Hochschule München
Fakultät 07 Informatik und Mathematik & 10 Betriebswirtschaft

Bachelorarbeit von **Nicolai Ruhnau**

Betreuer Prof. Dr. Ulrich Möncke
Bachelor **Wirtschaftsinformatik**

Bearbeitungsbeginn: 23.6.2016
Abgabetermin: 23.9.2016

Big-Data-Vorhersagbarkeit gruppenspezifischen Verhaltens – Chancen und Risiken

Big Data Predictability of Group Specific Behaviour – Opportunities and Risks

Hiermit erkläre ich, dass ich die Bachelorarbeit selbständig verfasst, noch nicht anderweitig für Prüfungszwecke vorgelegt, keine anderen als die angegebenen Quellen oder Hilfsmittel benutzt, sowie wörtliche und sinngemäße Zitate als solche gekennzeichnet habe.

München, 22.9.2016 _____

Hiermit erkläre Ich mein Einverständnis, dass die von mir erstellte Bachelorarbeit in die Bibliothek der Hochschule München eingestellt wird. Ich wurde darauf hingewiesen, dass die Hochschule in keiner Weise für die missbräuchliche Verwendung von Inhalten durch Dritte infolge der Lektüre der Arbeit haftet. Insbesondere ist mir bewusst, dass ich für die Anmeldung von Patenten, Warenzeichen oder Geschmacksmustern selbst verantwortlich bin und daraus resultierende Ansprüche selbst verfolgen muss.

München, 22.9.2016 _____

Abstract

Unklare Definitionen sowie Unsicherheiten bezüglich subjektiver Elemente und historischer Parallelen zur Statistik prägen die Debatte um automatisierte Datenverarbeitung. Wohlstandsfördernde, aber auch gefährliche Potentiale anonymer Vorhersagbarkeit sowie aktuelle Entwicklungen werden anhand von vielfältigen Beispielen betrachtet, insbesondere im Zusammenhang mit Geo-daten. Mittels ausführlichen Literaturvergleichs wird eine fachübergreifende Perspektive in Bezug auf Algorithmen des Machine Learning und auf Anonymisierung geschaffen, indem juristische, wirtschaftliche, technische und statistische Aspekte transparent gemacht werden. Ökonomische Analysen des Datenschutzes und Auswirkungen auf Individuen, Gruppen und das Gemeinwohl werden erörtert.

Schlüsselworte

Statistik – Automatisierte Datenverarbeitung – Geodaten – Algorithmen – Machine Learning - Anonymisierung – Ökonomische Analyse – Datenschutz

Abstract

Unclear definitions as well as insecurities about subjective elements and historic parallels to statistics influence the debate about automated data processing. Affluence-promoting, but also dangerous potentials of anonymous predictability as well as current developments are examined by means of a variety of examples, in particular in regard to spatial data. Using an extensive comparative study of literature an interdisciplinary perspective related to algorithms of machine learning and anonymization is established as legal, economic, technical and statistical aspects are made transparent. Economic Analyses of privacy and implications for individuals, groups and the common weal are discussed.

Keywords

Statistics – Automated Data Processing – Spatial Data – Algorithms – Machine Learning – Anonymization – Economic Analysis - Privacy

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Einleitung

Die Risiken der Vorhersagbarkeit von Verhalten wurden bislang vor allem im Hinblick auf die Privatsphäre von Individuen diskutiert. Die Betrachtung der Probleme anonymisierter Datenverarbeitung, bei denen Individuen zu Gruppen zusammengefasst sind, wurde dabei auch regulatorisch lange vernachlässigt. Besondere Brisanz erfährt diese Problematik dadurch, dass Informationstechnik unser Leben immer mehr durchdringt und damit umfassende Datenquellen für Konzerne schafft. Andererseits fehlt oft eine entsprechende Gegenüberstellung auch der Chancen anonymisierter Analysen, insbesondere im Zusammenhang mit Geodaten. Juristische, wirtschaftliche oder populärwissenschaftlich angehauchte Quellen behandeln Chancen und Risiken moderner Datenanalyse zwar, oft werden aber technische oder statistische Begriffe und Verfahren missverständlich oder unscharf definiert. Technisch ausgerichtete Literatur wiederum diskutiert Chancen und Risiken, wenn überhaupt, dann nicht ausführlich. Kontinuitäten moderner und traditioneller Datenanalyse werden bislang oft übersehen, die Ursachen von erkannten Risiken nicht genügend transparent gemacht. Dazu gehören auch die Fragen, ob moderne automatisierte Datenanalyse überhaupt noch subjektive Elemente besitzt und ob wirtschaftliche Analysen Hilfestellungen bei regulatorischen Reaktionen geben können.

Es handelt sich also um ein Querschnittsthema, das im Sinne einer integrierenden Sicht statistische, informationstechnische, volks- und betriebswirtschaftliche sowie rechtliche Aspekte vereinend durch Literaturvergleich beleuchtet wird. Verschiedene Beispiele verdeutlichen die Relevanz der Fragestellung und geben Einblick in aktuelle Entwicklungen.

Im ersten Teil werden die Begriffe des Themas bestimmt, angefangen mit dem Begriffen Daten und Informationen. Der Schwerpunkt liegt auf Definitionen zur Vorhersagbarkeit von Verhalten im Allgemeinen und in Bezug auf gruppenspezifisches Verhalten, sowie zu Big Data mit all seinen Facetten. Im zweiten Teil geht es darum grundsätzlich darzulegen, wie man im Sinne von Big Data technisch vorgeht, gruppenspezifisches Verhalten vorhersagen zu können. Es wird auch auf die Frage eingegangen, ob Datenanalyse nach dem Stand der Technik noch subjektive Elemente enthält. Es werden zudem beispielhaft besonders po-

puläre Algorithmen überblicksartig erklärt, wobei eine detaillierte mathematische Abhandlung den Rahmen dieser Arbeit sprengen würde. Im dritten Teil geht es um einen Überblick über die Chancen und Risiken, insbesondere in Bezug auf Geodaten. Im Bereich der Risiken wird der Schutz personenbezogener und anonymisierter Daten beschrieben. Dabei wird der Datenschutz in Deutschland von den Anfängen wie dem Volkszählungsurteil bis hin zu neueren Entwicklungen betrachtet. Zudem werden die Ursachen für das Entstehen der Gefahren untersucht.

Ich möchte mich bei Herrn Prof. Dr. Möncke für die Unterstützung und die Anregungen bei der Themenfindung und der Erstellung dieser Bachelorarbeit bedanken.

2 Begriffsbestimmungen

2.1 Daten und Informationen

Mit Datum ist eine mehr oder weniger subjektive Beobachtung bis hin zu einem objektiv wahren Fakt im Sinne einer Zahl gemeint, welche als Folge von Nullen und Einsen, d.h. binär repräsentiert ist und dadurch elektronisch kommuniziert werden kann[1].

Eine Information meint ein Datum, das mit Sinn oder Bedeutung verbunden ist[2].

2.2 Vorhersagbarkeit gruppenspezifischen Verhaltens

2.2.1 Vorhersagbarkeit von Verhalten

Vorhersagbarkeit von Verhalten meint die Fähigkeit, Aussagen darüber zu treffen, wie sich ein Mensch oder eine Gruppe von Menschen verhalten werden. Über die Analyse von Erfahrungswerten wird also eine Entscheidung darüber getroffen, für wie wahrscheinlich bestimmtes zukünftiges Verhalten gehalten wird[3].

Die Menschen, auf deren Erfahrungswerten die Analyse basiert, können von den Menschen, deren Verhalten prognostiziert wird, verschieden sein[4].

Da es sich um datenbasierte Entscheidungen handelt, geht es um Statistik[5]. Vorhersagen sind mit Wahrscheinlichkeiten verbunden. Selbstverständlich besitzt jeder eine gewisse Vorstellung davon, was Wahrscheinlichkeit bedeutet. Um allerdings die Qualität von Vorhersagen besser einschätzen zu können, ist eine Quantifizierung hilfreich, wie Laplace schon 1814 schreibt[6].

---

[1] vgl. Ballsun-Stanton, 2012, S. 353f; vgl. Voß, LIBREAS 2013, (5f).
[2] vgl. Vidgen et al., SJIS 1993, 97 (98).
[3] vgl. Bäumler/Gutsche, VuR 2008, 81 (81f).
[4] vgl. Kamp/Weichert, 2005, S. 10.
[5] vgl. Wernecke, 1995, S. 1.
[6] vgl. Laplace, 1814, S. 95: „On voit, par cet Essai, que la théorie des probabilités n'est, au fond, que le bon sens réduit au calcul; elle fait apprécier avec exactitude ce que les esprits justes sentent par une sorte d'instinct, sans qu'ils puissent souvent s'en rendre compte."

Im Allgemeinen haben sich zwei Definitionen von Wahrscheinlichkeit durchgesetzt, nämlich erstens die frequentistische und zweitens die bayessche. Erstere versteht Wahrscheinlichkeit als objektiv wahren Erwartungswert beliebig oft wiederholbarer Zufallsexperimente, wie z.B. das Werfen einer Münze[7]. Die Definition im Sinne Bayes hingegen versteht Wahrscheinlichkeit als subjektive Glaubwürdigkeit oder Unsicherheit auf Basis der vorhandenen Information[8]. Dies hat den Vorteil, dass auch Ereignisse von ihr erfasst sind, die sich nicht wiederholen[9].

Für die Klassifizierung von Individuen in gegebene Gruppen, um Verhalten vorherzusagen, erweist sich die bayessche Definition als passender[10]; sie ist auch grundsätzlich im Bereich moderner Datenanalysen tendenziell beliebter[11].

Üblicherweise quantifiziert man Wahrscheinlichkeit, indem man einer Aussage einen Wert zwischen 0 und 1 zuweist, d.h. man sagt im Sinne Bayes, dass man zu 0 oder 100% sicher ist, dass ein gewisses Ereignis eintritt. Der Wert 0,5 entspricht so einer 50%igen Wahrscheinlichkeit, also quasi einem Münzwurf. Eine solche Wahrscheinlichkeit scheint zunächst wenig wertvoll für Aussagen bezüglich gruppenspezifischen Verhaltens. Dennoch kann eine solche Aussage großen Wert haben, wenn man sie zur sonst üblichen Wahrscheinlichkeit des Ereignisses in Relation setzt. Im Fall der Kreditwürdigkeit genügt eine 50%ige Wahrscheinlichkeit, den Kredit rückzahlen zu können, dann nicht, wenn es genügend Kunden gibt, die eine höhere Wahrscheinlichkeit aufbringen. In anderen Fällen mit mehr Unsicherheit sind 50% Zuversicht vielleicht schon ein herausragend gutes Ergebnis.

Die Zielrichtung der Prognose kann beliebiges menschliches Verhalten betreffen. Anwendungsbeispiele betreffen Wahlen[12], Raumplanung und öffentliche

---

[7] vgl. Wasserman, 2004, S. 205; vgl. Teschl/Teschl, 2014, S. 241.
[8] vgl. Wasserman, 2004, S. 206.
[9] Murphy, 2012, S. 27.
[10] vgl. Tan et al., 2006, S. 227f; vgl. Murphy, 2012, S. 25f.
[11] vgl. Murphy, 2012, S. 25f; vgl. Wasserman, 2004, S. 206.
[12] vgl. Rusch et al., 2012.

Infrastruktur[13], Konsumverhalten[14], Risikobeurteilung im Hinblick auf Gesetzestreue, Kredite oder Versicherungen[15].

Grundsätzlich kann man sagen, dass je mehr Daten man hat, desto präziser Vorhersagen möglich sind, ähnlich dem Gesetz der großen Zahlen[16]. Insbesondere zufällige Schwankungen und fehlerhafte Ausreißer der Daten nehmen dadurch in ihrer Bedeutung ab.

2.2.2 Gruppenspezifisches Verhalten

Als Gruppe wird in dieser Arbeit eine Menge von Menschen verstanden, die möglichst ähnlich (homogen) zueinander, aber möglichst unterschiedlich (heterogen) bzw. abgegrenzt (separiert) gegenüber anderen sind[17]. Die Identität der Mitglieder ist irrelevant, d.h. sie sind anonym. Jede Gruppe besitzt beliebige Charakteristiken oder Eigenschaften, die auch als Merkmale (englisch attributes bzw. features[18]) bezeichnet werden können[19].

Im Sinne der mathematischen Mengen handelt es sich bei Gruppen um Teilmengen. Die Art der Zugehörigkeit einer Person zu einer Gruppe kann verschieden ausgestaltet sein[20]: eine Gruppierung kann exklusiv oder disjunkt sein, d.h. jede Person gehört zu genau einer Gruppe. Die Einteilung in Gruppen kann auch ein Überlappen erlauben, indem eine Person gleichzeitig zu mehreren Gruppen gezählt wird. Eine Gruppierung folgt der sogenannten Fuzzy Logic (wörtlich unscharfe Logik), wenn jedes Individuum zu jeder Gruppe einen Zugehörigkeitswert zwischen 0 und 1 aufweist, wobei 1 vollständige Zugehörigkeit impliziert und 0 größtmögliche Verschiedenheit[21]. Fuzzy Logic entspricht der Vorstellung, dass ein Element grundsätzlich nicht nur binär im Sinne von ganz oder gar nicht Teil einer Menge sein kann, sondern es auch in der Realität Übergänge,

---

[13] vgl. Batty, DHG 2013, 274.
[14] vgl. Larson et al., Int. J. Res. Mark. 2005, 395 (395).
[15] Kamp/Weichert, 2005, S. 12ff; vgl. Han et al., 2011, S. 443.
[16] Bolthausen/Wüthrich, ASTIN Bulletin 2013, 73; vgl. Junqué de Fortuny et al., Big Data 2013, 215 (223); vgl. Halevy et al., IEEE Intell. Syst. 2009, 8 (12); vgl. Banko/Brill, 2001, S. 32.
[17] vgl. Wernecke, 1995, S. 185; vgl. Falk et al., 2014, S. 259; vgl. Bäumler/Gutsche, VuR 2008, 81 (81); vgl. Han et al., 2011, S. 20.
[18] vgl. Aggarwal, 2015, S. 154.
[19] vgl. Wernecke, 1995, S. 185.
[20] vgl. Tan et al., 2006, S. 492f.
[21] vgl. Tan et al., 2006, S. 492.

