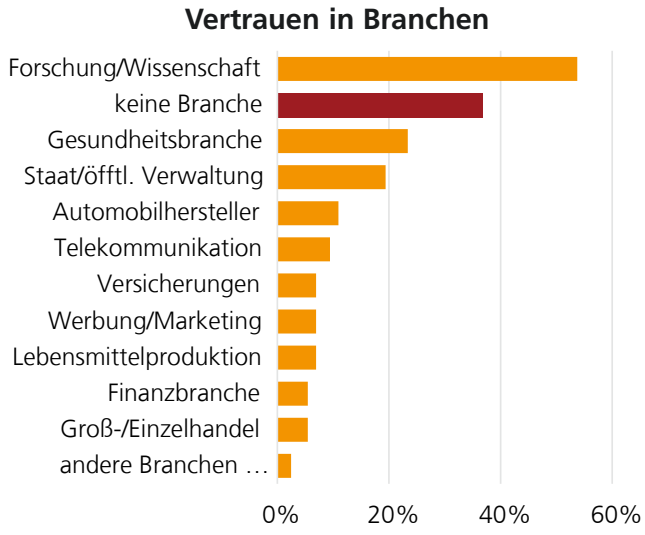

Abbildung 21: Welchen Branchen vertrauen die Teilnehmer bezüglich des Einsatzes von Big Data?

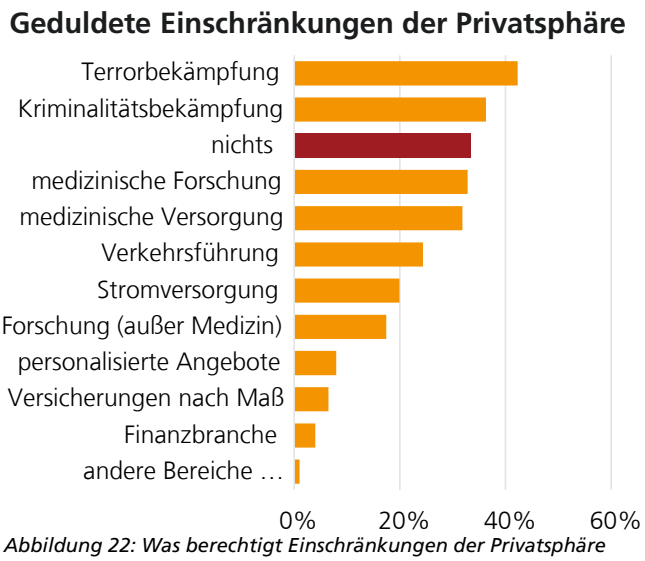

Abbildung 22: Was berechtigt Einschränkungen der Privatsphäre durch Big Data?

Hälfte der Teilnehmer ihr Vertrauen bekundet. Knapp jeder Fünfte vertraut dem Staat und der öffentlichen Verwaltung. Die Gesundheitsbranche hat unter den Industriezweigen den höchsten Vertrauenszuspruch von gut 20 Prozent. Den anderen aufgeführten Branchen wird beachtlich wenig Vertrauen ausgesprochen.

Eine Kritik an Big Data ist, dass diese Technologie die Privatsphäre potenziell verletzen könnte. Eine Frage war daher, ob es Anwendungen gebe, die eine Einschränkung der Privatsphäre rechtfertigten. Die meistgenannten Anwendungen waren Terrorismus- und Kriminalitätsbekämpfung, wie in Abbildung 22 zu sehen ist. Aus den Teilnehmerkommentaren zu dieser Frage geht jedoch hervor, dass diese Zustimmung nur mit Vorbehalt erfolgt. So dürfe eine Datenerhebung und -auswertung zur Terrorismus- oder Kriminalitätsbekämpfung beispielsweise »nur fallbezogen mit möglichst hoher Zugangssicherheit« erfolgen. Darüber hinaus sieht ein gutes Drittel aller Teilnehmer keine Anwendung als so wichtig an, dass dafür die Privatsphäre eingeschränkt werden dürfe. Dies zeigt, dass es sowohl ein starkes Sicherheitsbedürfnis als auch ein starkes Bedürfnis nach Privatsphäre gibt. Big-Data-Anwendungen zur Gefahrenabwehr müssen deshalb in einem transparenten politische Prozess diskutiert und – im Falle einer Einigung auf den Einsatz – im Sinne eines minimalen Privatsphäreneingriffs konzipiert sowie unter einem hohen Datenschutzstandard umgesetzt werden.

Nur einer von 20 Teilnehmern vertraut der Finanzbranche hinsichtlich Big Data.

Bekämpfung von Terrorismus und Kriminalität wird als stärkste Berechtigung für die Einschränkung der Privatsphäre gesehen.

Interessant ist hier auch, dass zwar in der vorhergehenden Frage der Medizinbranche mehr Vertrauen als dem Staat ausgesprochen wurde, hier nun aber Aufgaben des Staates vor denen der Medizin liegen. Ebenso werden Einschränkungen der Privatsphäre zugunsten der medizinischen Forschung und Versorgung eher akzeptiert als Einschränkungen zugunsten der Forschung auf anderen Gebieten, auch wenn in der

9 von 10 Teilnehmern sehen multinationale Internetunternehmen und Softwarehersteller als eine Gefahr für die Privatsphäre.

vorhergehenden Frage die Medizinbranche bei Weitem nicht das Vertrauen von Forschung und Wissenschaft erreicht hat.

Eine Reihe von Nutzern von Big-Data-Methoden und -Technologien stellt nach Meinung der Befragten eine Bedrohung für die Privatsphäre der Bürger dar (siehe Abbildung 23). Dabei sind »multinationale Internetunternehmen und Softwarehersteller« die Spitzenreiter. Nur jeder zehnte der Teilnehmer sieht in ihnen demnach keine Gefahr für die Privatsphäre. Auch gegenüber ausländischen und deutschen Geheimdiensten herrscht ein ausgeprägtes Misstrauen. Selbst dem Staat misstraut mehr als die Hälfte der Teilnehmer. Weitere genannte Gefährder waren Versicherungen, die Werbebranche, Banken und die organisierte Kriminalität. Oft wurde hier auch ein allgemeines Misstrauen in Unternehmen ausgesprochen. Nur eine deutliche Minderheit sieht keine Gefährdung der Privatsphäre durch Nutzer von Big Data.

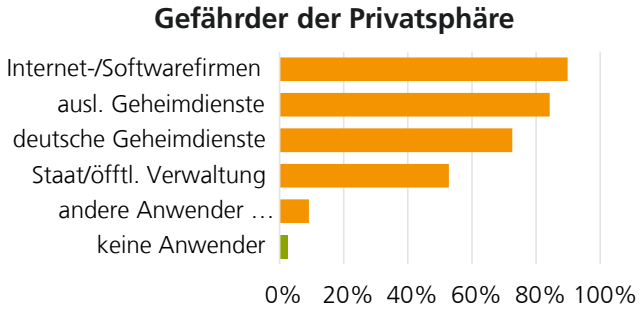

Abbildung 23: Welche (potenziellen) Nutzer von Big Data gefährden nach Meinung der Befragten die Privatsphäre von Bürgern?

7.3. Datenschutzrechte

Die Umfrageteilnehmer wurden gefragt, welche Aspekte des Datenschutzes ihnen bekannt sind, welche nach ihrer Ansicht ausreichend umgesetzt werden und zu welchen sie eine Unterstützung bei der Umsetzung wünschen. Abbildung 24 gibt einen Überblick über die Ergebnisse.

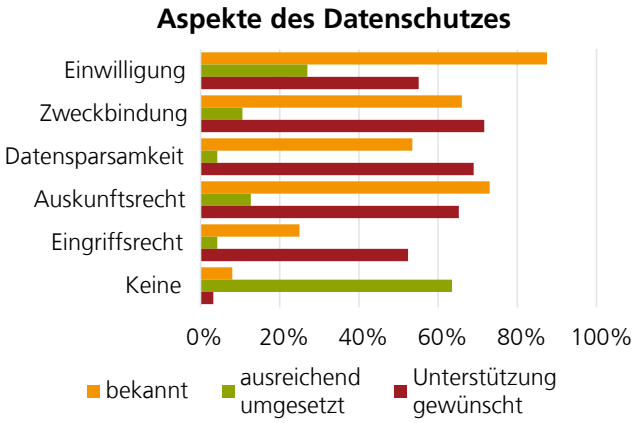

Abbildung 24: Bekanntheit und Umsetzung von Datenschutzaspekten.

Mehr als der Hälfte der Teilnehmer sind zumindest die Einwilligung, das Auskunftsrecht, die Zweckbindung und die Datensparsamkeit bekannt. Das Eingriffsrecht ist nur einem Viertel bekannt, und immerhin 8 Prozent kennen keines der genannten Rechte.

Ausreichend umgesetzt ist aus Sicht der Befragten nur ein geringer Teil der Aspekte. Fast zwei Drittel der Teilnehmer sind sogar der Meinung, dass kein Aspekt ausreichend umgesetzt ist. Nur die Einwilligung kommt auf einen Wert von über 25 Prozent, während die übrigen Aspekte nicht einmal 15 Prozent erreichen.