Unsicherheiten, Ungefähre Informationen und sich überschneidende Mengen gibt, die in sogenannten Fuzzy Sets abgebildet werden sollen[22].

Die Unterteilung der gegebenen Menschen muss nicht unbedingt vollständig sein. Ein Mensch kann auch keiner Gruppe zugeordnet sein, die Gruppierung wird dann partiell genannt[23].

In der Statistik verwendete verwandte Begriffe sind Kategorie, Taxonomie, Familie, Spezies, Segmente, Partition, Klasse und Cluster (englisch für Klumpen, Bündel, Anhäufung)[24].

Eine Gruppierung soll selbst entweder bereits eine bedeutende Erkenntnis darstellen, d.h. reich an Bedeutungen z.B. für das Verständnis über die Daten sein, oder jedenfalls nützlich[25]. Ziele sind insbesondere die Zusammenfassung, Aufteilung und Vorbereitung der Daten für eine bessere Übersicht und weitere Analyseschritte wie etwa die Vorhersage von Verhalten[26].

Bei gruppenspezifischen Verhalten geht also zunächst einmal nicht um die Beschreibung einer einzelnen Person mit ihren individuellen Eigenschaften, Interessen, Interaktionen und Geheimnissen, sondern darum, aus einer Menge von Personen mit Gemeinsamkeiten einen neuen, abstrakten und durch Aggregierung (englisch Aggregation, d.h. Zusammenfassen von Daten zu einem Ganzen)[27] anonymisierten Vergleichswert zu bilden, der wertend in Bezug zur Umwelt gesetzt werden kann[28]. Diese Teilmengen können bezüglich ihres Verhaltens analysiert und Korrelationen zwischen Merkmalen und beobachteter Verhaltenshistorie erkannt werden, ohne unbedingt auf präzise Informationen über einzelne angewiesen zu sein. Darauf aufbauend lassen sich Prognosen bilden. Wenn es nicht auf die Beurteilung des Verhaltens einer individuellen Person ankommt, sondern um das Verhalten von Menschenmassen geht - z.B. bei Planungsaufgaben, bei denen aggregierte Werte ausreichend sind - ist damit das Ziel der Vorhersagbarkeit gruppenspezifischen Verhaltens bereits erreicht.

---

[22] vgl. Noll, 2009, S. 25ff; vgl. Bosc/Prade, 1997, S. 1ff.
[23] vgl. Tan et al., 2006, S. 493.
[24] vgl. Tan et al., 2006, S. 487ff.
[25] vgl. Tan et al., 2006, S. 487ff.
[26] vgl. Aggarwal, 2015, S. 153f.
[27] vgl. Agrawal/Srikant, 2000; vgl. BITKOM, 2012, S. 45.
[28] vgl. Karg, ZD 2012, 255 (260).

Allerdings können die Erkenntnisse grundsätzlich auch auf Individuen angewandt werden. Man vergleicht eine ausgewählte, eventuell auch noch nicht zu einer Gruppe gehörende Person mit den vorhandenen Gruppen, um ihr Verhalten vorhersagen zu können. Durch die Übereinstimmung mit den Merkmalen der Gruppe wird das Individuum kategorisiert. Mit der Folge, dass das gesammelte Wissen über die Gruppe nun auch auf das betrachtete Individuum Rückschlüsse erlaubt.

Was indes für die jeweilige Zielsetzung der Vorhersage von Verhalten eine sinnvolle Gruppe darstellt, ist ein subjektiver Vorgang[29]. Dabei ist auch die Frage subjektiv zu beurteilen, wie viele Menschen eine Gruppe im Sinne der Vorhersage gruppenspezifischen Verhaltens ausmachen. Dem Gesetz der großen Zahlen folgend (s.o.) wird davon ausgegangen, dass mehr erfasste Personen grundsätzlich besser sind. Eine Gruppe sollte in diesem Sinne grundsätzlich als kleinstmöglichste Größenordnung mindestens 1000 Personen umfassen und dies auch nur, wenn die erfassten Personen vergleichsweise umfangreiche Datenmengen mit hoher Qualität mit sich bringen. Im Kontext von Big Data sinken allerdings die Kosten der Datenerzeugung[30]. Dadurch wird es sich oft wie von selbst ergeben, dass es sich um wesentlich mehr Personen innerhalb einer Gruppe handelt; bis hin zu Hunderten von Millionen oder gar Milliarden von Menschen.

Für die Nützlichkeit der Vorhersagbarkeit gilt auf andere Art und Weise ebenfalls. Auf je mehr Personen man die Erkenntnisse anwendet, umso mehr kann sich die oft aufwändige Investition in die Gewinnung von Erkenntnissen amortisieren und rentieren.

---

[29] vgl. Jain, Pattern Recogn. Lett. 2010, 651 (3).
[30] King, 2016, S. 1f.

2.3 Big Data Charakteristika

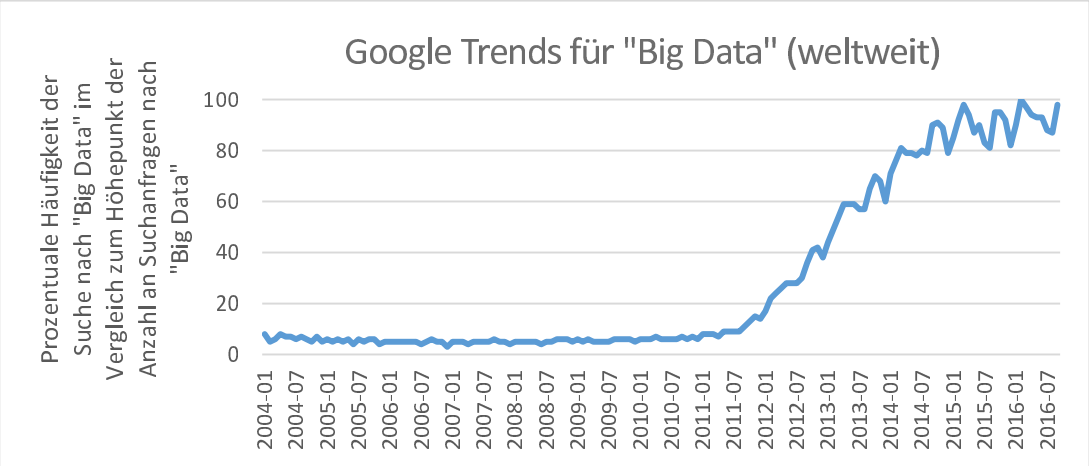

Abbildung 1: Google Trends[31]

Der Begriff Big Data - engl. „große Daten“ - mag zunächst selbsterklärend erscheinen. Doch gibt es trotz (oder vielleicht gerade wegen) der Popularität des Begriffs - seitdem z.B. die New York Times 2012 das Zeitalter von Big Data ausrief („The Age of Big Data“)[32] - keine strenge und erst recht keine messbare Definition[33]. Grundsätzlich liegt das Problem darin, dass der Begriff „groß“ mit quantifizierbaren Mengen assoziiert ist; diese klar zu umgrenzen oder gar historisch ins Verhältnis zu setzen ist allerdings ein subjektiver Vorgang[34]. Trotzdem ist es von Nutzen, sich an einer Begriffsbestimmung zu versuchen.

Grundsätzlich und Diskussionen zusammenfassend[35] lässt sich Big Data folgendermaßen definieren: die automatisierte Gewinnung von Erkenntnissen aus großen Datenmengen („Volume“), vielfältiger Herkunft („Variety“) und in hoher Geschwindigkeit („Velocity“)[36]. Es geht also nicht nur um Daten, sondern um Information (s.o.). Automatisierte Gewinnung von Erkenntnissen meint den konsequenten Wechsel weg vom manuellen Formulieren und Verifizieren von Hypothesen hin zum automatisierten Finden von Korrelationen, z.B. über das

---

[31] Google Incorporated, 18. September 2016.
[32] Lohr, The New York Times, 11. Februar 2012.
[33] Mayer-Schönberger/Cukier, 2013, S. 6, sowie S. 25: „answer questions that it didn’t consider in advance“; Manyika et al., 2011, S. 1.
[34] vgl. Ward/Barker, arXiv:1309.5821 [cs] 2013, (1).
[35] vgl. Ward/Barker, arXiv:1309.5821 [cs] 2013.
[36] Laney, Meta Group (Gartner), 2001; BITKOM, 2014, S. 19; Schroeck et al., 2012, S. 3; Zikopoulos et al., 2011, S. 4; vgl. ORACLE, 2016, S. 4; vgl. National Institute of Standards and Technology, 2015, S. 5.

sogenannte Data Mining[37]. Der Fortschritt im Bereich Datenanalyse wird von manchen in der Wissenschaft als derartig gewaltig angesehen, dass behauptet wird, Moore's Law sei im Vergleich Größenordnungen darunter anzusiedeln[38]. Der Begriff ist notorisch unscharf, da er viel mit Marketing zu tun hat[39].

Die Tätigkeit desjenigen, der diese Datenmanipulationen durchführt, ist dabei zunehmend interdisziplinär[40] und künstlerisch[41] - es ist bisweilen von „data artists" die Rede[42]. Dass Statistik zu Big Data gehört, wird auch durch den Wunsch mancher Autoren verdeutlicht, die ein viertes Kriterium „Richtigkeit" („Veracity") sehen wollen, das die Zuverlässigkeit von Informationen und Vorhersagen anspricht[43]. Übertragen auf die Vorhersagbarkeit von Verhalten werden Analysen nur dann durch mehr Daten besser, wenn diese Daten auch eine gewisse Mindest-Qualität aufweisen. Dennoch kann gegebenenfalls allein die schiere Menge von Daten, obwohl sie nicht strukturiert sind, wertvoll sein[44].

Eine andere Definition spricht von Datenmengen, die von herkömmlicher Technologie nicht zu verarbeiten und analysieren sind[45]. In anderen Worten gehören zu Big Data auch die verwendeten Technologien[46], bei denen es insbesondere auch um verteilte Datenverarbeitung geht.

Ein praktisches Beispiel für Big-Data-Analysen ist die Technologie Google Translate. Der Erfolg lag nicht daran, besonders viele Experten für Fremdsprachen zusammenarbeiten zu lassen, sondern die Grundidee war schlicht: möglichst viele Quellen durchsuchen, die bereits übersetzt worden sind, und durch Vergleich im Sinne von Korrelation eine Wahrscheinlichkeit für jedes Wort errechnen, was es in der anderen Sprache bedeutet[47]. Solche Quellen sind z.B. von der EU oder den UN in mehreren Sprachen vorliegende Dokumente, Websites von internationalen Unternehmen, Wikipedia und Google Books, das sich mit seinen in viele Sprachen übersetzten und sehr langen Texten besonders anbie-

---

[37] Mayer-Schönberger/Cukier, 2013, S. 55.
[38] vgl. King, 2016, S. 2.
[39] Lohr, The New York Times, 11. Februar 2012.
[40] vgl. Han et al., 2011, S. 29.
[41] vgl. Aggarwal, 2015, S. 5.
[42] Teradata Perspectives, Forbes, 30. Januar 2015; Davis, Forbes, 30. Juni 2015.
[43] vgl. Freiknecht, 2014, S. 13; vgl. Schroeck et al., 2012, S. 5.
[44] vgl. Halevy et al., IEEE Intell. Syst. 2009, 8 (12); vgl. Banko/Brill, 2001, S. 32.
[45] vgl. Manyika et al., 2011, S. 1; vgl. BITKOM, 2012, S. 21.
[46] Zikopoulos et al., 2011, S. 51ff; ORACLE, 2016, S. 4; Mayer-Schönberger/Cukier, 2013, S. 6.
[47] vgl. Mayer-Schönberger/Cukier, 2013, S. 36–39.

tet. Machine Learning Experten bei Microsoft werden scherzend zitiert, dass die Qualität von Übersetzungen mit jedem entlassenen Sprachwissenschaftler, der das Team verlässt steige[48].

Die neuen Möglichkeiten bieten sich für Vorhersage von Verhalten an, wobei im Folgenden einzelne Elemente der Definition mit mehr Detail erläutert werden.

2.3.1 Volume

Um die Jahrtausendwende setzte die Digitalisierung zunehmend ein und führte zu immer größer werdenden Datenmengen. Digitalisierung meint einerseits, dass bisherige und neue Informationen vorrangig digital statt analog gespeichert werden[49]. Andererseits vervielfacht sich die Menge an digital gespeicherter Information, und zwar in exponentieller Weise[50]; je nach Meinung verdoppelt sich alle 12[51] bis 24[52] Monate. Daneben führte Moores Law zu extremen Vergünstigungen im Bereich der Rechenkraft, auch wenn sich dieser Effekt in Zukunft wohl abschwächen wird[53]. All das hat neben anderen Faktoren dazu geführt, dass die Kosten der Datenerzeugung und -verarbeitung stark gesunken sind[54].

Die seit ungefähr zehn Jahren allgemein beliebt gewordenen sozialen Medien z.B. erzeugen gigantische Datenmengen. Auf Facebook und WhatsApp sind jeweils mehr als eine Milliarde Benutzer mindestens monatlich aktiv[55], auf dem in China beliebten Äquivalent Weixin/WeChat über 800 Millionen[56]. Auf Facebook wurden 2010 täglich über 100 Millionen Fotos hochgeladen; die gesamten Bilder machten damals mehr als 20 Petabyte aus[57]. Diese Zahl ist weiter gestiegen, so

---

[48] vgl. Mayer-Schönberger/Cukier, 2013, S. 142.
[49] vgl. Abbildung in Lohr, The New York Times, 1. Februar 2013.
[50] Hilbert/López, Science 2011, 60 (64).
[51] Helbing, 2015, S. 76.
[52] Gantz/Reinsel, 2012, S. 1.
[53] Simonite, Massachusetts Institute of Technology. Technology Review, 13. Mai 2016.
[54] King, 2016, S. 1f.
[55] Facebook, 2016, S. 6.
[56] Tencent, 2016, S. 3.
[57] Beaver et al., 2010.

wurden 2015 jeden Tag über 400 Millionen, auf WhatsApp sogar über 1000 Fotos übermittelt[58].

Auch in traditionellen Unternehmen finden sich neue Datenmassen. Es wird geschätzt, dass der amerikanische Einzelhandelskonzern Walmart bereits 2011 eine Datenbank über die Kundentransaktionen von über 2,5 Terabyte besaß[59]. Im gleichen Jahr erzeugte Twitter über 7 TB an pro Tag, Facebook 10 TB[60]. Bis 2020 wird von manchen eine weltweite Datenmenge von 35 Zetabyte erwartet[61].

Dass es so viel Daten gibt, liegt auch an der zunehmenden Verbreitung von mobilen internetfähigen Geräte. Von der Bevölkerung haben weltweit mehr als 25% und deutschlandweit[62] sogar 50% ein Smartphone.

Die teilweise neuartige Beschaffenheit der Daten um uns herum mag nicht alleiniges Definitionsmerkmal sein[63]. Dennoch impliziert die enorme Masse an Daten viele bedeutsame Eigenschaften, die zum Begriff Big Data gezählt werden. Zwar ist es ein Wechselspiel zwischen dem Angebot, Daten günstiger zu generieren, zu verwalten und zu analysieren und der Nachfrage und damit dem tatsächlichen Auftreten von Datenmengen mit besonderen Eigenschaften. Dennoch ist rein logisch gedacht eine Erkenntnisgewinnung aus großen Datenmengen ohne die Tatsache großer Datenmengen nicht möglich. Auch wenn diese Ansicht nicht von allen geteilt wird[64], hat es in diesem Sinne durchaus seine Berechtigung, dass mit Big Data ein Begriff Verbreitung gefunden hat, der als Schlagwort vom Wortlaut her lediglich auf die Größe anspielt und nicht auf die sich indirekt daraus ergebenden Möglichkeiten.

Datenmengen, die von herkömmlicher Technologie nicht zu verarbeiten und analysieren sind, werden von manchen Autoren mit einer Menge von vielen Terabyte oder Petabyte (1 Terabyte = 1000 Gigabyte, 1 Petabyte = 100.000 Gigabyte) beziffert[65]. Das sind Datenmengen, die die der größten Bibliothek der Welt,

---

[58] Meeker, 2016, S. 90.
[59] Troester, 2011, S. 1.
[60] Zikopoulos et al., 2011, S. 5.
[61] Zikopoulos et al., 2011, S. 5.
[62] Crampton, comScore Insights, 8. Juli 2016.
[63] vgl. King, 2016, S. 2.
[64] vgl. King, 2016, S. 2.
[65] Schroeck et al., 2012, S. 4.

der amerikanischen Library of Congress, um das 100fache überschreiten[66]. Erfasste Datenausschnitte können derart groß sein, dass sie der Gesamtpopulation gleichzusetzen sind[67].

Dabei wird es jedoch oft passieren, dass wenige Datenpunkte häufig vorkommen, die meisten jedoch vergleichsweise selten - damit behalten die Kernfragen klassischer Statistik, nämlich wie man aus wenigen Daten generalisierte Erkenntnisse gewinnen kann, auch in Zeiten von Big Data erhebliche Relevanz[68]. Bei der Analyse vom Einkaufsverhalten in Supermärkten muss man sich so z.B. mit dem Problem auseinandersetzen, dass der durchschnittliche Einkaufsweg nur 25% aller Regale betrifft[69]. Es wird mitunter zu Recht bezweifelt, dass „Big" in absoluten Zahlen groß bedeutet, sondern es um Relativität gehe[70]. Datenmengen müssen nicht einmal Gigabyte erreichen, um eine Analyse als ein Big Data-Problem zu qualifizieren und Big Data-Technologien vorteilhaft zum Einsatz kommen zu lassen. Es kommt mehr auf die Anzahl der Datenpunkte bzw. in Tabellen gesprochen der Zeilen und Spalten an – die Aufzeichnung von Taxifahrten in Boston über sieben Monate von 2012 mag nur 174 Megabyte groß sein[71], dennoch wurde die Analyse dieser über 2 Millionen Zeilen vom Massachusetts Institute of Technology als Big Data-Wettbewerb ausgerufen[72].

Das sogenannte Internet der Dinge („Internet of Things") ist eine weitere Quelle für Massen an Daten. Damit ist das immer größer werdende Netzwerk von Geräten gemeint, die selbstständig, d.h. ohne Benutzerinteraktion, Daten aufnehmen und weitergeben können[73]. Damit sind insbesondere alle Arten von Sensoren gemeint. Zum Wahrnehmen und identifizieren von Geräten dient dabei z.B. die kabellose Technologie RFID (Radio-Frequency Identification), die immer beliebter wird[74]. So regelt im Bereich „Smart Home" ein intelligenter Thermostat die Temperatur; tragbare Fitnessgeräte versuchen Puls und Blutdruck zu messen.

---

[66] Helbing, 2015, S. 76.
[67] vgl. Mayer-Schönberger/Cukier, 2013, S. 29.
[68] vgl. Murphy, 2012, S. 2.
[69] vgl. Sorensen, MR 2003, 30 (32).
[70] Mayer-Schönberger/Cukier, 2013, S. 29.
[71] City of Boston.
[72] Massachusetts Institute of Technology, bigdata@CSAIL, 12. November 2013.
[73] Ashton, RFID Journal, 22. Juni 2009; Internet of Things Global Standards Initiative, 2012, S. 3.
[74] Das/Harrop, 2015.

Auch mit dem Internet kommunizierende Autos, z.B. vernetzt, Carsharing oder selbstfahrende Autos zählen dazu.

Sensoren und viele andere Quellen senden ihre Zustandsänderungen derart oft, dass man von Datenströmen oder Streams spricht. Findet die Analyse in Echtzeit statt, spricht man von Complex Event Processing[75]. Auch Zeitreihenanalysen sind mit diesem Begriff verbunden[76].

2.3.2 Velocity

Social Media Anwendungen verarbeiten Daten heutzutage üblicherweise in Echtzeit – Nachrichten werden mit wenigen Millisekunden Verzögerung versandt, Videos innerhalb von Sekunden in höchster Qualität geladen. Allerdings geht es beim Kriterium Velocity nicht nur um die Geschwindigkeit der Generierung bzw. Übertragung von Daten, sondern insbesondere auch um deren Analyse[77]. Technologien wie Hadoop, MapReduce und Cloud Computing im Allgemeinen helfen hierbei, indem sie die Datenverwaltung verteilbar und skalierbar machen.

Schon seit über 50 Jahren wurden Datenmengen erzeugt und kombiniert, wie das Beispiel der innovativen „Chicago Area Transportation Study“ von 1954-1962 zeigt[78]. Zur Verbesserung des Straßennetzes wurden unter hohen Kosten über 250.000 Befragungen von Haushalten, Taxis und Lastwagenfahrern durchgeführt und mit umfangreichem kartographischem Material kombiniert[79]. Hunderte Mitarbeiter und Spezialisten waren von Nöten. Der gigantische und teure Großrechner benötigte Stunden für Berechnungen die mit Smartphones heute trivial erscheinen müssen[80], waren jedoch jedes Mal aufs Neue maßgeschneidert mit einzigartigen Lochkarten vorzubereiten und dauerten dadurch viele Monate[81]. Heute würde man vielleicht auf kostspielige Befragungen verzichten und einfach GPS-Daten von Taxis mit digitalen Karten kombinieren, um in Echtzeit

---

[75] Luckham, Information Age 2004, 20; Chandy et al., 2011, S. 6.
[76] vgl. Flouris et al., J. Syst. Software 2016, (2).
[77] Schroeck et al., 2012, S. 4.
[78] Plummer, 2010.
[79] vgl. Plummer, 2010, S. 3, 18.
[80] Gewirtz, ZDNet, 13. April 2012.
[81] vgl. Plummer, 2010, S. 18, 24.

automatisiert komplexe Vorhersagen graphisch ansprechend präsentieren zu können[82].

Hadoop ist ein Software-Rahmenwerk, das die flexibel den Ansprüchen anpassbare parallele Verarbeitung von großen Datenmengen auf Hardware ermöglicht, die von der Qualität her günstig und leicht ersetzbar (Commodity Hardware), aber dafür in hoher Quantität vorhanden ist[83]. Parallele Verarbeitung meint damit eine Verteilung. Flexibel meint, dass die Verkleinerung oder Vergrößerung der Rechenkapazität sehr schnell und einfach möglich ist. Dies entspricht Skalierbarkeit. Dazu verwendet Hadoop ein spezielles System zur Ablage von Daten namens „Hadoop Distributed File System“, sowie MapReduce. Hadoop wurde 2007 veröffentlicht.

MapReduce ist ein Programmiermodell und eine damit assoziierte Implementierung zur Verarbeitung und Erzeugung großer Datenmengen, veröffentlicht 2003[84]. Es geht ausdrücklich darum, die durch die oben beschriebenen vielfältigen und massenhaft auftretenden Daten beherrschbar zu machen. Auf tausenden von günstigen Maschinen wird die Bearbeitung der Daten automatisiert aufgeteilt. Dabei wird auch mit redundanter, also mehrfacher Verarbeitung derselben Daten gearbeitet, um eine hohe Geschwindigkeit zu ermöglichen[85]. So kann auch ein Ausfall einer Maschine automatisiert ausgeglichen werden. Viele NoSQL-Datenbanken benutzen ebenfalls MapReduce. Der Videoanbieter Netflix benutzt z.B. hunderte von miteinander verknüpften Rechnern, sogenannten Knoten (englisch node)[86].

Cloud Computing, wörtlich in etwa „Berechnung in einer Wolke“, ist nur vage definiert[87]. Manche sagen, die Idee dahinter sei “so alt und simple wie das Internet selbst“[88]. Es ermöglicht die Auslagerung von Infrastruktur und sonstige IT-Services, die sehr leistungsfähig und flexibel skalierbar, standardisiert und über

---

[82] vgl. Alghuraybi et al., IJDTA 2016, 191.
[83] vgl. Freiknecht, 2014, S. 20f.
[84] Dean/Ghemawat, CACM 2008, 107 (1).
[85] Dean/Ghemawat, CACM 2008, 107 (11); vgl. Freiknecht, 2014, S. 20.
vgl. *Datastax Corporation*, 2013, S. 10.
[87] vgl. Huth/Cebula, 2011, S. 1.
[88] „In a way, the cloud is as old and simple as the Internet itself. The cloud is really just about accessing storage or software remotely from a computer via the Internet. It’s a modern twist on an old concept [...]“, Martin, USA TODAY, 6. November 2011.

das Internet vergleichsweise einfach benutzbar sind[89], inklusive selbstregelnder Eigenschaften und umfassender Werkzeuge zur Überwachung der Auslastung (Monitoring Tools)[90]. Es ist damit sogar Verbrauchern möglich, für wenig oder sogar ohne Geld Speichermöglichkeiten und vieles mehr zu auszulagern[91]. Dropbox bietet z.B. eine private kostenlose Cloud für jedermann, Facebook sogar in unbegrenztem Speicherumfang; wobei die eigenen Daten dort als mindestens teilweise öffentlich einzuordnen sind. Sie können ganz nach individuellem Ressourcenbedarf in Hinsicht auf Zeit und Intensität gebucht werden können. Beliebt sind z.B. Cloud Services von Amazon, welche im Übrigen auf Hadoop aufbauen[92]. Amazon wirbt explizit damit, dass durch Cloud Computing z.B. saisonale Schwankungen etwa zu Weihnachten ausgeglichen werden können[93].

Technologien wie Docker ermöglichen es zudem, die Auslagerung von Infrastruktur auch von einem Software-Standpunkt her wesentlich zu vereinfachen. Die Cloud muss nicht zeitaufwändig für den Betrieb der Anwendung vorbereitet werden, es wird stattdessen einfach ein exaktes Abbild der lokalen Software erschaffen und innerhalb von Sekunden gestartet[94]. Veränderungen und Updates können so sehr einfach und flexibel an die ausgelagerte Infrastruktur weitergeleitet werden.