63 % der Teilnehmer sehen keinen Datenschutzaspekt ausreichend umgesetzt.

Die meiste Unterstützung wird bei der Zweckbindung gewünscht, hier sind es 72 Prozent. Aber auch bei den übrigen Aspekten wünscht mehr als die Hälfte der Teilnehmer Unterstützung.

Zusammengefasst kann gesagt werden, dass die rechtlichen Grundlagen des Datenschutzes einem guten Teil der Befragten bekannt sind, jedoch Unzufriedenheit hinsichtlich ihrer Umsetzung besteht und die Mehrheit sich Unterstützung dabei wünscht, die Rechte durchzusetzen. Eine gewisse Unsicherheit bezüglich des Datenschutzrechts ist daran zu erkennen, dass ein Anteil von 6 bzw. 7 Prozent der Umfrageteilnehmer die beiden letztgenannten Fragen übersprungen hat.

Für 74 % der Teilnehmer scheitert Datenschutz bei der Durchsetzung.

Zu den vorhergehenden Ergebnissen passt auch, dass etwa drei Viertel der Teilnehmer der Meinung sind, dass Datenschutz an der Durchsetzung der Regeln scheitert. Zusätzlich meint die Hälfte der Teilnehmer, dass die Regelungen zum Datenschutz geändert werden müssen, um effektiv zu sein. Eine Minderheit von 8 Prozent ist der Meinung, dass die bestehenden Regeln ihren Zweck erfüllen.

Der Schutz der Privatsphäre soll nach Möglichkeit international geregelt werden: 76 Prozent wünschen sich globale Datenschutzregeln, wobei allerdings 70 Prozent fordern, dass dabei keine Kompromisse hinsichtlich des hohen Standards des deutschen Datenschutzes gemacht werden.

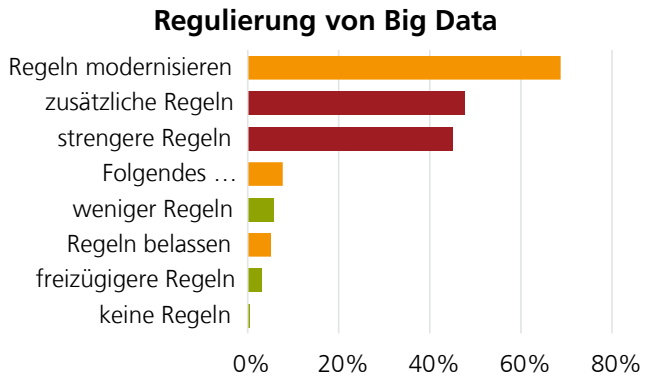

Abbildung 25: Wünsche zur Regulierung von Big Data hinsichtlich des Datenschutzes.

Der Schutz der Privatsphäre sollte vom einzelnen Bürger (87 %) selbst durchgeführt, vom Staat (86 %) durchgesetzt und von Unternehmen (58 %) beachtet werden. Vergleichsweise niedrige 45 Prozent der Teilnehmer fordern, dass unabhängige Dritte auf den Schutz der Privatsphäre achten sollen. In den Freitextantworten wurden auch die Forschung, die Justiz, die Datenschutzbeauftragten der Länder sowie »jeder« bzw. »alle« genannt.

Wenn es um den Einsatz von Big-Data-Methoden geht, sind gut zwei Drittel der Teilnehmer der Meinung, dass die in Deutschland geltenden Gesetzte modernisiert werden müssen (siehe Abbildung 25). Knapp die Hälfte fordert zusätzliche Reglungen, und fast ebenso viele fordern eine Verschärfung der Regeln. Nur 5 Prozent sind mit dem herrschenden Recht zufrieden. Ebenso in der Unterzahl sind diejenigen, die eine Liberalisierung fordern.

69 % der Teilnehmer wünschen eine Modernisierung der Gesetze.

Anonymisierung ist ein verbreitetes Mittel zum Schutz der Privatheit. Dabei ist ein wichtiger Faktor, wie groß eine Gruppe von Personen mit identischen Identifizierungsmerkmalen sein muss, damit von anonymisierten Daten gesprochen werden kann. (k-Anonymität, siehe Abschnitt 3.3). Daher wurde die folgende Frage gestellt: *»In der deutschen Rechtsprechung gilt man bereits als anonym, wenn man innerhalb eines Kreises von mindestens 5 Personen nicht eindeutig identifizierbar ist, also bspw. innerhalb eines Fünf-Personen-Haushalts. Unter wie vielen Personen empfinden Sie sich als ausreichend anonym?«*

Als ausreichend anonym werden Gruppen der Größe 100 empfunden.

Die Antworten, welche in Abbildung 26 gezeigt werden, sind breit gestreut – teilweise mit extremen Forderungen. Ein paar Teilnehmer gingen bis zum Maximum des Eingabefeldes und nannten zehn Millionen (oder 9.999.999) Personen. Ein paar andere Teilnehmer gaben die Zahl 1 an, wobei manche davon vermutlich zum Ausdruck bringen wollten, dass sie Anonymisierung für unwichtig erachten, während den übrigen wahrscheinlich die Frage unklar war. Der Median der Antworten liegt bei 25, d. h., dass die Hälfte der Antwortenden eine Gruppengröße von 25 oder weniger als ausreichend betrachtet, während die andere Hälfte mindestens eine Gruppengröße von 25 fordert. 15 Prozent der Antwortenden gaben die Größe 100 an. Für insgesamt 80 Prozent ist bei einer Gruppengröße von 100 die Privatsphäre ausreichend geschützt, da sie 100 oder weniger angaben.

12 Prozent der Umfrageteilnehmer haben diese Frage nicht beantwortet, was darauf hindeutet, dass die Frage schwierig ist. Auch aus den Kommentaren geht hervor, dass Anonymisierung ein schwieriges Thema ist. Daneben werfen manche Teilnehmer ein, dass für sie die Wahl der Gruppengröße von der Sensibilität der erfassten Daten abhänge. Die Diversität der Personengruppe wird als weitere relevante Größe angeführt.

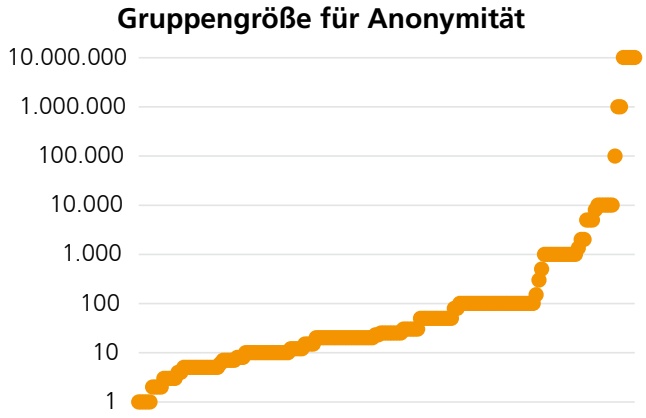

Abbildung 26: Wie viele Personen müssen für einen Datensatz in Frage kommen, damit dieser als anonym angesehen werden kann? Dargestellt sind die Antworten der Teilnehmer sortiert in aufsteigender Reihenfolge.

7.4. Scoring und Profiling

Die Methoden von Unternehmen, die Scoring und Profiling betreiben, stehen oft in der Kritik (siehe Abschnitt 4.2). Bei der in Abbildung 27 dargestellten Frage sollten die Teilnehmer bestimmen, wie weit entsprechende Systeme in die Vergangenheit blicken dürfen sollten. Dabei stimmten 39 Prozent für »gar nicht« und weitere 26 Prozent für »maximal ein Jahr«. Nur 7 Prozent der Teilnehmer sprachen sich für eine unbegrenzte Sicht in die Vergangenheit aus. Die Befragten gaben in ihren Kommentaren jedoch an, dass ihre Einschätzung auch vom Anwendungszweck abhänge.