Skalierung für gigantische Datenmengen ist auch einer der Hauptgründe der Entwicklung der sogenannten NoSQL-Bewegung im Bereich Datenbanken. Der Name steht für Unterschiede zu den traditionellen relationalen Datenbanken, die die Anfragesprache SQL (Structured Query Language) zum Durchsuchen von Daten verwenden und bedeutet „not only SQL“. Die sogenannten relationalen Datenbank-Management-Systeme (RDBMS) waren seit den 70er Jahren der de-facto-Standard (1979 wurde das erste kommerziell erhältliche RDBMS von Oracle veröffentlicht); NoSQL Ansätze waren seit der Jahrtausendwende zu beobachten[95]. Traditionelle RDBMS sind traditionell auf große und teure, da qualitativ

---

[89] Varia, 2008.
[90] vgl. Mell/Grance, 2011, S. 2.
[91] vgl. Martin, USA TODAY, 6. November 2011.
[92] vgl. Varia, 2008.
[93] vgl. Varia, 2008.
[94] vgl. Willis, 2015, S. 5.
[95] vgl. IBM Cloudant, 2010, S. 2.

hochwertige und zentralisierte Hardware wie etwa Großrechner ausgerichtet[96]. Mit der Anforderung, nie dagewesene Datenmengen in höchster Geschwindigkeit mit zugleich hoher Verfügbarkeit (d.h. niedriger Ausfallrate) und Flexibilität zu verarbeiten, geriet diese klassische Kombination von Hardware und Software immer mehr an ihre Grenzen. Als Unternehmen wie Google, Yahoo, Amazon und Facebook bemerkten, dass selbst die schnellsten verfügbaren Großrechner ihre Anforderungen nicht mehr erfüllen konnten oder schlicht zu teuer waren, entwickelten sie entsprechende dezentralisiertere bzw. verteiltere Datenbanksysteme, die schließlich als NoSQL-Datenbanken bekannt wurden[97].

Der Geschwindigkeitsvorteil kann, muss sich aber nicht aus einer Lockerung der Anforderung an die Korrektheit bzw. Aktualität der Daten (sogenannte Konsistenz), damit nicht darauf gewartet werden muss, bis veränderte Daten im System verteilt werden, bevor Datenanfragen beantwortet werden können[98]. Hier ist auch eine Abwägung gefragt, ob z.B. ein Unternehmen für sein System für Flugreservierungen selten auftretende eventuelle Überbuchungen als Preis für schnellere Geschwindigkeit und höhere Verfügbarkeit als die Konkurrenz akzeptiert[99]. RDBMS bleiben das am meisten verbreitete Datenbanksystem, doch geht die Tendenz klar in die Richtung, dass NoSQL-Datenbanken wie etwa MongoDB (erste Veröffentlichung 2009) oder Cassandra (erste Veröffentlichung 2008) immer populärer werden[100].

Zum Bereich NoSQL werden auch graphorientierte Datenbanken gezählt wie etwa Neo4J (erste Veröffentlichung 2007). Aufbauend auf Entwicklungen der hierarchischen und netzwerkorientierten Datenmodelle der 70er[101] werden Daten als Graphen darzustellen, d.h. Informationen werden durch Knoten und Kanten im Sinne der mathematischen Graphtheorie gespeichert. Dies hat den Vorteil, die Vernetzung von Informationen auf natürliche Art und Weise abbilden zu können und dadurch diese Daten auch schneller durchsuchen zu können[102]. Beispiele für Netze sind etwa das Internet mit seinen Links, Facebook mit

---

[96] vgl. Burd, ;login: 2011, (6).
[97] vgl. Burd, ;login: 2011, (5f).
[98] vgl. Burd, ;login: 2011, (7f).
[99] vgl. Cattell, SIGMOD Rec. 2011, 12 (24).
[100] Vgl. das Ranking unter http://db-engines.com/en/ranking
[101] vgl. Angles/Gutierrez, CSUR 2008, 1:1 (3).
[102] vgl. Angles/Gutierrez, CSUR 2008, 1:1 (5).

„Freundschaften“ und Likes, oder Wal-Mart mit den Verknüpfungen zwischen Kunden und gekauften Produkten[103].

Für Vorhersagen über das Verhalten von Menschen und Gruppen, z.B. für Empfehlungssysteme, ist die Analyse ihrer Beziehungen eine naheliegende Erkenntnisquelle. Hier brillieren graphorientierte Datenbanken[104]: Z.B. müssen in RDBMS für Fragen „Wie gut ist die Kreditwürdigkeit der Freunde der betrachteten Person“ die entsprechenden Tabellen (etwa „Personen“, „Freundschaften“, „Kredite“) mit einem Aufwand kombiniert (Table Join) und durchsucht werden, der mit der Anzahl der Freunde und Kredite immer weiter und stark ansteigt. Das liegt daran, dass die Beziehungsdaten noch nicht vorberechnet vorliegen, sondern erst zur Laufzeit, d.h. zur Suchanfrage berechnet werden. In graphorientierten Datenbanken liegen die Beziehungsdaten bereits berechnet vor als „Zeiger“ (Pointer); der Aufwand des Durchlaufens der Verbindungen vergleichsweise konstant[105]. Außerdem ist die Formulierung der entsprechenden Suchanfrage tendenziell komfortabler[106].

Das Dursuchen von Daten zum Auffinden von Informationen ist im Übrigen zu einer eigenen Disziplin von Software geworden. So haben sich Technologien wie Apache Lucene (erste Veröffentlichung 1999) oder das darauf aufbauende Elasticsearch (erste Veröffentlichung 2004) auf das Durchsuchen von vielen Terabyte in Echtzeit und mehr spezialisiert. Damit wird z.B. die Vervollständigung von Sucheingaben bei Wikipedia in Echtzeit[107] für viele Millionen von Benutzern realisiert.

Viele der genannten Technologien sind im Übrigen kostenlos benutzbar oder open-source, wenn auch die individuelle technische Unterstützung meist bezahlt werden muss[108].

Die Zusammenführung von Daten zu Analysezwecken hat verschiedene historische Vorläufer. In den 80ern wurde das Konzept der sogenannten statistischen Datenbanken populär, in den 90er Jahren das sogenannte Data Warehouse (DW)

---

[103] vgl. neo4J, 2015, S. 2f.
[104] vgl. Eifrem, 2015, S. 11ff.
[105] vgl. Eifrem, 2015, S. 17f.
[106] vgl. neo4J, 2015, S. 3ff.
[107] vgl. Gormley/Tong, 2015, S. 1.
[108] IBM Cloudant, 2010, S. 2f.

und Online Analytical Processing (OLAP)[109]. In beiden Fällen geht es vor allem um stark strukturierte, Daten mit mehreren fest definierten Kategorien bzw. Dimensionen. Als Datenmodell sind der statistischen Datenbank vor Allem mehrdimensionale Tabellen eigen, dem DW hingegen der sogenannte Datenwürfel (Data Cube)[110]. In beiden Fällen dominieren Hierarchien die Struktur der Daten[111]. Es handelt sich um eine Sammlung von Daten aus verschiedenen Quellen unter einem vereinheitlichten Schema[112]. Die abstrakten Ziele und Techniken sind ähnlich, doch werden mit statistischen Datenbanken im konkreten vor allem Daten der Statistik wie etwa über den Zensus assoziiert, mit DW in erster Linie Unternehmensdaten, die z.B. über Ort und Zeit mehrdimensional untersucht werden können[113]. Beide Datensysteme unterstützen insbesondere die Aggregierung von Daten[114].

Während in statistische Datenbanken bzw. DW/OLAP vergleichsweise schlichte Zusammenfassungen der Daten im Voraus berechnen (insb. Aggregierungen) und diese Auswertungen anbietet, werden im Data Mining komplexere und insbesondere stark automatisierte Formen der Erkenntnisgewinnung angewandt[115]. Data Warehouses sind im Übrigen weit verbreitet und sind als Datenbasis für weitere Analyseschritte des Data Mining durchaus sehr nützlich[116]. Insbesondere skalieren sie gut, wenn es sich ausschließlich um strukturierte Daten handelt[117].

2.3.3 Variety

Vor dem Aufkommen von Big Data waren tabellenorientierte relationale Datenbanken die Regel. Dabei muss beim Anlegen der Datenbank möglichst schon im Vorhinein feststehen, nach welchen Spalten die Daten strukturiert sind, um sie z.B. durchsuchen zu können. Spätere Änderungen sind aufwändig. Traditionel-

---

[109] vgl. Shoshani, 1997, S. 185.
[110] vgl. Shoshani, 1997, S. 185ff.
[111] vgl. Han et al., 2011, S. 149f.
[112] vgl. Han et al., 2011, S. 10.
[113] vgl. Han et al., 2011, S. 148f.
[114] vgl. Shoshani, 1997, S. 191f.
[115] vgl. Han et al., 2011, S. 167; vgl. Hamel, 2005, S. 8ff.
[116] vgl. Han et al., 2011, S. 6ff.
[117] Han et al., 2011, S. 26.

lerweise handelte es sich vor allem um Daten in Form von Text, Zahlen und Zeitangaben.

Mit sozialen Medien, Massive Open Online Courses (wie etwa Khan Academy oder Coursera.org)[118], sowie dem Streaming von Medien (wie etwa Netflix für Filme und Serien, Spotify für Musik) erhalten Datenarten wie MP3, Bilder und Videos mehr Aufmerksamkeit. Um diese Inhalte katalogisieren, navigieren und durchsuchen zu können, fügt man flexibel jedem Video oder Bild Stichworte in Textform, sogenannte „tags" an. Diese entstehen spontan und unterliegen keiner einheitlichen Regelung. Entscheidender Unterschied ist also, dass die Struktur der Daten nicht im Vorhinein definiert ist, sondern erst im Nachhinein entsteht[119]. Tags können z.B. auch von Benutzern erstellte Bewertungen oder Kommentare sein, wie sie etwa Amazon oder das Videoportal YouTube bieten.

Diese Flexibilität passt zum Paradigma, dass die bezüglich einer Datensammlung zu stellenden Fragen und damit auch die Hierarchien von Information nicht im Vorhinein starr feststehen[120].

Sensoren erzeugen außerdem nicht nur viele Daten wie oben angedeutet, sondern auch besondere Arten von Daten. Durch GPS-Sensoren in Smartphones und Autos entstehen Abfolgen von Geodaten, die mit zeitlichen und sonstigen Angaben verbunden werden können. Geodaten haben auch für viele vergleichsweise alten Industrien herausragende Bedeutung, wie etwa der Kauf des Kartendienstes HERE von Nokia für 2,6 Milliarden € durch BMW, Daimler-Benz und Audi zeigt[121]. Dieser Dienst soll insbesondere für die Entwicklung selbstfahrender Autos und LKWs eine große Rolle spielen. Auch für den sportlichen Bereich gibt es Anwendungen für Verhaltensvorhersagen, insbesondere in Echtzeit, indem etwa vorhergesagt wird, zu wem mit welcher Wahrscheinlichkeit ein Pass erfolgen wird[122].

Durch RFID in den Abonnement-Karten für den ÖPNV in London wird etwa das Ein- und Aussteigen erfasst[123]. Andere Geräte übermitteln Daten zum Luftdruck

---

[118] vgl. Mayer-Schönberger/Cukier, 2013, S. 115.
[119] Mayer-Schönberger/Cukier, 2013, S. 42f.
[120] Mayer-Schönberger/Cukier, 2013, S. 45.
[121] vgl. Daimler AG, 2016, S. 76.
[122] vgl. Yue et al., 2014, S. 2.
[123] vgl. Batty, DHG 2013, 274 (275); van't Hof, 2006, S. 10.

oder der Temperatur. Die Sensoren des Internets der Dinge nehmen auf verschiedenste Art und Weise ihre Umwelt war und erzeugen somit neue Arten von Daten. In der Realität wird es zudem oft vorkommen, dass Datensätze nur ganz bestimmte Merkmalswerte besitzen, und viele andere nicht (sparse data), womit RDBMS Probleme haben (technisch gesehen sehr viele NULL-Werte)[124].

Viele NoSQL-Datenbanken sind ideal für die Verwaltung von unstrukturierten Daten, indem sie diese Art von Daten als sogenanntes Dokument mit einem Schlüssel und dazugehörigen Merkmalen verwalten[125]. Diese Merkmale sind nicht unbedingt im Voraus definiert. Ein Dokument kann z.B. ein einfacher unformatierter Text, ein PDF-Dokument oder ein Video sein. Flexibel können dadurch auch im Nachhinein (Meta-)Strukturen zu Daten erstellt werden, etwa indem man wie oben beschrieben Tags erstellt. Dies ist aber nicht zwingend, grundsätzlich können die Daten auch in roher Form vergleichsweise effizient verwaltet werden. Auch das klassische Data-Warehouse kann zwar grundsätzlich mit derartigen komplexe Datentypen („Complex Types of Data“) umgehen, weshalb es sich durchaus weiterhin als Basis für Analysen anbietet[126]. Dennoch werden häufig NoSQL-Datenbanken von Vorteil sein, gerade, wenn unstrukturierte Daten die Mehrheit ausmachen und es um besonders hohe Geschwindigkeit bei gleichzeitig großen Datenmengen geht (s.o.).

RFID finden darüber hinaus in Bibliotheken Einsatz, um die Inventur und das Ausleihen von Büchern zu vereinfachen[127], oder gar per App den kürzesten Weg zum gewünschten Buch angezeigt zu bekommen[128].