Eine andere Frage bezog sich auf Versicherungsbeiträge:
»Durch Scoring und Profiling können Versicherungen Risiken bei individuellen Personen besser identifizieren. Die Folge können leicht

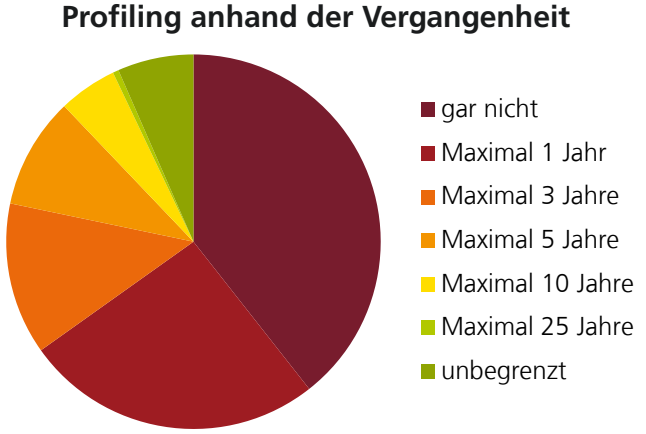

Abbildung 27: Wie lange sollten Scoring- und Profiling- Systeme in die Vergangenheit blicken dürfen?

niedrigere Kosten für die Allgemeinheit und stark höhere Kosten für risikobehaftete Individuen sein. Risiken können hier verletzungsintensive Hobbys sein, aber auch in der Familie häufig auftretende Krankheiten. Wie ist Ihre Meinung hierzu?«

Hier waren 40 Prozent der Teilnehmer dafür, dass das Risiko und die damit verbundenen Kosten allgemein von der Gemeinschaft getragen werden sollten. Die Mehrheit mit 53 Prozent bevorzugte, selbst verschuldete Risiken vom Individuum tragen zu lassen. Nur 7 Prozent waren dafür, dass Risikoträger die Kosten für jegliches eigenes Risiko tragen sollten, ob selbst verschuldet oder nicht.

Das ist eine knappe Mehrheit für eine konkrete Nutzung von Big Data durch die Versicherungsbranche, der in den übrigen Fragen nur wenig Vertrauen ausgesprochen wurde. Dies kann als ein Indiz dafür gesehen werden, dass ein Einsatz von Big Data inklusive einem Eingriff in die Privatsphäre durchaus akzeptiert wird, wenn die Bürger dadurch auch einen konkreten Nutzen, in diesem Fall individuell potenziell niedrigere Versicherungskosten, vor Augen haben. Die Brisanz dieses Themas lässt sich jedoch daran erkennen, dass zu dieser Frage die meisten Kommentare abgegeben wurden. Darunter waren Kommentare, die das Solidaritätsprinzip vehement verteidigten, und solche, die Zweifel an einer praktikablen Grenzziehung zwischen selbstverschuldeten und nicht selbstverschuldeten Risiken äußerten.

Jeder Zweite möchte selbst verschuldete Versicherungsrisiken anderer nicht tragen.

7.5. Nutzerverhalten

Nutzer beeinflussen durch ihr Verhalten bei der Internetnutzung maßgeblich, wer welche Daten über sie sammelt bzw. sammeln kann. Deshalb hat der Fragebogen Aspekte der Internetnutzung abgefragt, etwa zur Nutzung von Suchmaschinen, E-Mail-Diensten, Chat-Diensten und sozialen Medien, zur Auseinandersetzung mit den Nutzungsbedingungen von Diensten sowie zum Einkaufen im Internet.

In Abbildung 28 ist zu sehen, dass Google mit Abstand die meistgenutzte Suchmaschine ist. Sie wird von beinahe 9 von 10 Teilnehmern genutzt. An zweiter Stelle kommt die privatsphärenfreundliche Suchmaschine DuckDuckGo, die von knapp einem Drittel der Teilnehmer genutzt wird. An dritter Stelle kommt die Suchmaschine Bing, welche von jedem fünften Teilnehmern genutzt wird. Insgesamt verwenden 45 Prozent der Teilnehmer privatsphärenfreundliche Suchmaschinen (neben DuckDuckGo auch Ixquick, Startpage und ein Teil der Suchmaschinen, die als »andere . . . « zusammengefasst sind) und immerhin 11 Prozent verwenden ausschließlich solche Suchmaschinen – dies sind fast alle Teilnehmer, die auf Google verzichten. Somit scheint das Bedürfnis nach Privatheit die wesentliche

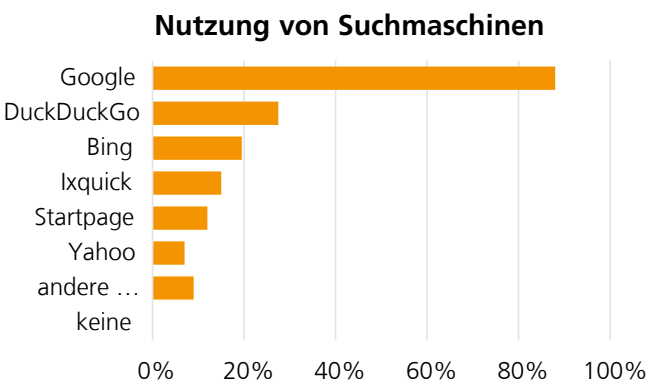

Abbildung 28: Welche Suchmaschinen nutzen die Teilnehmer?

Motivation für einen Verzicht auf Google zu sein. Umgekehrt nutzen 42 Prozent der Teilnehmer ausschließlich Google.

Mehr als die Hälfte (54 %) der Teilnehmer gibt an, E-Mail-Verschlüsselung zu nutzen. Diese Quote ist überraschend hoch. Vielleicht wurde hier teilweise E-Mail-Verschlüsselung mit der aus Sicherheitsaspekten notwendigen Transportverschlüsselung verwechselt, da letztere durch die Kampagne »E-Mail made in Germany« (http: //www.e-mail-made-in-germany.de/) in den Medien häufig Thema war, jedoch nicht immer klar von der eigentlichen E-Mail-Verschlüsselung abgegrenzt wurde. Allerdings zeigen die Antworten der Nutzer entsprechender E-Mail-Dienste im Vergleich zu anderen Teilnehmern keine auf eine solche Verwechslung hindeutende Verschiebung der Häufigkeiten. Ein anderer Erklärungsversuch der Quoten zur Nutzung von E-Mail-Verschlüsselung ist, dass möglicherweise überdurchschnittlich viele Umfrageteilnehmer eine gewisse Affinität zur IT-Sicherheit haben.

In der Umfrage wurden die Teilnehmer gefragt, ob sie die AGB von Onlinediensten lesen und ob sie aufgrund der Bedingungen schon auf die Nutzung von Diensten verzichtet haben. Eine deutliche Mehrheit von 62 Prozent liest AGB teilweise, weitere 16 Prozent lesen diese immer. Die meisten Teilnehmer (84 %) haben aufgrund der Nutzungsbedingungen schon auf Dienste verzichtet. Dies trifft sogar auch auf mehr als die Hälfte (56 %) der Teilnehmer zu, die AGB nicht lesen.

7.6. Zusammenhänge

Wie am Anfang dieses Kapitels beschrieben, konnten die Teilnehmer mittels Schieberegler ihre Empfindung ausdrücken, ob sie in Big Data eher eine Chance oder Bedrohung sehen (siehe Abbildung 19). Im Folgenden wird diese Einschätzung anderen Antworten gegenübergestellt. Dabei stellen die Punkte in den nachfolgenden Abbildungen die Einstellungen des Schiebereglers dar, d. h. weiter links bedeutet mehr Chance und weiter rechts bedeutet mehr Gefahr. Die Größe der Punkte entspricht der Anzahl gleicher Antworten.

Betrachtet man in Abbildung 29 die Beziehung zwischen dem Alter der Teilnehmer und deren Einstellung zu Big Data, dann fällt auf, dass die Altersgruppe über 65 Jahren die positivste Einstellung aufweist und als einzige Gruppe im Durchschnitt auf der Seite der Chance liegt. Am kritischsten ist die Gruppe von 31 bis 50 Jahren. Bei allen Gruppen gibt es allerdings eine breite Streuung der Antworten.

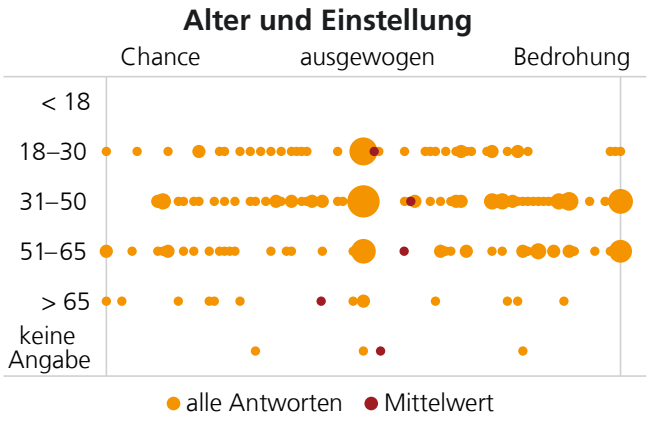

Abbildung 29: Wie steht das Alter der Befragten mit der Einstellung zu Big Data in Beziehung?