2.3.4 Automatisierte Gewinnung von Erkenntnissen

Automatisierte Methoden zur Datenanalyse, von Strukturen bis hin zu Vorhersagen, werden unter dem Begriff Machine Learning (Maschinelles Lernen) zusammengefasst[129]. Automatisiert lässt sich auch mit dem Wort implizit übersetzen – statt der Software explizit zu sagen, welche Hypothese getestet werden

---

[124] Lindemuth/Lembke, 2013, S. 1.
[125] vgl. Sharma, arXiv:1509.08035 [cs] 2015, (3).
[126] Han et al., 2011, S. 166.
[127] Shashid, LPP 2005.
[128] vgl. Choudari et al., IJIRCCE 2013, 30 (31).
[129] vgl. Murphy, 2012, S. 1.

soll, lernt sie es selbst[130]. Die Vorgehensweise des Computers ist dabei für den, der sie festlegt, nicht zwangsläufig eine absolut unverständliche, geradezu magische „Black Box“ wie mitunter propagiert wird[131], auch wenn das für Außenstehende der Fall sein mag[132]. Allerdings stimmt es, dass es mitunter aufwändig sein kann, Nachvollziehbarkeit zu gewährleisten, womit es auch um Interpretierbarkeit geht[133]. Auf jeden Fall bleiben viele Schritte manuell einzustellen und schrittweise zu überprüfen, wie im Folgenden deutlich wird.

Menschen können Daten vergleichsweise gut im zwei- oder dreidimensionalen Raum analysieren, das bedeutet bei wenigen zu vergleichenden Merkmalen[134]. Stilmittel wie etwa verschiedene Farben, Größen und Formen von Datenpunkten ermöglichen zusätzliche Dimensionen. Dimensionen lassen sich zudem noch reduzieren, indem hochkorrelative Werte zusammengefasst werden. Allerdings ist auch dieser Schritt oft besser automatisiert[135]. Schnell wird man es allerdings mit mehr Dimensionen zu tun haben, gerade wenn es um nicht im Vorhinein und vollständig strukturierte Daten geht. Wie man es auch dreht und wendet, die automatisierte Datenverarbeitung besitzt grundsätzlich starke Vorteile.

Grundsätzlich ist Machine Learning allerdings keineswegs als neu zu bezeichnen, sondern Jahrzehnte alt; bedeutende Meilensteine für das Machine Learning im engeren Sinn wurden bereits in den 80er Jahren gesetzt[136]. Dies gilt auch für Algorithmen wie etwa das Clustering[137].

Knowledge Discovery in Databases (KDD) bzw. Data Mining sind konkrete Anwendungen, die Machine Learning benutzen[138]. KDD bedeutet wörtlich übersetzt die Entdeckung von Erkenntnissen aus Datenbanken. In manchen Definitionen wird hervorgehoben, dass das durch den Prozess gewonnene Wissen nützlich sein[139] und es um die nichttriviale Identifizierung valider, neuartiger, und poten-

---

[130] vgl. Kubat, 2015, S. xi.
[131] vgl. Mayer-Schönberger/Cukier, 2013, S. 179.
[132] vgl. Executive Office of the President, 2016, S. 8.
[133] vgl. Goodman/Flaxman, arXiv:1606.08813 [cs, stat] 2016, (6ff).
[134] vgl. Jain, Pattern Recogn. Lett. 2010, 651 (3).
[135] vgl. Aggarwal, 2015, S. 40f.
[136] vgl. Kubat, 2015, S. xi ff.
[137] vgl. Jain, Pattern Recogn. Lett. 2010, 651 (5f).
[138] Coenen, Knowl. Eng. Rev. 2011, 25 (2).
[139] Fayyad et al., CACM 1996, 27 (28).

tiell nützlicher Strukturen in den untersuchten Daten gehen müsse[140]. In eine ähnliche Richtung gehen Definitionen, die die gewonnenen Erkenntnisse als zunächst verborgen qualifizieren („hidden“) [141]. Es geht also nicht um das Auffinden offensichtlich bestehender Strukturen im Sinne des „Information Retrieval“, z.B. über konventionelle Suchanfragen in tabellenorientierten Datenbanken (SQL-Queries)[142].

Außerdem wird darauf hingewiesen, dass die Methoden und Techniken in Software gegossen sind und über Software umgesetzt werden[143].

Eine Ansicht sieht Data Mining und KDD implizit als Synonym[144]. Wiederum einschränkend wird dabei mitunter vertreten, Data Mining bezeichne im engeren Sinne einen Teilschritt von Data Mining im weiteren Sinne[145]. Data Mining wird von anderen Ansicht dagegen ausschließlich als Teilschritt der KDD gesehen. Data Mining bezeichne demnach lediglich die Algorithmen, die aus den vorbereiteten Daten neue Informationen gewinne, ohne diese Ergebnisse zu visualisieren oder analysieren[146]. Diese Ansicht gibt zu, dass die Begriffe in der Tat sehr häufig synonym verwendet werden[147].

Oft wird betont, dass es um besonders große Datenmengen beliebiger Herkunft geht[148]. Andererseits kann man die neuartigen Methoden und Techniken auch auf konventionelle, d.h. auf vergleichsweise überschaubare und strukturierte Daten anwenden, um zu neuen Einsichten zu gelangen[149]. Mit vielfältigen Daten sind z.B. Sensordaten (Ort, Zeit, Temperatur), kontinuierlich fließende Daten (streams), Multimedia-Daten (Bilder, Videos, mp3) und ganz allgemein das Internet gemeint (soziale Medien, Suchanfragen usw.)[150]. Es gibt hier Parallelen zum Begriff Big Data.

---

[140] Fayyad et al., CACM 1996, 27 (30); Tan et al., 2006, S. 3.
[141] vgl. Coenen, Knowl. Eng. Rev. 2011, 25 (1).
[142] vgl. Coenen, Knowl. Eng. Rev. 2011, 25 (1); vgl. Tan et al., 2006, S. 3.
[143] Coenen, Knowl. Eng. Rev. 2011, 25 (1).
[144] Han et al., 2011, S. 8; Aggarwal, 2015, S. 1; Meinicke, DSRITB 2014, 183 (195).
[145] vgl. Han et al., 2011, S. 8.
[146] Fayyad et al., CACM 1996, 27 (28).
[147] Fayyad et al., CACM 1996, 27 (29).
[148] Han et al., 2011, S. 8; Tan et al., 2006, S. 1; Fayyad et al., CACM 1996, 27 (28).
[149] Beaver et al., 2010, S. 1.
[150] Han et al., 2011, S. 14; vgl. Aggarwal, 2015, S. 6ff.

Passend zu den großen Datenmengen wird überdies auch vertreten, dass der Aspekt der Automatisierung zentral sei, d.h. die vom manuellen menschlichen Eingaben losgelöste Gewinnung von Erkenntnissen[151]. Dies hängt mit der oben erwähnten Verbindung zum Machine Learning zusammen. Trotz der Automatisierungstendenzen bleiben technisches Verständnis und manuelle Entscheidungen z.B. bezüglich der konkreten Vorgehensweise unabdingbar[152].

Neben der Automatisierung zeigt sich aber auch in den Datenquellen ein großer Unterschied zum klassischen statistischen Vorgehen[153]. Dieses geht eher von Stichprobenartig für einen bestimmten Zweck manuell erhobenen und im Vorhinein klar strukturierten Daten aus, wie z.B. Meinungsumfragen[154]. KDD möchte sich hingegen mit der Gesamtheit der Massen an Daten befassen, die nicht nur, aber doch häufig rein automatisiert, unstrukturiert und mit offen gehaltener Verwendungsabsicht erstellt wurden[155].

Die Idee, unabhängig von Annahmen und Hypothesen Daten zu erkunden oder auszukundschaften, ist im Übrigen bereits seit den 70er Jahren umfassend unter dem Begriff exploratorische Datenanalyse (EDA) diskutiert worden[156]. Dabei geht es auch, aber nicht nur, um Methoden der Visualisierung[157]. Die Zielsetzungen sind also ähnlich. Manche gehen dabei soweit, Big-Data-Analyseverfahren wie die KDD gewissermaßen als EDA nach dem aktuellen Stand der Technik zu bezeichnen[158]. Dieser Auffassung ist zu entgegnen, dass die EDA eher mit der klassischen Statistik assoziiert wird im Sinne einer Vorstufe zur Bildung von besseren Hypothesen[159]. EDA wird aber zum Teil auch als Gegensatz zur klassischen Statistik und im Sinne einer von konkreten Methoden losgelösten Philosophie definiert, durch Unvoreingenommenheit natürliche und versteckte Muster in den Daten zu entdecken[160]. Treffender ist die Aussage, KDD ist automatisierte und in vielen Belangen wesentlich weiterentwickelte EDA. Damit können die

---

[151] Tan et al., 2006, S. 1f; Fayyad et al., CACM 1996, 27 (29); Meinicke, DSRITB 2014, 183 (195); Goebel/Gruenwald, SIGKDD Explor. 1999, 20 (20).
[152] vgl. Edelstein, 1999, S. 1–2.
[153] vgl. Wernecke, 1995, S. 12.
[154] vgl. Wernecke, 1995, S. 15f.
[155] Aggarwal, 2015, S. 1ff.
[156] Tukey, 1977.
[157] vgl. Wernecke, 1995, S. 26ff; vgl. Tan et al., 2006, S. 97.
[158] vgl. Richter, DuD 2016, 581.
[159] vgl. Tan et al., 2006, S. 97f.
[160] vgl. National Institute of Standards and Technology, 2003.

Erkenntnisse aus der jahrzehntelangen Erfahrung mit der EDA durchaus für die KDD fruchtbar gemacht werden[161].

Um also aus gegebenen Daten über Menschen Gruppen und darauf aufbauend Prognosen zu bilden, bieten sich KDD und Data Mining offensichtlich an. Die Klassische Statistik hat mitnichten ausgedient, sondern fügt sich als einflussreicher Teil harmonisch ein, nicht nur, um die automatisiert gewonnenen Erkenntnisse im Nachhinein statistisch zu überprüfen[162].

[161] vgl. Goebel/Gruenwald, SIGKDD Explor. 1999, 20 (22f).
[162] vgl. Siebes, 1994, S. 553f.

3 Technische Vorgehensweise

Sowohl in der klassischen Statistik, als auch neueren Methoden werden deskriptive, explorative und induktive („predictive“) Aufgaben unterschieden[163].

Eine Strukturierung in Gruppen charakterisiert die Daten und gehört damit grundsätzlich zum Feld der deskriptiven[164] (beschreibenden) und der damit sehr eng verbundenen explorativen (erkundenden) Statistik[165]. Synonym im Bereich Machine Learning wird der Begriff „Unsupervised Learning“ verwendet[166]. Das „lernende“ Vorgehen ist „unbeaufsichtigt“, da die Gruppen selbstständig, gleichsam einer natürlichen Ordnung folgend gebildet werden, ohne mit bereits vorhandenen Strukturen verglichen zu werden. Wenn man die Strukturierung automatisiert[167] aus einer gegebenen Datenmenge erschafft, spricht man von Clustering[168].

Im Gegensatz dazu gehört die Vorhersage von Verhalten gehört induktive (schließende, inferentielle) Statistik, d.h. aus der gegebenen Beschreibung der Daten werden Schlussfolgerungen gezogen. Dazu gehört auch die Ermittlung, zu welcher Gruppe eine neue Person gezählt wird[169]. Im Jargon des Machine Learning wird diese Art von Algorithmen „Supervised Learning“ genannt[170]. Das „Lernen“ ist „überwacht“, da zu strukturierende Daten mit bereits vorhandenen Strukturen, nämlich den bereits definierten Gruppeneigenschaften verglichen werden.

Allerdings gibt es auch Mischformen, wie etwa das „Probabilistic Clustering“ [171].

---

[163] vgl. Han et al., 2011, S. 15; vgl. Teschl/Teschl, 2014, S. 209; Siraj/Ali, 2011, S. 54.
[164] Siraj/Ali, 2011, S. 54; vgl. Han et al., 2011, S. 15.
[165] vgl. Coenen, Knowl. Eng. Rev. 2011, 25 (3); vgl. Teschl/Teschl, 2014, S. 209; vgl. Jain, Pattern Recogn. Lett. 2010, 651 (2).
[166] Murphy, 2012, S. 2.
[167] Han et al., 2011, S. 444.
[168] vgl. Wernecke, 1995, S. 185; vgl. Aggarwal, 2015, S. 153.
[169] vgl. Siraj/Ali, 2011, S. 54.
[170] Murphy, 2012.
[171] vgl. Han et al., 2011, S. 467ff.

3.1 Teilschritte der Knowledge Discovery in Databases

Der Prozess der KDD wird logisch in verschiedene Aufgaben unterteilt, die iterativ, also flexibel schrittweise wiederholt und verfeinert werden[172].

Die KDD beginnt zunächst mit der Spezifizierung des fachlichen Ziels[173], hier also damit, welches gruppenspezifische Verhalten genau vorhergesagt werden soll. Danach sind die entsprechenden Daten zu gewinnen bzw. zu sammeln.

Wieviel Aufwand dies bedeutet, hängt stark vom Einzelfall ab. Deshalb sparen manche Autoren Details hierzu aus[174]. Andere legen darauf mehr Wert[175]. Aus wirtschaftlichen Aspekten wird es meist sinnvoll sein, für den Prozess der KDD eine von der normalen Datenbasis eines Unternehmens separate, gewissermaßen duplizierte Datenbasis zu schaffen[176]. Auch lassen sich so gegebenenfalls rechtliche Ziele besser berücksichtigen, wenn z.B. ausschließlich aggregierte bzw. anonymisierte Daten übernommen werden. Da es um gruppenspezifisches Verhalten geht, wird man oft auf Informationen verzichten können, die Individuen identifizieren.