Die Nutzung von Suchmaschinen (siehe Abbildung 28) korreliert auf naheliegende Weise mit der Haltung zu Big Data: In Abbildung 30 ist zu erkennen, dass privatsphärenfreundliche Suchmaschinen Nutzer haben, die im Durch-

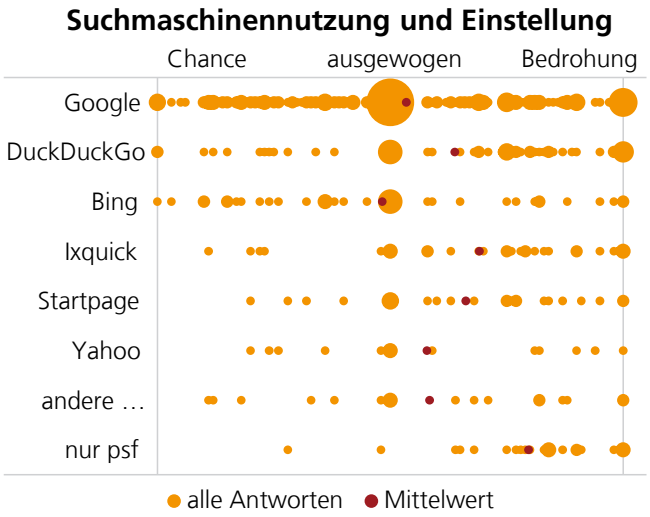

Abbildung 30: Wie steht die Nutzung von Suchmaschinen mit der Einstellung zu Big Data in Beziehung? Die Bezeichnung »nur psf« steht hier für Personen, die ausschließlich privatsphärenfreundliche Suchmaschinen nutzen.

schnitt eher kritisch zu Big Data stehen. Jedoch gibt es auch bei diesen Nutzern eine breite Streuung bei der Einstellung zu Big Data. Lediglich für die Personen, die ausschließlich privatsphärenfreundliche Suchmaschinen nutzen, überwiegt die Bedrohung meist klar, sodass die Streuung nicht ganz so stark ausfällt.

Die Ansichten zur gesetzlichen Regulierung (siehe Abbildung 25) stehen in einem deutlichen statistischen Zusammenhang mit der Einstellung zu Big Data, wie in Abbildung 31 zu sehen ist. Personen, die weniger Regulierung wünschen, sehen im Durchschnitt Big Data eher als Chance, während Personen, die eine Verschärfung von Regeln wünschen, im Durchschnitt eher eine Bedrohung sehen. Teilnehmer, die unter »Folgendes … « Regulierungswünsche selbst formuliert haben, haben ebenfalls eine überwiegend kritische Einstellung. Bei dieser Frage ist wieder eine breite Streuung innerhalb der Gruppen zu erkennen, sodass manche Teilnehmer, die eine Verschärfung der Regeln fordern, neben den Gefahren von Big Data auch große Chancen sehen, während andere trotz der Forderung nach mehr Freizügigkeit auch die Risiken wahrnehmen.

Ein Zusammenhang zeigt sich auch zwischen der Haltung zur Verteilung von Risiken bei Versicherungen, welche weiter oben bereits aufgeführt wurde, und der Einstellung zu Big Data. Diejenigen, die für ein Tragen der Risiken durch die Gemeinschaft sind, sehen tendenziell mehr Bedrohung durch Big Data als der allgemeine Durchschnitt. Wer jedoch für ein Tragen aller Risiken durch das Individuum ist, sieht durchschnittlich etwas mehr Chance als Bedrohung. Diejenigen, welche nur selbst verschuldete Risiken vom Individuum tragen lassen wollen, haben eine ähnliche Verteilung wie die Gesamtheit der Teilnehmer. In allen Gruppen ist die Streuung wieder groß.

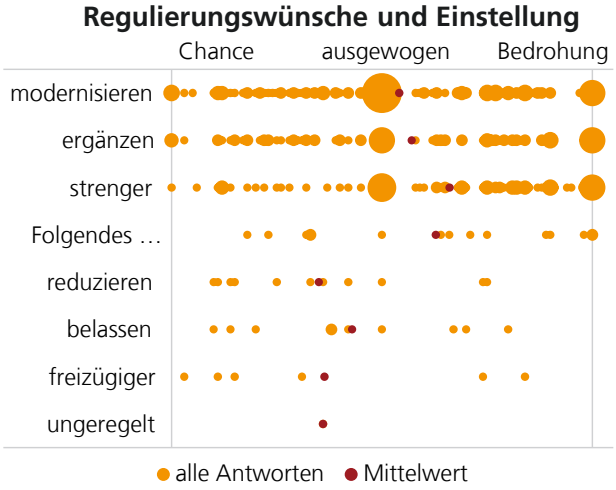

Abbildung 31: Wie steht die Haltung zur gesetzlichen Regulierung von Big Data mit der Einstellung zu Big Data in Beziehung?

Der Fragebogen war in drei Themenblöcke unterteilt, die sich den Chancen von Big Data, dem Privatsphärenschutz und dem Nutzerverhalten widmeten (siehe oben). Für eine aggregierte Darstellung der Ansichten der Teilnehmer wurde bei der Auswertung für jeden Teilnehmer zu jedem Themenblock ein Score bestimmt. Die Scores sind gewichtete Summen der angekreuzten Antworten. Der erste Block steht für den Score »Chance«, der zweite

Block für den Score »Gefahr« und der dritte Block für den Score »Selbstdatenschutz«.

Bei dem Score »Chance« gibt es viele Teilnehmer, die einen sehr niedrigen Wert erzielen, aber kaum Teilnehmer mit einem mittleren oder hohen Wert. Umgekehrt gibt es bei dem Score »Gefahr« viele Teilnehmer mit einem hohen Wert, aber kaum Teilnehmer, die hier einen niedrigen Wert haben. Bei dem Score »Selbstdatenschutz« gibt es weder starke Ausschläge in die eine noch in die andere Richtung. Die Werte befinden sich alle in einem mäßigen Bereich, wobei es eine leichte Tendenz zu wenig Selbstdatenschutz gibt.

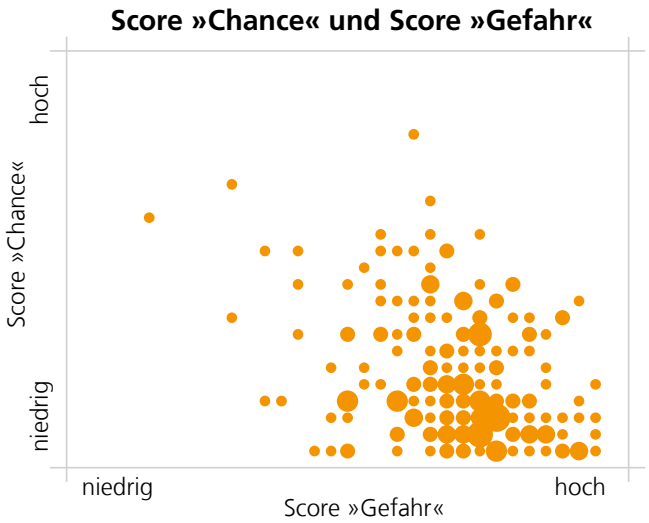

Abbildung 32: Gegenüberstellung von Score »Chance« und Score »Gefahr«. Die Größe der Punkte entspricht der Anzahl der Teilnehmer mit den zugehörigen Scores.

Betrachtet man die Beziehung von Score »Chance« und Score »Gefahr« in Abbildung 32, so fällt eine flächige Streuung auf, d. h. es gibt keinen starken Zusammenhang zwischen diesen beiden Größen. Dennoch gibt es eine leichte Tendenz, dass ein höherer Score »Gefahr« zu einem niedrigeren Score »Chance« gehört und umgekehrt.

Ähnliche Zusammenhänge bestehen auch zwischen anderen Größen. Etwas stärker ausgeprägt ist der Zusammenhang zwischen dem Score »Selbstdatenschutz« und der Differenz aus Score »Gefahr« und Score »Chance«. Die Streuung zeigt hier, dass bei den Teilnehmern die Ansichten zu Big Data und Privatsphäre unterschiedlich stark das Nutzerverhalten beeinflussen.

Am stärksten ist der Zusammenhang zischen der Differenz aus Score »Gefahr« und Score »Chance« und der Einstellung des bereits ausführlich behandelten Schiebereglers. Aber selbst hier gibt es eine große Streuung.

Insgesamt ist also festzustellen, dass zwischen der Einstellung zu Big Data, den Wünschen nach Regulierung und dem eigenen Verhalten eine gewisse Korrelation vorliegt, aber die Grundansichten der Teilnehmer nicht einfach auf eine bipolare Achse abgebildet werden können, da es quer zu dieser Achse eine breite Streuung gibt. So gibt es durchaus Teilnehmer, die eine Nutzung der Möglichkeiten von Big Data befürworten und gleichzeitig den Schutz der Privatsphäre für wichtig erachten.
Beide Interessen, die hier aufeinandertreffen, müssen ernst genommen werden. Eine allgemeine Akzeptanz ist nur zu erreichen, indem man Big Data mit einem effektiven Datenschutz vereinbart.

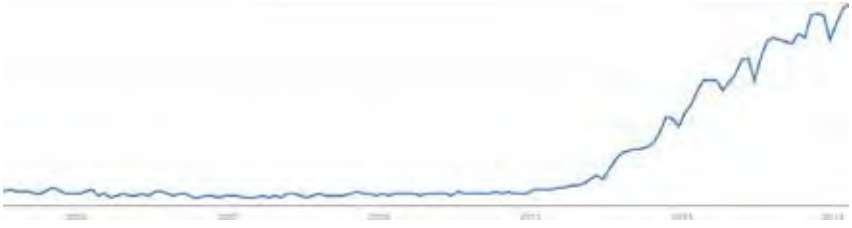

Abbildung 33: Globales Interesse an dem Suchbegriff »Big Data« im zeitlichen Verlauf gemäß Google Trends. Quelle: Google, http://www.google.de/trends/explore#q=big%20data

8. ÖFFENTLICHE WAHRNEHMUNG

Big Data findet nicht erst seit dem NSA-Skandal ein wachsendes Interesse in der Öffentlichkeit. Betrachtet man den Graphen in Abbildung 33, der die Häufigkeit von Suchanfragen zu Big Data in Google ausdrückt, so lässt sich weltweit ein signifikanter Zuwachs bereits ab dem Jahr 2011 erkennen. Ähnlich sieht der Verlauf für Deutschland aus.

Um die öffentliche Wahrnehmung von Big Data in Deutschland zu erheben, wurden im Rahmen dieser Studie mehrere Untersuchungen durchgeführt. Als Datengrundlage wurden Texte zum Thema Big Data gesammelt und hinsichtlich ihrer Ausprägung untersucht. Insgesamt wurden ca. 12.000 deutschsprachige Texte aus unterschiedlichen Internetquellen zusammengetragen. Zum einen wurden Leserkommentare und Tweets bezüglich Stimmung und Thematik analysiert (Abschnitt 8.1); zum anderen wurden Presseartikel bekannter Onlineportale hinsichtlich ihrer Einstellung zu Big Data untersucht (Abschnitt 8.2). Die Untersuchungsergebnisse beschreiben numerisch und nominal, wie die Gesellschaft das Thema Big Data wahrnimmt. Ein besonderer Fokus wird auf die Einschätzung von Chancen und Risiken gelegt.

Da sich die gesammelten Texte in ihrer Rohform nicht für eine Analyse eigneten, mussten diese zunächst aufbereitet werden. Neben dem Entfernen von Zusatzinformationen wie bspw. Markup-Tags (HTML, XML, CSS etc.) wurden Texte von zu geringer Länge herausgefiltert. Ebenfalls ignoriert wurden Texte, welche sehr ähnlich zu bereits aufgenommenen waren (sogenannte Near-Duplicates), wobei 80 Prozent übereinstimmende Wörter als Schwellenwert festgelegt wurden. Infolgedessen standen für die Analyse 1.595 Tweets, 440 Nachrichtenartikel und 7.408 Kommentare zur Verfügung.

8.1. Tweets und Leserkommentare

Kommentarfunktionen und Tweets bieten Bürgern eine effektive Möglichkeit, ihre Meinung öffentlich kundzutun. Deshalb wurden für die Analyse Kommentare zu Online-Artikeln mit Bezug zu Big Data sowie Tweets mit dem Hashtag #bigdata erhoben. Die Kommentare gehören zu Artikeln aus den Jahren 2013 und 2014, während die Tweets im November und Dezember 2014 erstellt wurden.

8.1.1. Analysemethodik

Zu den verwendeten Analysemethoden zählten Klassifikations- und Clusteringverfahren (siehe Abschnitt 2.3). Anhand computerlinguistischer Methoden wurden unter anderem Wortarten (Nomen, Adjektive etc.) und Phrasentypen (Nominalphrasen etc.) ermittelt.

Klassifikationsverfahren: Mittels Klassifikationsverfahren wurden die Texte hinsichtlich ihrer Stimmung in die Klassen »positiv«, »negativ« und »neutral« unterteilt. Dazu wurden drei Wortlisten eingesetzt, welche einige Tausend positive, negative bzw. neutrale Wörter enthielten. Zu jedem Wort innerhalb einer Liste war der Wortstamm, die Wortart sowie ein Gewicht zwischen 0 und 1 angegeben.

Um zu bewerten, zu welcher der drei Klassen ein Text nun gehört, wurde aus den Gewichtungen derjenigen Wörter, die in dem Text gefunden wurden, jeweils eine Summe für positive

und negative Stimmung gebildet. Bei Adjektiven wurde hierbei als zusätzliches Gewicht 1 hinzuaddiert. Die Idee dahinter ist, dass mittels Adjektiven positive oder negative Emotionen ausgedrückt werden können. Diese wiederum können die Stimmung besser ausdrücken als Verben oder Nomen. Um sich einen Eindruck des Verfahrens machen zu können, wird in Tabelle 1 der Satz »Unerwünschte Werbung hat in den vergangenen Jahren dramatisch zugenommen.« mit entsprechenden Gewichtungen aufgezeigt. Da keines der neun Wörter in der Liste positiver Wörter enthalten ist, jedoch »Unerwünschte« und »dramatisch« als negativ klassifiziert werden, wird der Satz als insgesamt als negativ eingestuft.

Tabelle 1: Stimmungsanalyse eines Beispielsatzes

Stimmung	Bewertung der Wörter									Gesamt
Unerwünschte Werbung hat in den vergangenen Jahren dramatisch zugenommen.										
Positiv	0	0	0	0	0	0	0	0	0	0
Negativ	0,34+1	0	0	0	0	0	0	0,37+1	0	2,71

Texte, die größenteils aus Sarkasmus bestanden, wurden mit einer Heuristik (annähernde Gleichverteilung von positiv und negativ geladenen Phrasen je Text) ermittelt und für die weitere Analyse nicht berücksichtigt, sodass die verbliebenen Texte inhaltsgetreu in die drei Klassen fielen.

Clusteringverfahren: Mittels eines Clusteringverfahrens wurden die Texte in Gruppen aufgeteilt, um thematisch unterschiedliche Gruppen zu identifizieren. Dafür wurden Nomen als thematisch identifizierende Merkmale ausgewählt.

8.1.2. Erkenntnisse zur Stimmung

Zusammengefasst brachte die Stimmungsanalyse die Erkenntnis, dass sich Bürger im Internet überwiegend negativ zum Thema Big Data äußern. Die Chancen, welche die Bürger im Kontext von Big Data erkannt haben, unterlagen deutlich den damit verbundenen Risiken.

Es fiel zunächst auf, dass über die Hälfte der gesammelten Tweets von den eingesetzten Methoden als positiv bewertet wurden. Eine genauere Betrachtung der entsprechenden Texte zeigte jedoch, dass diese häufig mit Werbung in Form von URLs und Hashtags behaftet waren. Hierbei kann davon ausgegangen werden, dass die Mehrheit der als positiv erkannten Texte nicht von Privatpersonen verfasst wurde, sondern von kommerziellen Anbietern zum Bewerben von Software-Lösungen, Büchern, Seminaren etc. Nach der Einschränkung auf Tweets, die keine Werbung enthalten, wurden nahezu alle Texte als negativ oder neutral bewertet.

Bei den Kommentaren wurde die überwiegende Mehrheit (ca. 85 %) als negativ gewertet, während ein kleiner Teil (ca. 10 %) als neutral und der Rest (ca. 5 %) als positiv gekennzeichnet wurde. Zu ergänzen ist, dass bei den Kommentaren keine Werbung gefunden wurde.

8.1.3. Erkenntnisse zur Thematik

Bei den Tweets wurden insgesamt sechs Cluster identifiziert, die den folgenden Themengruppen zugeordnet werden können: Versicherungswesen, Marketingbranche, Gesundheit und Medizin, Bildung, Gesellschaft sowie Terror- und Verbrechensbekämpfung. Die meisten Tweets konnten

den Clustern »Versicherungswesen« und »Marketingbranche« zugewiesen werden.

Bei den Kommentaren wurden dagegen fünf bis acht Cluster ausfindig gemacht, die thematisch gesehen teilweise nur schwer zu unterscheiden waren. Infolgedessen teilten sich manche Cluster mehrere Themengebiete. Die Themengebiete erstreckten sich hierbei über Politik, Wissenschaft, Werbebranche, Gesellschaft, innovative Technologien sowie technisches Expertenwissen.

Abbildung 34: Wortwolke des Clusters »Gesellschaft« der untersuchten Kommentare.

Der Cluster »Gesellschaft« im Kontext der Kommentare fiel durch seine markant negative Form auf. Seine Schlagworte sind als Wortwolke in Abbildung 34 zu sehen. Eindeutig zu erkennen sind hierbei die negativ behafteten Wörter »Überwachung«, »Geheimdienste« und »Kontrolle«, die gemäß ihrer Häufigkeit in den untersuchten Kommentaren aus der Wolke hervorstechen.

8.1.4. Originalton

Um sich ein näheres Bild von den Meinungen der Bürger zu machen, ohne auf die eingesetzten abstrahierenden Verfahren zurückzugreifen, sind im Folgenden einige kurze und prägnante Sätze wortwörtlich aufgeführt, die in den Kommentaren zu den verschiedenen Artikeln vorgefunden wurden. Ein kurze Einleitung zu Beginn jedes Zitats versucht den Kontext dieser Sätze zusammenzufassen.