Über die genaue Reihenfolge der folgenden Teilschritte der Datenvorbereitung herrscht Uneinigkeit. Es kann mit Teilen der Datenreduktion und -transformation (data reduction and transformation) begonnen werden, indem zuerst die gewünschten Merkmale aus den rohen Daten ausgewählt und die Daten übertragbar gemacht werden (feature extraction and data portability[177] bzw. selection[178]), um diese Daten dann zu säubern und zu integrieren (data cleaning and integration) und schließlich zu reduzieren und transformieren (data reduction and transformation). Die meisten Autoren fangen allerdings sofort mit der Datensäuberung und -integration der gesamten Rohdaten an[179].

---

[172] vgl. Aggarwal, 2015, S. 3ff; vgl. Han et al., 2011, S. 6ff; vgl. Fayyad et al., CACM 1996, 27 (30f); vgl. Tan et al., 2006, S. 3.
[173] vgl. Aggarwal, 2015, S. 3ff; vgl. Han et al., 2011, S. 6ff; vgl. Fayyad et al., CACM 1996, 27 (30f); vgl. Tan et al., 2006, S. 3.
[174] vgl. Han et al., 2011, S. 6ff; vgl. Tan et al., 2006, S. 3.
[175] Fayyad et al., CACM 1996, 27 (30); Edelstein, 1999, S. 23f; Aggarwal, 2015, S. 3.
[176] vgl. Edelstein, 1999, S. 2; vgl. Aggarwal, 2015, S. 53.
[177] vgl. Aggarwal, 2015, S. 27f.
[178] vgl. Edelstein, 1999, S. 25.
[179] vgl. Han et al., 2011, S. 85f; vgl. Fayyad et al., CACM 1996, 27 (30).

3.1.1 Datensäuberung

Die Säuberung der Daten betrifft Fehler, die bei der Gewinnung entstanden sind. Viele Daten haben fehlende, ungenaue, veraltete, unglaubwürdige, unverständliche oder auf sonstige Weise fehlerbehaftete Werte. Damit sind nicht nur grundsätzlich fehlerbehaftete manuell erstellte Daten gemeint – auch Sensordaten sind notorisch ungenau; Fotographische Scanner erkennen nicht jedes Muster korrekt, batteriegetriebene GPS-Sensoren z.B. eines Smartphones senden außerdem aus Energiemangel nicht durchgehend[180]. Das Fehlen von Daten kann auch vergleichsweise verborgen sein („disguised missing data"), indem etwa bei Fehlen eines Datums automatisch immer das Datum 1. Januar eingetragen wird[181], z.B. auch weil Benutzer aus Anonymitätserwägungen willentlich keine Angaben machen[182]. Hier sind subjektive Entscheidungen gefragt, ob derartige Werte ignoriert oder wenn möglich korrigiert werden sollen, gegebenenfalls auch durch Schätzung[183]. Dabei ist aber zu beachten, dass gerade das Fehlen von Werten signifikante Aussagekraft haben kann[184]. Um z.B. störende, fehlerhafte Ausreißer als „Rauschen" visuell erkennen und beseitigen zu können, werden vielfach visuelle Methoden wie die explorativen Datenanalyse benutzt[185].

Die Verarbeitung von Daten mit viel Rauschen (noisy data) und vielen fehlenden Werten (sparse data) bleibt aber auch im Zeitalter von Big Data und Machine Learning anspruchsvoll und schwierig[186].

3.1.2 Datenintegration

Um möglichst nützliche bzw. verlässliche Gruppendefinitionen und Vorhersagen zu bilden, werden oft Daten aus verschiedenen Quellen zusammengeführt. Das kann Herausforderungen bereiten, die im speziellen als Problem der Daten-

---

[180] vgl. Aggarwal, 2015, S. 34f; vgl. Han et al., 2011, S. 88ff.
[181] Han et al., 2011, S. 84.
[182] Aggarwal, 2015, S. 34.
[183] vgl. Tan et al., 2006, S. 40f.
[184] Edelstein, 1999, S. 25.
[185] vgl. Han et al., 2011, S. 89.
[186] vgl. Junqué de Fortuny et al., Big Data 2013, 215 (223).

integration bezeichnet werden[187]. Dabei geht es insbesondere um Inkonsistenzen zwischen Datenquellen. Bei redundant vorhanden Daten kann oft nur ein Fachspezialist erkennen, welche von mehreren Quellen die richtigen oder wenigstens am wenigsten fehlerbehafteten Werte liefert. Auch Synonyme und allgemein missverständliche Datenbezeichnungen können dabei zu Problemen bei der Synthese der Daten führen.

3.1.3 Datenreduktion und -transformation

Schließlich gehört zur Vorbereitung auch die Frage der Datenreduktion und Transformation[188]. Zum einen benötigt man oft nur einen Teil der Daten, zum anderen sollen unterschiedliche Skalen und Wahrscheinlichkeitsverteilungen vergleichbar gemacht werden.

Zunächst einmal sind die Rohdaten oft noch nicht einmal im gleichen, von der KDD/Data Mining Software lesbaren Format. Dies betrifft insbesondere Sensordaten, aber auch Bilderdaten werden für eine schnellere Verarbeitung komprimiert oder gar der Inhalt in Worte übersetzt[189]. Dokumente in Textform, die Informationen über Gruppenindividuen enthalten, müssen unter Umständen aufwändig und mit viel fachspezifischem Spezialwissen in eine andere Form gebracht werden; man denke nur an Wortendungen, Spitznamen oder Abkürzungen[190].

Aufgrund der Komplexität und den vielen zu beachtenden subjektiv zu beurteilenden Feinheiten und manuellen Arbeitsschritten ist diesen Aufgaben eine künstlerische Natur zu attestieren[191]. Die Zielsetzung und Auswahl bzw. Kombination der Daten kann einen äußerst kreativen Akt darstellen. Die Schritte der KDD bis hierhin können bis zu 90% der gesamten Arbeit ausmachen[192].

---

[187] Han et al., 2011, S. 93ff; Edelstein, 1999, S. 25.
[188] Aggarwal, 2015, S. 37f; Han et al., 2011, S. 99ff, 111ff.
[189] Aggarwal, 2015, S. 54f; vgl. Han et al., 2011, S. 86.
[190] vgl. Aggarwal, 2015, S. 55; vgl. Han et al., 2011, S. 123.
[191] Aggarwal, 2015, S. 30.
[192] vgl. Edelstein, 1999, S. 23.

3.2 Analytische Verarbeitung

Nun kann die Analytische Verarbeitung mit Hilfe von Algorithmen bzw. das Data Mining im engeren Sinn beginnen[193], um zur Vorhersagbarkeit gruppenspezifischen Verhaltens zu gelangen.

Dabei können die Algorithmen grundsätzlich in vier Kategorien aufgeteilt werden[194].

1. Gruppierung (Clustering)
2. Identifikation von Strukturen (Pattern Extraction)
3. Klassifizierung
4. Outlier Detection

Eine wunschgemäße Strukturierung in Gruppen wird mit Big Data selten bereits gegeben sein. Dabei begnügt man sich nicht mit den eher manuellen Methoden der EDA, sondern greift auf die mächtigen Methoden des Clustering zurück, die Automatisierung nutzen[195]. Was der passende Algorithmus für das gegebene Analyseziel ist, hängt vom Fall und subjektiven Vorstellungen ab[196]. Clustering kann neben dem Gewinn von Erkenntnissen auch zu Datenkomprimierungszwecke dienen[197].

Die Ähnlichkeit zwischen zwei Datenpunkten, d.h. Personen mit ihren Merkmalen, wird üblicherweise dadurch berechnet, indem man ihre Merkmale in Matrizen übersetzt und miteinander vergleicht[198]. Grundsätzlich wird Ähnlichkeit dabei als Abstand bzw. Distanz definiert[199]. Wie das Abstandsmaß berechnet wird, hat erheblich Auswirkung auf die ermittelten Ergebnisse[200]. Üblich sind insbesondere die (quadrierte) Euklidische oder die Manhattan-Distanz. Insbesondere für Daten, die viele fehlende Merkmale aufweisen, ist im Übrigen eine probabilistische Zuordnung von Merkmalen sinnvoll (Probabilistic Cluste-

---

[193] Aggarwal, 2015, S. 63ff; Han et al., 2011, S. 243ff.
[194] Coenen, Knowl. Eng. Rev. 2011, 25 (2ff); vgl. Han et al., 2011, S. 15–21.
[195] Tan et al., 2006, S. 847ff; Aggarwal, 2015, S. 153ff; Han et al., 2011, S. 443ff.
[196] vgl. Estivill-Castro, SIGKDD Explor. 2002, 65 (65); vgl. Jain, Pattern Recogn. Lett. 2010, 651 (3).
[197] vgl. Tan et al., 2006, S. 489; vgl. Jain, Pattern Recogn. Lett. 2010, 651 (4f).
[198] vgl. Murphy, 2012, S. 875.
[199] vgl. Wernecke, 1995, S. 187.
[200] vgl. Aggarwal, 2015, S. 63.

ring)[201]. Die fehlenden Daten werden geschätzt und die Korrektheit dieser Schätzung mit einer Wahrscheinlichkeit versehen[202].

Pattern extraction[203]: Hiermit werden interessante Korrelationen zwischen beobachteten Eigenschaften und Verhaltensweisen innerhalb der Gruppen aufgedeckt und schließlich Prognosen mit Wahrscheinlichkeiten aufgestellt. Techniken der klassischen Statistik sind z.B. Korrelationsanalyse und lineare Regression[204], sowie die Multivariate Varianzanalyse[205].

Classification[206]: In diesem Schritt geht es darum, zu welcher Gruppe eine in den Daten (noch nicht zu einer Gruppe gehörende) Person zu zählen ist. Um dies festzustellen, werden die konkreten zu dieser Person gesammelten Werte mit den abstrakten Werten der vorhandenen Gruppen verglichen und kategorisiert. Die klassische Statistik nennt diese Aufgaben Diskriminanzanalyse[207].

Outlier Detection[208]: Dadurch können außergewöhnliche Personen, Gruppen und Verhaltensweisen aufgespürt werden.

3.2.1 Clustering

Da es um einen auch subjektiven Vorgang geht[209], ist es kein Wunder, dass es eine Vielzahl an Varianten[210] von Cluster-Algorithmen gibt.

Ziel ist in jedem Fall, eine intrinsische bzw. natürliche Ordnung der Daten zu finden; dabei handelt es sich um einen subjektiven Vorgang[211].

Die Clustering-Methoden lassen sich in zwei große Gruppen aufteilen; partitionierende und hierarchische Algorithmen[212]:

---

[201] vgl. Han et al., 2011, S. 467.
[202] vgl. Estivill-Castro, SIGKDD Explor. 2002, 65 (66).
[203] Tan et al., 2006, S. 327ff; Aggarwal, 2015, S. 93ff; Han et al., 2011, S. 243ff.
[204] vgl. Wernecke, 1995, S. 113ff.
[205] vgl. Wernecke, 1995, S. 148ff.
[206] Tan et al., 2006, S. 145ff; Aggarwal, 2015, S. 285ff; Han et al., 2011, S. 327ff.
[207] vgl. Wernecke, 1995, S. 161.
[208] Aggarwal, 2015, S. 237ff; Han et al., 2011, S. 543ff; vgl. Tan et al., 2006, S. 651.
[209] Jain, Pattern Recogn. Lett. 2010, 651 (3).
[210] Estivill-Castro, SIGKDD Explor. 2002, 65.
[211] Jain, Pattern Recogn. Lett. 2010, 651 (3).
[212] vgl. Tan et al., 2006, S. 491; vgl. Han et al., 2011, S. 448ff; vgl. Jain, Pattern Recogn. Lett. 2010, 651 (6); vgl. Murphy, 2012, S. 875; vgl. Fraley/Raftery, Comput. J. 1998, (579).

Partitionierende Methoden sind strukturell einfacher, denn sie unterteilen die vorhandenen Daten in Gruppen, ohne Untergruppen zuzulassen. Algorithmen, die sich überlappende oder fuzzy Zugehörigkeiten zulassen, sind grundsätzlich möglich, wenn auch komplizierter.

Hierarchische Methoden gruppieren die Personen verschachtelt gemäß einer Baumstruktur. Die Wurzel des Baums enthält alle Gruppen, die Blätter sind unter Umständen einelementige Mengen (Singletons), d.h. Personen. Wiederum lassen sich Bottom-Up bzw. agglomerative und Top-Down bzw. divisive Methoden unterschieden. Bei der Bottom-Up-Methode geht man von unten nach oben vor: es gibt zu Anfang genauso viele Gruppen, wie es Personen gibt. Schritt für Schritt werden ähnliche Personen zu Gruppen und ähnliche Gruppen zu größeren Gruppen zusammengefügt, bis alle Personen erfasst und alle Gruppen mit einer gemeinsamen Obergruppe, der Wurzel des Baums, umfasst sind. Bei der Top-Down Methode geht man umgekehrt vor: zu Anfang gibt es nur eine einzige Gruppe, die alle Personen umfasst, und die dann geteilt wird.

Welche Anzahl an Clustern bzw. Gruppen optimal ist, ist eine vor allem subjektiv und manuell festzulegende Frage[213].

k-means clustering (K-Mittelwert-Clustering) ist der populärste Clustering-Algorithmus und kann auf über 50 Jahre Geschichte zurückgreifen[214]. Charakteristisch ist die sogenannte Centroid-Methode[215]: nach der jeweiligen Teilung der Gruppen im Sinne von Top-Down, wird diejenige Gruppe, zu dessen Zentrum die jeweilige Person den geringsten Abstand im Sinne euklidischer Distanz hat, zur Gruppe der betrachteten Person. Dies wird iterativ solange wiederholt, bis sich die Centroiden nicht mehr verändern.

3.2.2 Pattern Extraction

Das Ergebnis der Analyse von Zusammenhängen wird im Ergebnis des KDD üblicherweise eine aussagenlogische Antwort zur Folge haben („wenn A, dann B“), die mit Werten für die beiden Maße für Interessantheit Support (Unterstüt-

---

[213] vgl. Fraley/Raftery, Comput. J. 1998, (580).
[214] vgl. Jain, Pattern Recogn. Lett. 2010, 651; Bock, JEHPS 2008, (1).
[215] vgl. Wernecke, 1995, S. 200; vgl. Tan et al., 2006, S. 497.

zung) und Confidence (Zuversicht) verknüpft ist[216]. Interessantheit meint Nützlichkeit und Sicherheit der entdeckten Regeln. Der Wert für Support gibt an, wie häufig die betrachtete Aussage auf die Daten anwendbar ist. Der Wert für Confidence gibt an, wie oft in den beobachteten Daten die Aussage wahr ist.

Beispielsweise wird bei einer Analyse von gekauften Produkten zu Kunden herauskommen, dass der Kauf von Tomatensauce den Kauf von Spaghetti impliziert, mit den Werten Support = 2% und Confidence = 50%. Das bedeutet, dass dieser Fall in 2% aller erfassten Zusammenhänge vorkommt und diese Aussage in 50% aller beobachteten Fälle wahr ist.

Üblicherweise setzt man fest, ab welchen Schwellenwerten Daten als interessant gelten sollen. Hier werden oft fachliche Experten zu Rate gezogen werden[217].

In jedem Fall ist zu beachten, dass es um Korrelationen geht, und nicht um Kausalität. D.h. es geht um das gleichzeitige Auftreten von zwei Ereignissen, nicht um einen Zusammenhang im Sinne Ursache-Wirkung[218]. So können etwa zwei Ereignisse miteinander aus reinem Zufall korrelieren, oder aufgrund einer dritten, beide Ereignisse verursachenden Variable.

3.2.3 Classification

Classification ist ein zweistufiger Prozess[219]. In der klassischen Statistik wird die Vorgehensweise unter dem Stichwort Diskriminanzanalyse betrachtet[220].

Zunächst wird ein mathematisches Modell (Training Model[221]) definiert, durch die bereits erfassten Merkmale zu den bereits definierten Gruppen zugeordnet werden können[222]. Die Güte des Modells wird dabei anhand einer Konfusionsmatrix bewertet:

---

[216] vgl. Tan et al., 2006, S. 328ff; vgl. Han et al., 2011, S. 245ff.
[217] vgl. Han et al., 2011, S. 245.
[218] vgl. Tan et al., 2006, S. 330.
[219] vgl. Han et al., 2011, S. 329; vgl. Aggarwal, 2015, S. 286.
[220] vgl. Wernecke, 1995, S. 162.
[221] Aggarwal, 2015, S. 286.
[222] vgl. Tan et al., 2006, S. 146; vgl. Wernecke, 1995, S. 162.

		Vorhergesagte Gruppe	
		Gruppe = 1	Gruppe = 0
Tatsächliche Gruppe	Gruppe = 1	f_{11}	f_{10} (Typ II / β)
	Gruppe = 0	f_{01} (Typ I / α)	f_{00}

Abbildung 2: Konfusionsmatrix[223]

f_{ij} steht für die jeweilige Anzahl an Personen von Gruppe i die zur Gruppe j zugeordnet wurden. Das Konzept ist vergleichbar mit dem Prinzip der klassischen Statistik, Hypothesen anhand von Fehlern von Typ I bzw. α und II bzw. β (false positives und false negatives) zu bewerten[224]. Im Anschluss kann die Genauigkeit (Accuracy) des Modells berechnet werden. Nur Modelle mit hoher Trefferwahrscheinlichkeit sind nützlich[225]. Dabei ist aber zu beachten, dass sich das Modell beim Lernen auch übertrieben an gewisse Eigenheiten der betrachteten „Trainingsdaten" anpassen kann, die keine generelle Gültigkeit haben und schließlich bei der Anwendung des Modells auf noch nicht erfasste Personen sehr viele Fehler macht[226]. Dieses Phänomen nennt man Model Overfit. Das Gegenteil, ein zu grobes Model wird als Model Underfit bezeichnet.

Ein weit verbreitetes mathematisches Modell zur Classification sind Entscheidungsbäume (Decision Trees)[227], die die Personen nach ihren Merkmalen in Gruppen einteilen. Jeder innere Knoten ist eine Entscheidung, die binär mit ja oder nein beantwortet wird und damit der jeweiligen Abzweigung gefolgt wird. Die Blätter legen schließlich die Gruppenzugehörigkeit fest[228].

Andere Methoden der Classification sind regelbasierte Verfahren, die viele aussagenlogische Regeln („wenn, dann") aufstellen[229] und Probabilistische Verfahren, die mit Wahrscheinlichkeiten z.B. nach Bayes funktionieren[230].

---

[223] Tan et al., 2006, S. 149.
[224] vgl. Wernecke, 1995, S. 62f.
[225] vgl. Wernecke, 1995, S. 166.
[226] vgl. Han et al., 2011, S. 330.
[227] vgl. Tan et al., 2006, S. 150.
[228] vgl. Han et al., 2011, S. 330f.
[229] vgl Aggarwal, 2015, S. 298ff; vgl. Han et al., 2011, S. 355ff.
[230] vgl. Han et al., 2011, S. 350ff; vgl. Aggarwal, 2015, S. 306ff.

3.3 Interpretation und Visualisierung

Die Interpretation und Visualisierung der Ergebnisse schließt die KDD ab. Visualisierung ist dabei von großer Bedeutung[231], indem die Ergebnisse nachvollziehbar und unmissverständlich zum Ausdruck gebracht werden[232]. Einerseits ist sie auch Interpretation der Ergebnisse, andererseits macht sie die Ergebnisse auch besser interpretierbar. Menschen besitzen nämlich herausragende visuelle Fähigkeiten, Muster zu erkennen[233]. Wiederum können Mittel der explorativen Datenanalyse weiterhelfen. Manche vergleichen die Bedeutung des künstlerischen Aspekts mit dem des klassischen Industrie-Designs und ziehen Parallelen zur Wichtigkeit von Steve Jobs und Jonathan Ivy von Apple[234].

Die Bedeutung und die Verfügbarkeit von Anwendungen, die durch Visualisierung bessere Einsicht in die Daten verschaffen, wächst[235]. Sogenannte Dashboards (Instrumententafeln) bereiten Informationen von Anwendungen in Echtzeit auf, z.B. zur Anzahl und Herkunft von Besuchern einer Website. Visualisierung von Big Data ist derart wichtig geworden, dass der oben erwähnte Wettbewerb des Massachusetts Institute of Technology zwei getrennte Kategorien für die Gewinner hatte: Visualisierung und Vorhersage[236].

Ein Beispiel für eine die Visualisierung einer Big-Data-Vorhersage von gruppenspezifischem Verhalten ist die Darstellung von Satelliten- und Verkehrsüberwachungsdaten in Echtzeit[237].

In jedem Fall ist zu beachten, dass ausschließlich Korrelationen, keine Kausalitäten dargelegt werden. Die Daten, auf denen die Analyse aufbaut, sind zudem häufig fehlerhaft. Zudem ist das gesamte Verfahren von subjektiven und allgemein fehleranfälligen Entscheidungen geprägt. Dies gilt umso mehr, als es nicht um rein technische oder empirische Fragestellungen geht, sondern um die Vorhersage von menschlichem Verhalten. Nachvollziehbar darzustellen, welche Merkmale entscheidend waren für eine Prognose, kann ein Problem für sich

---

[231] Schroeck et al., 2012, S. 12.
[232] Han et al., 2011, S. 56ff.
[233] Goebel/Gruenwald, SIGKDD Explor. 1999, 20 (22).
[234] vgl. Davis, Forbes, 30. Juni 2015.
[235] vgl. BITKOM, 2014, S. 75f.
[236] Massachusetts Institute of Technology, bigdata@CSAIL, 12. November 2013.
[237] vgl. Executive Office of the President, 2014, S. 6.

sein, da besonders gute Vorhersagen oft auf besonders komplexen Algorithmen basieren[238].

Aus all diesen Gründen sind Vorhersagen gruppenspezifischen Verhaltens mit Vorsicht zu genießen, auch wenn sie auf gigantischen Datenmengen basieren und mit neuesten statistischen Methoden automatisiert analysiert wurden.

[238] vgl. Goodman/Flaxman, arXiv:1606.08813 [cs, stat] 2016, (6ff).

4 Chancen und Risiken

In den vorherigen Kapiteln wurden bereits verschiedene innovative und wirtschaftlich erfolgreiche Big Data Lösungen genannt. Im Folgenden wird sich auf angewandte Beispiele rund um Geodaten konzentriert, um das Potential der Verbesserung der Situation verschiedener Stakeholder anschaulich zu machen. Werden in Amerika eher die Risiken staatlicher Bevormundung hervorgehoben[239], so werden in insbesondere Deutschland[240] bzw. in der EU traditionell die Risiken von Freiheit bzw. die Chancen von hoheitlichen Eingriffen betont[241] – hier sollen aber auch die Chancen von Big Data nicht zu kurz kommen[242]. Einige Beispiele wurden bereits im Kapitel zu Big Data erwähnt. Unternehmen können durch Big Data ihre Wettbewerbsfähigkeit steigern und neuartige oder stark verbesserte Services anbieten, Verbraucher kommen in den Genuss dieser Dienste und erhöhen damit ihren Wohlstand. Hoheitliche Akteure auf Landes-, Regions- oder Gemeindeebene erhalten durch die Vorhersagbarkeit gruppenspezifischen Verhaltens durch Big Data ebenfalls die Möglichkeit, die Wohlfahrt zu steigern, indem sie öffentliche Aufgaben effizienter wahrnehmen können.

Die Kehrseite von Chancen sind Risiken, deren Art und Ursachen im Anschluss näher beleuchtet werden.

4.1 Chancen

Big Data kann neue, nie dagewesene Möglichkeiten schaffen. Erst durch die Analysefähigkeit von großen Mengen an Sensordaten ist z.B. der Bau der ersten unterirdischen Stromtrasse Deutschlands möglich geworden[243].

Die Big-Data-Analyse von Geodaten ermöglicht zudem die Entwicklung von neuartigen Methoden, die z.B. Stadtbewohner zur einfacheren Kommunikation mit den Behörden über Probleme zu befähigen. Die Stadt Boston hat eine App für

---

[239] vgl. Cavoukian, iapp.org Privacy Perspectives, 18. April 2014.
[240] vgl. Author, The Huffington Post, 17. Juli 2014.
[241] vgl. Determann, NVwZ 2016, 561 (565).
[242] vgl. BITKOM, 2012, S. 46.
[243] vgl. Wetzel, Welt Online, 22. September 2015.

Smartphones entwickelt, mit der Straßenschäden wie etwa Schlaglöcher der Stadtverwaltung gemeldet werden können[244]. Ein ähnliches Angebot bietet das Ordnungsamt Berlin seit dem 1. Juli 2016 an, um auf illegalen Müll, falsch parkende Autos und Lärmbelästigung aufmerksam zu machen[245]. Verbunden mit Daten über besondere Vorkommnisse wie Festivals oder Konzerte in der Stadt könnten z.B. genauere Vorhersagen über den Bedarf an Reinigungs- und Sicherheitspersonal getroffen werden.

Über RFID-Chips unterm Einkaufswagen kann das Verhalten von Kunden beobachtet werden[246]. Kombiniert mit den Zahlungsdaten können so Kundenkategorien gebildet und das Angebot darauf hin optimiert werden[247].

Dynamische geografische Daten werden auch zur Entscheidung über die Vergabe von Krediten integriert[248]. Dadurch kann die Kreditvergabe potentiell genauer oder einfacher die Kreditwürdigkeit von Kunden einschätzen, was zu einer Verbesserung der Konditionen und Gerechtigkeit für Kreditnehmer führen kann und damit volkswirtschaftliches Potential bietet, was wiederum Steuereinnahmen generiert.

Eine interessante Quelle für Daten, die verschiedenste Bereiche umspannt, umfangreiche Datensätze enthält und vor Allem für jedermann frei zugänglich ist, sind hoheitliche Daten, die im Sinne von Open Data verfügbar gemacht werden. Ziel ist die Förderung von Demokratie durch Transparenz, die Verbesserung der wirtschaftlichen Möglichkeiten und die Verbesserung der Lebensverhältnisse der Bürger, wie es in der US-amerikanischen Open Government Initiative von Präsident Barack Obama 2014 heißt[249].