In der Vernetzung von Geräten, Häusern, Fahrzeugen, Städten etc. sehen IT-Konzerne die Möglichkeit, viele Dinge einfacher und komfortabler zu machen und viele Aufgaben effizienter zu lösen. Viele Bürger hierzulande scheinen diese Vision nicht zu teilen und äußern sich darüber besorgt, wütend, frustriert und verärgert:
»Das Jahr 2014 wird für mich ein »Rückschritts- Jahr« werden. Weniger Dienste mit »Abgreif-/ Tracking«-Faktor nutzen. Mobile Kommunikation reduzieren. Zurück zu Geräten die auf eine Sache spezialisiert sind und nicht alles nur mögliche abdecken. Ich hoffe es denken noch andere Bürger so. Weil dann drückt sich das auch in Umsatzzahlen aus und könnten »Schmerzen« bei euch erzeugen.« [44, Kommentar von MG16373]

Über das zunehmende Sammeln von Daten, welches dem proklamierten Ziel der Terrorismusbekämpfung dient, äußern sich Bürger überwiegend negativ:
»Big Data wird wahrscheinlich weder Terror noch Kriminalität verhindern können. [: : :] Big Data ist ja wohl eine Reaktion darauf, dass Milliarden in Sozialprogrammen nur selten etwas bewirkten, Aggression und Gewalt eine menschliche Eigenschaft bleiben.« [18, Kommentar von Werner Katz]

»Zu was für einem Staat, zu welchem Verhalten der Menschen führt das? Dazu, dass alle, bewusst oder unterschwellig, darauf hinarbeiten, nicht aufzufallen, dem Algorithmus keinen Grund zu liefern, den eigenen Namen zwecks genauerer Überprüfung auszuspucken. Und eines

ist sicher: Mit einem freiheitlichen Rechtsstaat hat das nichts zu tun, und auch nicht mit einem Staat, in dem ich mich frei und wohl fühle.« [18, Kommentar von Christoph Wirtz]

Gleichzeitig waren in der Onlinebefragung Terror- und Kriminalitätsbekämpfung die meistgenannten Bereiche, für die eine Einschränkung der Privatsphäre akzeptiert wird – knapp vor der Ansicht, kein Anwendungsbereich rechtfertige eine Einschränkung der Privatsphäre (siehe Abschnitt 7.2). Hier wird noch einmal deutlich, dass es wichtig ist, ausgewogene Lösungen bei diesem sensiblen Thema zu finden.

Kritik wurde auch daran geübt, dass Big Data öffentlich stark gefördert werde, während dies bei Datenschutzmaßnahmen nur unzureichend der Fall sei:
»Gibt's auch Förderung um zu verhindern, dass man selbst in diesen Big Data Pool schwimmt? Stromzähler, Unfallmelder im Auto, Smartphone und Apps, Mautbrücken, [...], RFID, Save Harbor, Prism, E-Klopapier und das mit meinem sauer verdienten Steuergeld gefördert ... dann kann ich auch gleich einen Dieb dafür bezahlen, dass er mich beklaut. Am besten geschützt sind Daten, die nicht erfasst sind!« [53, Kommentar von MarkusR]

8.2. Big Data in der Presse

Um zu verstehen, inwiefern Big Data in der (deutschen) Presse verbreitet ist und welches Bild von Big Data dort gezeichnet wird, wurden bei fünf bekannte Nachrichtenportalen Artikel zum Thema Big Data aus den Jahren 2013 und 2014 untersucht (siehe Abbildung 35). Die dabei gewonnenen Erkenntnisse werden im Folgenden näher erläutert.

8.2.1. Spiegel Online

Die Mehrheit der Artikel auf *Spiegel Online* befasst sich eher mit den Chancen durch Big Data als mit den damit verbundenen Risiken. So wird hier beispielsweise der Nutzen von Big Data für eine personalisierte Medizin propagiert [38]. In einem anderen Artikel wird Big Data als Präventionswerkzeug betrachtet, welches Aufstände oder Krisen in instabilen Ländern vorherzusagen vermag, um eine rechtzeitige Evakuierungen von Mitarbeitern dort ansässiger Unternehmen zu ermöglichen [37].

Aber auch Risiken werden angesprochen. Ein Artikel behandelt Googles Übernahme der Firma Nest, die intelligente Thermostate und Rauchmelder herstellt [45]. Mit den anfallenden Daten lasse sich erkennen, ob eine Person zu Hause ist und in welchem Raum sie sich aufhält. Da jene Firma von Google aufgekauft wurde, könnte Google die anfallenden Haushaltsdaten mit den Daten der zahlreichen Google-Dienste (siehe Abschnitt 3.1) verknüpfen, um noch umfangreichere und detailliertere Nutzerprofile zu erhalten. Der Artikel spricht vom »Datenschutz-Alptraum« der »total vernetzten Zukunft« und von einer »völlig neue[n] Dimension« für »potenzielle Sicherheitsprobleme«.

8.2.2. Zeit Online

Das Portal *Zeit Online* berichtet sowohl von Chancen als auch von Risiken von Big Data, wobei tendenziell ein leichtes Übergewicht beim Thema Risiken liegt. Hinsichtlich der Chancen für Big Data findet sich ein Artikel im Kontext des Einzelhandels [29]. Darin wird Big Data als nützliche

Unterstützung zur Bestimmung von Konsumgewohnheiten und Lebensstilen angesehen. Alle Erkenntnisse über das Konsumverhalten würden letztlich dem Ziel dienen, Marktanteile zu gewinnen und zu sichern. Dies geschehe zum Vorteil der Kunden, denn man strebe an, »deren Loyalität durch Qualität, Preis oder Verpackungsgröße zu gewinnen.«

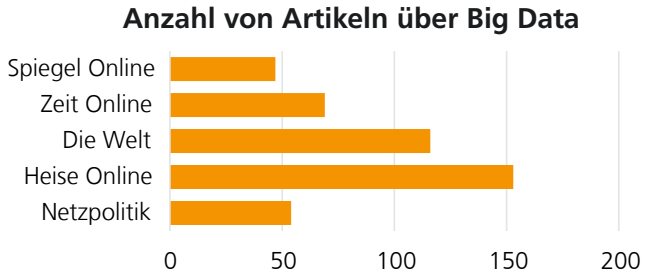

Abbildung 35: Anzahl der Artikel über Big Data in bekannten deutschsprachigen Nachrichtenportalen.

Eine kritische Sicht auf Big Data im Einzelhandel geht aus einem anderen Artikel hervor [25]. Hier liegt der Fokus auf dem Phänomen »Preisdiskriminierung«. Mittels Big Data (möglicherweise gewonnen aus Facebook-, Google- oder Smartphone-Aktivitäten der Kunden) ließen sich individuelle Preise berechnen. Diese Technologie halte nun Einzug in herkömmliche Läden. Im konkreten Fall würden Supermarktkunden individuelle Rabattcoupons basierend auf ihrer Einkaufshistorie erhalten. Dahinter verberge sich aber das Ziel, jedem Kunden den individuellen Höchstpreis zu berechnen. Durch Datensammlungen und -verknüpfungen entstünden gläserne Kunden und viele Geschäfte seien heute schon »veritable[] Überwachungsdienste[]«.

8.2.3. Die Welt

Die Artikel des Portals *Die Welt* thematisieren gleichermaßen Chancen und Risiken von Big Data. In einem Artikel aus dem medizinischen Umfeld wird erklärt, wie Big Data helfen könne, personalisierte Krebsbehandlungen zu ermöglichen [41]. So sollen die Krankengeschichte und die Genomdaten des Patienten mit allen zur Verfügung stehenden medizinischen Fachbeiträgen und Medikamenten abgeglichen werden, um dem behandelnden Arzt eine auf den Patienten individuell zugeschnittene Behandlungsempfehlung zu erstellen.

Unter den Artikeln, welche die Risiken von Big Data beleuchten, finden sich auch solche, die die Fehlbarkeit von Big Data untersuchen. Beispielsweise berichtet ein Artikel über den Einsatz von Big Data im Kontext von Tippspielen zur Fußballweltmeisterschaft [54]. Vorhersagen seien auf Basis von »der Fußballhistorie, Wirtschaftszahlen teilnehmender Länder und dem Transferwert der Spielerkader« getroffen worden, aber kläglich gescheitert. Der Autor macht daran fest, Big Data habe seine Grenzen: Favoritenstürze und weitere Überraschungen ließen sich mittels Big Data nicht vorhersagen.

8.2.4. Heise Online

Die Artikel auf *Heise Online* sind größtenteils positiv gestimmt und betrachten Big Data überwiegend im technischen Kontext. Ein Artikel, der Big Data als Chance ansieht, erläutert z. B. ein Projekt in Chicago, bei dem Sensoren installiert werden sollen, um Umweltdaten wie Luftqualität, Temperatur, Lautstärkepegel und Feinstaubwerte zu erfassen [28]. Mithilfe der zahlreich anfallenden Daten erhoffe man sich, die Metropole »besser [zu] verstehen« und gleichzeitig »sicherer, effizienter und sauberer« zu machen. US-Medien würden das Vorhaben als »freundlichen Big Brother« bezeichnen, da bei den gespeicherten Daten keine Personenzuordnung möglich sei und der Datenschutz eingehalten werde.

Ein anderer Artikel erzählt davon, wie mittels Big Data das Kerngeschäft der renommierten *New York Times* mit Abonnements bewahrt werden solle [40]. Man wolle erkennen, womit neue Abonnenten gewonnen werden können und welche Verhaltensmuster auf eine Kündigung hindeuten. Die Zeitung beschäftige dazu Analysten, die mit maschinellem Lernen helfen sollen, die geschäftlichen Probleme der überregionalen Tageszeitung zu beheben.

8.2.5. Netzpolitik

Die Artikel, die in dem Portal *Netzpolitik* veröffentlicht wurden, betrachten Big Data überwiegend kritisch. So könnten bspw. mit nur vier Ortsangaben die meisten Einzelpersonen in zunächst anonymen Mobilfunk-Bewegungsdaten identifiziert und so aussagekräftigen Bewegungsprofilen zugeordnet werden [35]. Diese Gefahr für die Privatsphäre stellt der Autor den zahlreichen Interessen Dritter an der Sammlung von Bewegungsdaten gegenüber.

Ein anderer Artikel mahnt vor einem Freiheitsverlust bedingt durch den zunehmenden Einsatz von vorhersagenden Algorithmen [4]. Die Vorhersagekraft von Algorithmen wird mit einem Beispiel demonstriert, das schwangere Frauen frühzeitig an Änderungen in ihrem Einkaufsverhalten erkenne, um ihnen zur passenden Zeit Werbung für Babysachen schicken zu können. Solche Vorhersagen würden Zielgruppen basierend auf Verhaltensähnlichkeiten identifizieren. Die Prognosesysteme könnten aber irrtümliche Urteile fällen und würden durch ihre Präsenz das Verhalten von Menschen beeinflussen und ihnen die Entscheidungsfreiheit nehmen.

9. SCHLUSSWORT

Big Data ist ein Begriff, der heute unter verschiedenen Vorbedingungen diskutiert wird. Politik und Wirtschaft sehen darin eine Technologie, die es zu nutzen gilt, um zukunftsfähig zu bleiben. Kritiker von Big Data sehen darin hingegen einen Trend, der mehr Überwachung, den Verlust der Kontrolle über die eigenen Daten und sogar eine Entmenschlichung wichtiger gesellschaftlicher Aspekte mit sich bringt.

Dieses Dokument zeigt, dass in all diesen Argumenten eine gewisse Wahrheit liegt: Big Data birgt wie viele andere Technologien Chance und Risiko in sich. Es ist jedoch nicht zu erwarten, dass einer der beiden Aspekte die Überhand gewinnt, denn die Technologie weist alle Eigenschaften eines Dual-Use-Phänomens auf: Die gleichen Algorithmen bieten beispielsweise Chancen für eine bessere Nutzung von Ressourcen als auch Risiken unkontrollierter Überwachung – abhängig davon, wozu sie eingesetzt werden. Und so ist Big Data nicht nur eine technische Entwicklung, sondern auch eine gesellschaftliche Herausforderung. Es gilt, die Interessen einer gewinnorientierten industriellen Datenverarbeitung mit dem Schutz der Privatsphäre zu vereinen.

Zurzeit existieren hinsichtlich Big Data viele Ängste und Bedenken, wie die Berichterstattung und deren Kommentierung im Internet zeigt (siehe Kapitel 8). Betrachtet man die Impulse aus dem Bürgerdialog (siehe Kapitel 6) und der Onlinebefragung (siehe Kapitel 7), so wird eines deutlich: Ein erster Schritt muss das Schaffen einer höheren Transparenz sein. Vermutlich erwecken Werbebotschaften zu Big Data den Eindruck einer sich verselbständigenden, übermächtigen Technologie, die sich in sämtliche Bereiche des Lebens drängen will. Setzt man dieser abstrakten Darstellung konkrete Anwendungsfälle entgegen, steigt die Zustimmung bei den Bürgern deutlich. Ein Arzt soll natürlich durch Big Data bessere Diagnosen stellen können, nur soll der Computer den Arzt nicht ersetzen, sondern ihn unterstützen. Genauso soll Verbrechens- und Terrorbekämpfung verbessert werden, wenn dabei die Grenzen der allgemeinen Privatsphäre nicht regelmäßig überschritten werden.

Sinnvolle und faire Regeln zum Umgang mit personenbezogenen Daten sind nötig, um Big Data zu einem allgemein akzeptierten Instrument zu machen. Aus der Onlinebefragung geht hervor, dass ca. zwei Drittel der Teilnehmer der Meinung sind, dass kein Datenschutzaspekt ausreichend umgesetzt werde. Bei Datensparsamkeit und Eingriffsrechten wird ein besonders hoher Nachholbedarf gesehen. Außerdem fordern fast 70 Prozent eine Modernisierung der Gesetze, mit denen der Einsatz von Big Data geregelt wird. Gewünscht werden hier beispielsweise kurze Vorhaltezeiten von Daten und höhere Anforderungen an die Anonymisierung personenbezogener Daten.

Besonders ausgeprägt ist die Diskussion über Transparenz und Vereinbarkeit mit dem Datenschutz beim Thema Scoring und Profiling (siehe Kapitel 4). Das Interesse von Unternehmen an möglichst genauen Berechnungen beispielsweise der Ausfallwahrscheinlichkeit von Zahlungen ist nachvollziehbar. Eine breite Ablehnung und ein ausgeprägtes Misstrauen gegen die Entscheidungen der Algorithmen basiert vor allem auf drei Faktoren: Erstens erscheint es vielen Bürgern erschreckend, Menschen anhand eines Zahlenwerts, des Score, abzubilden. Zweitens besteht die Angst, dass mit Big Data sensible Daten wie Kontakte und Äußerungen in sozialen Netzwerken in den Score einfließen können. Drittens ist die Art und Weise, wie der Score berechnet

wird, oft nicht nachvollziehbar, obwohl das BDSG dies vorschreibt. Das Resultat sind Bürger, die sich ungerechtfertigten Pauschalierungen ausgesetzt fühlen. Ein Weg zu Scoring und Profiling, der gleichzeitig die Geschäftsinteressen und -geheimnisse der Scoring-Betreiber schützt und Transparenz für Betroffene schafft, ist eine der großen zukünftigen Herausforderungen für Big Data.

Der Bürgerdialog und das vorliegende Begleitdokument sind ein Projektergebnis des *European Center for Security and Privacy by Design* (EC SPRIDE, http://www.ec-spride.de/). Das Kompetenzzentrum EC SPRIDE wird vom BMBF gefördert und ist eine Kooperation der TU Darmstadt und des Fraunhofer SIT. Die Forschung von EC SPRIDE will helfen, Sicherheit und Privatsphärenschutz schon bei der Entwicklung von Software und IT-Systemen sicherzustellen.

Die vorliegende Arbeit schließt den ersten Teil des Projektes »Big Data und Privatsphärenschutz vom Bürgerdialog bis zur risikobehafteten explorativen Grundlagenforschung« ab. Im zweiten Teil des Projektes werden Datenschutzmechanismen im Kontext von Big Data gemeinsam mit Experten der TU Darmstadt untersucht. Es gilt festzustellen, ob die Impulse und Wünsche der Bürger heute bereits realisierbar sind oder ob neue Methoden zum Datenschutz entwickelt werden müssen, beispielsweise zum Verbessern der Anonymität. Das Ergebnis wird ein Dokument sein, welches Datenschutzmechanismen detaillierter und technischer als hier erörtert und als Grundlage für eine Umsetzung in praktischen Anwendungen dienen kann.

Zu guter Letzt wollen wir hier allen danken, die uns bei dieser Studie durch die Teilnahme an dem Bürgerdialog und der Onlinebefragung unterstützt haben.

LITERATUR

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

[2] *Beuth, Patrick: Algorithmen: Die Polizei als Hellseher. Zeit Online, August 2011. http:// www.zeit.de/digital/datenschutz/2011-08/ predictive-policing.*

[3] *Beuth, Patrick: Snowden-Enthüllungen: Alles Wichtige zum NSA-Skandal. Zeit Online, Oktober 2013. http://www.zeit.de/digital/ datenschutz/2013-10/hintergrund-nsaskandal/ komplettansicht, Inhalt zuletzt aktualisiert am 05.01.2015.*

[4] *Biermann, Kai: Algorithmen Allmächtig? Freiheit in den Zeiten der Statistik. Netzpolitik, Juli 2014. https://netzpolitik.org/2014/ algorithmen-allmaechtig-freiheit-in-denzeiten- der-statistik/.*

[5] *Biermann, Kai: Überwachungsaffäre: NSAAusschuss sieht nur schwarz. Zeit Online, September 2014. http://www.zeit.de/ politik/ deutschland/2014-09/nsa-ausschuss-aktengeschwaerzt.*

[6] *BITKOM: Potenziale und Einsatz von Big Data. Studienbericht, BITKOM, Mai 2014. http: //www.bitkom.org/de/publikationen/38338_ 79283.aspx.*

[7] *BITKOM-Arbeitskreis Big Data: Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte. Leitfaden, BITKOM, September 2012. http: //www.bitkom.org/de/publikationen/38337_ 73446.aspx.*

[8] *BITKOM-Arbeitskreis Big Data: Management von Big-Data-Projekten. Leitfaden, BITKOM, Juni 2013. http://www.bitkom.org/ de/publikationen/38337_76511.aspx.*

[9] *BITKOM-Arbeitskreis Big Data: Big-Data- Technologien – Wissen für Entscheider. Leitfaden, BITKOM, Februar 2014. http://www.bitkom.org/de/publikationen/38337_78776. aspx.*

[10] *Bodden, Eric, Siegfried Rasthofer, Philipp Richter und Alexander Roßnagel: Schutzmaßnahmen gegen datenschutz-unfreundliche Smartphone-Apps. Datenschutz und Datensicherheit, 37(11):720–725, November 2013. http://www.dud.de/Premium-Inhalt/40/2706/ Schutzma-#223;nahmen-gegen-datenschutz--unfreundliche-Smartphone-Apps.html.*

[11] *Brühl, Jannis und Florian Fuchs: Gesucht: Einbrecher der Zukunft. Süddeutsche.de, September 2014. http://www.sueddeutsche.de/ digital/polizei-software-zur-vorhersagevon- verbrechen-gesucht-einbrecher-derzukunft- 1.2115086.*

[12] *Bull, Hans Peter: Es war einmal ein Datenschutz-Märchen. Süddeutsche.de, November 2014. http://www.sueddeutsche.de/digital/pkw-maut-der-bundesregierunges- war-einmal-ein-datenschutz-maerchen-1. 2200854.*

[13] *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.*

[14] *Christl, Wolfie: Kommerzielle digitale Überwachung im Alltag. Studie, Cracked Labs, November 2014. http://crackedlabs.org/ studie-kommerzielle-ueberwachung/info.*

[15] *Clarke, Roger: Introduction to dataveillance and information privacy, and definitions of terms, August 1997. http://www.rogerclarke.com/ DV/Intro.html, Inhalt zuletzt aktualisiert im Oktober 2013.*

[16] *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-inofficialconsolidated- LIBE.pdf.*

[17] *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.*

[18] *Fienbork, Matthias: Evgeny Morozov zu Big Data Warum entsteht Terror? FAZ.net, Juni 2013. http://www.faz.net/-hbj-7aavl.*

[19] *Finn, Rachel L., David Wright und Michael Friedewald: Seven types of privacy. In: European Data Protection: Coming of Age, Seiten 3–32. Springer, 2013. http://link.springer.com/ chapter/10.1007/978-94-007-5170-5_1.*

[20] *Foschepoth, Josef: Überwachtes Deutschland – Post- und Telefonüberwachung in der alten Bundesrepublik. Vandenhoeck & Ruprecht, 2012. http://www.v-r.de/de/title-1-1/ ueberwachtes_deutschland-1007436/, 4., durchgesehene Auflage 2014.*

[21] *Fröhlich, Christoph: Googles große Zwangseingemeindung. Stern.de, Januar 2013. http: //www.stern.de/digital/online/ googlepflicht- fuer-youtube-und-co-googlesgrosse- zwangseingemeindung-1952778.html.*

[22] *Friend, Zach: Predictive policing: Using technology to reduce crime. FBI Law Enforcement Bulletin, April 2013. http://leb.fbi.gov/ 2013/april/predictive-policing-usingtechnology- to-reduce-crime.*

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

[24] *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/abs/nature07634.html.*

[25] *Grassegger, Hannes: Konsum: Jeder hat seinen Preis. Zeit Online, Oktober 2014. http:// www.zeit.de/wirtschaft/2014-10/absolute-preisdiskriminierung.*

[26] *Greenwald, Glenn: No Place to Hide: Edward Snowden, the NSA and the U.S. Surveillance State. Metropolitan Books (Henry Holt), Mai 2014. http://us.macmillan.com/books/ 9781627790734.*

[27] *Henschen, Doug: Analytics at work: Q&a with tom davenport. Interview with thomas davenport, InformationWeek, April 2010. http://www. informationweek.com/news/software/bi/ 222200096.*

[28] *Holland, Martin: »Freundlicher Big Brother«: Umweltsensoren in Chicago zählen Mobilgeräte. Heise Online News, Juni 2014. http://heise. de/-2240788.*

[29] Jungclaussen, John F.: Schwerpunkt: Big Data: Oma will kein Megapack. Zeit Online, Januar 2013. http://www.zeit.de/2013/02/ Tesco- Verbraucherverhalten-Auswertung-Big-Data- Supermarkt.

[30] Kannenberg, Axel: NRW testet ab 2015 Software zu Kriminalitäts-Vorhersagen. Heise Online News, November 2014. http://heise.de/- 2468412.

[31] 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-privatheitde/ texte/veroeffentlichungen-desforums/ themenpapiere-white-paper/ Forum_ Privatheit_White_Paper_Selbstdatenschutz_ Web.pdf.

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

[33] 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.

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

[35] Meister, Andre: Einzigartig in der Masse: Aus Mobilfunk-Bewegungsdaten können ganz einfach Einzelpersonen identifiziert werden. Netzpolitik, März 2013. https://netzpolitik. org/2013/ einzigartig-in-der-masse-ausmobilfunk- bewegungsdaten-konnen-ganzeinfach- einzelpersonen-identifiziertwerden.

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

[37] Peil, Florian: Computermodelle: So wollen Forscher Revolutionen vorhersagen. Spiegel Online, Juli 2013. http://www.spiegel. de/ netzwelt/web/big-data-extremereignissemittels- statistik-vorhersagen-a-912347. html.

[38] Pietsch, Wolfgang: Big Data in der Medizin: Sprechstunde beim Superrechner. Spiegel Online, Juli 2013. http://www.spiegel. de/wissenschaft/medizin/big-datawundermittel- auch-fuer-die-medizin-a- 911333.html.

[39] 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.

[40] Regalado, Antonio und Ben Schwan: Mit Big Data gegen Aboverluste. Technology Review (Heise Online), März 2014. http://heise. de/- 2138251.

[41] Ridderbusch, Katja: Medizin: Wie riesige Datenmengen den Krebs besiegen sollen. Die Welt, Dezember 2014. http://www. welt.de/ wirtschaft/article135026492/Wie-riesige- Datenmengen-den-Krebs-besiegen-sollen. html.

[42] Rijmenam, Mark van: The Los Angeles Police Department is predicting and fighting crime with big data, April 2014. https:// datafloq. com/read/los-angeles-police-departmentpredicts- fights-crim/279.

[43] 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.

[44] Sokolov, Daniel A. J. und Martin Holland: Cisco: Internet of Everything = Internet mal 10. Heise Online News, Januar 2014. http://heise. de/-2077874.

[45] Stöcker, Christian: Nest-Übernahme: Google will in Ihr Schlafzimmer. Spiegel Online, Januar 2014. http://www.spiegel. de/netzwelt/ gadgets/nest-uebernahme-google-will-inihr-schlafzimmer-a-943406.html.

[46] Stuart, Tessa: Santa Cruz's predictive policing experiment, Februar 2012. http://www. santacruz.com/news/santa_cruzs_predictive_ policing_experiment.html.

[47] Sweeney, Latanya: K-anonymity: A model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):557–570, Oktober 2002. http: //www. worldscientific.com/doi/10.1142/ S0218488502001648.

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

[49] Türpe, Sven, Annika Selzer, Andreas Poller und Mark Bedner: Denkverbote für Star-Trek- Computer? Datenschutz und Datensicherheit, 38(1):31–35, Januar 2014. http://www.dud. de/ Premium-Inhalt/40/2765/Denkverbote-f-#252; r-Star-Trek-Computer-.html.

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

[51] 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.

[52] Wienand, Lars: Autobahnschütze: RLPDatenschützer fordert Gesetzesänderung für Massen-Kennzeichen-Erfassung. Interview mit Edgar Wagner, Rhein-Zeitung, August 2014. http://www. rhein-zeitung.de/ region_artikel,-Autobahnschuetze-RLPDatenschuetzer- fordert-Gesetzesaenderungfuer- Massen-Kennzeichen-Erfassung-_arid, 1192889.html.

[53] Wilkens, Andreas: 2,5 Milliarden Euro sollen EU-Wirtschaft bei Big Data voranbringen. Heise Online News, Oktober 2014. http:// heise.de/- 2421299.

[54] Zschäpitz, Holger: Tippspiel: Big Data hat leider keine Ahnung von Fußball. Die Welt, Juli 2014. http://www.welt.de/finanzen/ verbraucher/ article130355461/Big-Data-hat-leiderkeine-Ahnung-von-Fussball.html.