Die Stadt New York bietet beispielsweise umfassende Daten über Taxi-Fahrten ab 2009 in standardisiertem Format an, wahlweise per Download oder Schnittstelle zur Anwendungsprogrammierung[250]. Jede Fahrt besitzt die Attribute Fahrt-, Taxi- und Fahrernummer, Start- und Endpunkt inklusive Datum und Zeit, Kosten, Trinkgeld und Maut. Alleine die Fahrten der Jahre 2009, 2011 und

---

[244] vgl. Executive Office of the President, 2014, S. 51.
[245] rbb-online, 12. Juni 2016.
[246] vgl. Larson et al., Int. J. Res. Mark. 2005, 395.
[247] vgl. Sorensen, MR 2003, 30.
[248] vgl. BITKOM, 2015, S. 80f.
[249] vgl. Executive Office of the President, 2014, S. 11ff.
[250] The City of New York, NYC Open Data, 2015 2009.

2012 belaufen sich auf über 500 Millionen, wobei die Größe vergleichsweise überschaubar mit 120 Gigabyte bleibt[251].

Diese und ähnliche Daten mit modernen Mitteln auszuwerten bietet Potential insbesondere für die Raum- und Stadtplanung. Der Vergleich mit der „Chicago Area Transportation Study" von oben treibt den Unterschied Analysemethoden vor Big Data auf die Spitze, indem Verhaltensvorhersagen heutzutage mit minimalen Kosten und sogar von Einzelpersonen in Echtzeit durchgeführt werden können. Vorteile liegen damit insbesondere auch in der Verwendung aktueller dynamischer Quellen, auf die zeitnah reagiert werden kann. Vorhersagen, wo sich wann wie viele Menschen per Taxi hinbewegen und ein Vergleich mit dem existierenden öffentlichen Nahverkehr könnten z.B. Hinweise darauf geben, wie es um die Auslastung etwa von Bussen bestellt ist und wo neue Linien Sinn machen würden – gegebenenfalls sogar über zusätzliche Busse, die flexibel einsetzbar in Bereitschaft gehalten werden. Geodaten bieten auch Potentiale bei der Standortanalyse, die oft noch aufwändig manuell durchgeführt wird. Sämtliche zur Verfügung stehenden Verkehrsdaten kombiniert könnten von Stadtplanern oder Unternehmen auch als Indiz für die Bewertung der wirtschaftlichen oder kulturellen Anziehungskraft eines Stadtteils benutzt werden[252].

Daten über die Straßennutzung können auch von Navigationsgeräte kommen. Hersteller räumen sich großzügig Rechte ein, Fahrtdaten zu speichern und an andere weiterzuverkaufen[253]. Der Versanddienstleister UPS verwendet in seinen Fahrzeugen ebenfalls GPS-Sender, um z.B. optimale Routen zu berechnen. Der Fahrtenvermittler Uber verkauft die gesamten Daten jedes Benutzers u.a. an eine große Hotelkette, die dadurch zielgerichtet werben kann. Über die GPS-Daten von vielfrequentierten Gebieten, die also eine gewisse Repräsentativität aufweisen, kann Uber z.B. eine Liste der beliebtesten Restaurants aufstellen und verkaufen[254]. Anbieter von Kreditkarten verkaufen wiederum die Information, wer wann und wo was gekauft hat an Anbieter im Bereich E-Commerce[255]. Die Stadt München könnte theoretisch die Daten des kürzlich eingeführten

[251] Ferreira et al., IEEE VCG 2013, 2149 (3).
[252] vgl. Tasse/Hong, 2014.
[253] DANA 2011, 86.
[254] Hirson, Forbes, 23. März 2015.
[255] Edwards, Business Insider, 16. April 2013.

Fahrradverleihsystems der Münchner Verkehrsgesellschaft benutzen[256], um dringend benötigte oder überlastete Radwege zu erkennen – ohne auch nur einen einzigen Fragebogen zu dieser Thematik gestalten und verschicken zu müssen, wie es früher der Fall gewesen wäre.

Ungeplante, natürlich entstehende Trampelpfade sind häufig effizienter, als vom Menschen geplante, die durch Wünsche nach Ästhetik oder Voreingenommenheit z.B. Kreuzungen gegenüber Kreisverkehren oder Weggabelungen bevorzugen[257]. Mit Hilfe dieses Wissens kann z.B. durch GPS-Daten die Wegeführung Fußgängerzonen, Gebäuden oder Veranstaltungsarealen für mehr Komfort oder den Katastrophenfall mit weniger Aufwand besser, da organischer gestaltet werden[258]. Allgemein könnte über GPS-Sensoren auch eine sich anbahnende Katastrophe noch in den Anfängen in Echtzeit erkannt und verhindert werden[259]. Dies gilt aber nicht nur für Fußgänger, die Wegeführung jedes Verkehrsmittels könnte grundsätzlich davon profitieren.

4.2 Risiken

Viele Risiken im Zusammenhang mit der Erfassung und Auswertung von Daten zur Verhaltensvorhersage finden ihre Berücksichtigung im Recht.

In der rechtswissenschaftlichen Diskussion werden sie unter den Begriffen Profiling[260] und Scoring[261] behandelt. Der Begriff Scoring steht dabei grundsätzlich im Zusammenhang mit Prognosen[262], meist zur Bonität, d.h. bezüglich der Wahrscheinlichkeit der Zurückzahlung eines Kredites[263]. Als Profil wird dabei zum einen die anonyme Personenkategorie[264], sowie die personenbezogene

---

[256] vgl. Siddle, 10. April 2014.
[257] Helbing, SZ Magazin 2010.
[258] Helbing, SZ Magazin 2010; vgl. Rauner/Stockrahm, Zeit Online, 25. Juni 2012.
[259] vgl. Rauner/Stockrahm, Zeit Online, 25. Juni 2012.
[260] Härting, CR 2014, 528 (529); vgl. Weichert, ZD 2013, 251 (255); vgl. Hildebrandt/Gutwirth, 2008, S. 19; vgl. Möncke, 1999, S. 40.
[261] vgl. ULD/GP Forschungsgruppe, 2014, S. 19ff; vgl. Deutscher Bundestag, Entwurf eines Gesetzes zur Änderung des Bundesdatenschutzgesetzes. Drucksache 16/10529, 2008, S. 9.
[262] vgl. Türpe et al., DuD 2014, 31 (33).
[263] vgl. Deutscher Bundestag, Entwurf eines Gesetzes zur Änderung des Bundesdatenschutzgesetzes. Drucksache 16/10529, 2008, S. 9; vgl. Möller/Florax, MMR 2002, 806 (806).
[264] vgl. Hildebrandt, DuD 2006, 548 (548).

Sammlung von Merkmalen verstanden[265]. Das Verfolgen des Verhaltens von Nutzern bezüglich ihrer Internetnutzung[266] oder Aufenthaltsortes[267] wird unter dem Begriff Tracking diskutiert. Tracking ist nicht zu verwechseln mit dem Begriff Tracker, der zu Zwecken der Deanonymisierung innerhalb von statistischen Datenbanken dient[268]. Die Versuche, im kommerziellen Bereich möglichst individuell passende Werbung, Produkte und Services anzubieten, werden dabei auch unter dem Begriff Personalisierung bzw. Personalization[269] (bzw. im Englischen seltener: Personalizing[270]) diskutiert.

Die Ursachen von Risiken von Big Data-Vorhersagen gruppenspezifischen Verhaltens sind technischer, wirtschaftlicher und grundlegend menschlicher Natur. Dabei ist auch eine wichtige Frage, inwiefern Risiken durch subjektive Wertungen oder durch besonders hohe Genauigkeit der Vorhersage entstehen[271].

Eine mögliche Gliederung der Risiken von Profiling orientiert sich an zwei grundsätzlichen Schritten der Datenverarbeitung: der Erstellung einer Datenbasis und Auswertung dieser Daten durch Algorithmen[272]. Dieser Ansicht wird modifizierend gefolgt, indem statt Daten- zwei Schritte der Informationsverarbeitung unterschieden werden: Erstens die Risiken der Big-Data-Gewinnung von der Vorhersagbarkeit und zweitens die Risiken der wertenden Anwendung von Vorhersagbarkeit. Beide Gebiete werden im Allgemeinen, aber auch für Geodaten im Speziellen betrachtet.

Im ersten Teil geht es um die Vereinbarkeit mit deutschem Datenschutz mit einzelnen Verweisen auf Entwicklungen des US- und EU-Rechts. Im zweiten Teil wird die Frage bearbeitet, ob auch rein anonymisierte Datenverarbeitung Risiken haben kann. Dazu gehören auch unbewusste, der Intuition widersprechende Fehlentscheidungen und Fehlentwicklungen aufgrund von bewusstem wirtschaften, wie etwa Fragen von Diskriminierung, deren mögliche wirtschaftliche

---

[265] vgl. Türpe et al., DuD 2014, 31 (33); vgl. Möncke, 1999, S. 40.
[266] Conrad/Hausen, in: Auer-Reinsdorff/Conrad, § 36 Rn 52ff; vgl. Weichert, ZD 2013, S. 255.
[267] vgl. Weichert, ZD 2013, 251 (255).
[268] Denning/Denning, ACM TODS 1979, 76; vgl. Möncke, 1999, S. 46.
[269] Cranor, 2003.
[270] vgl. Weichert, ZD 2013, 251 (255).
[271] vgl. Türpe et al., DuD 2014, 31 (34).
[272] vgl. Härting, CR 2014, 528 (529ff); vgl. Federal Trade Commission, 2016, S. 4f.

und gesellschaftliche Auswirkungen und Argumente für die Ausgestaltung regulatorischer Reaktionen.

4.2.1 Big-Data-Gewinnung der Vorhersagbarkeit

Abbildung 3: Arrivals[273]

Die Gewinnung der Vorhersagbarkeit betrifft direkt die Zielrichtung des Datenschutzes: die informationelle Selbstbestimmung, die schon 1971 propagiert wurde[274]. Es geht um das „Recht des Einzelnen, selbst über die Preisgabe und Verwendung seiner Daten zu bestimmen“[275]. Wie das Bundesverfassungsgericht im Volkszählungsurteil 1983 ausgeführt hat, geht es um die Entscheidungsfreiheit von Individuen, damit diese sich frei in ihrem Verhalten entfalten können, als Voraussetzung für die Demokratie[276]. Dazu gehören insbesondere die Prinzipien der Zweckbindung und Datensparsamkeit[277]. Personenbezogene Daten dürfen nur zu einem präzise im Voraus klar bestimmten und eingegrenzten Zweck verwendet werden. Es dürfen ausschließlich die minimal zur Zweckerfüllung notwendigen Daten gespeichert werden, die Sammlung „auf Vorrat“ ist ausdrücklich nicht erlaubt. Es geht um ein rechtliches Gegengewicht[278] aus Sorge vor der Steuerbarkeit von Individuen durch fortschrittlichere Verarbeitung von Daten[279]. Das Gericht befürchtet „eine umfassende Registrierung und Katalogisierung der Persönlichkeit durch die Zusammenführung einzelner Lebens- und

---

[273] Gregorius, Wikimedia Commons, 16. Juni 2011.
[274] vgl. Steinmüller et al., 1971, S. 39ff.
[275] Simitis, in: Simitis, § 1 Rn 25.
[276] vgl. BVerfG, *Volkszählungsurteil*, NJW 1984, 419 (422).
[277] vgl. BVerfG, *Volkszählungsurteil*, NJW 1984, 419 (422).
[278] vgl. Simitis, in: Simitis, § 1 Rn 27.
[279] vgl. Simitis, in: Simitis, § 1 Rn 26.

Personaldaten zur Erstellung von Persönlichkeitsprofilen der Bürger“ bis hin zum „Totalabbild“[280], ohne Kontrollmöglichkeit des Einzelnen. Der dadurch erzeugte psychische Druck wirkt schließlich freiheitshemmend auf sein Verhalten ein[281] - es genügt schon die Angst vor „Verdatung“[282].

Die Behandlung von Daten ist nach ihrer Art zu unterscheiden: personenbezogene und nicht personenbezogene, d.h. anonymisierte Daten[283]. Nur in ersterem Fall ist das Datenschutzrecht anwendbar[284]. In der Literatur wird die Frage der Analyse von anonymisierten Datensätzen unter dem Stichwort Privacy-Preserving Data Mining diskutiert[285].

Der Charme von Big Data liegt freilich darin, dass man gar nicht mehr sagen kann, für wozu welche Daten (z.B. von Sensoren) alles gut sein könnten, sondern dass es ganz im Gegenteil eher darum geht, durch möglichst viele Daten genaue und künstlerisch-kreative[286] Analysen durchzuführen. Die Wiederverwendung (reuse of data)[287] und Kombination von Daten ist eher als Regelfall, denn als Ausnahme anzusehen[288]. Dies widerspricht dem Gebot der Zweckbindung und erst recht dem der Datensparsamkeit[289].

Abhilfe kann grundsätzlich nur noch durch zwei Möglichkeiten geschaffen werden. Erstens ist eine wirksame Einwilligung der betroffenen Person denkbar, zweitens die Anonymisierung.

Die grundsätzliche Frage, ob anonymisierte Datenverarbeitung ohne Einwilligung zulässig sein sollte, hängt mit der Vorstellung zusammen, inwiefern es ein „Dateneigentum“ geben sollte, durch das andere von der Nutzung von Informationen ausgeschlossen werden können[290]. Gegen ein Totalverbot spricht zum einen die zutiefst menschliche Eigenschaft der Neugier und sozialen

---

[280] vgl. BVerfG, *Volkszählungsurteil*, NJW 1984, 419 (424).
[281] vgl. BVerfG, *Volkszählungsurteil*, NJW 1984, 419 (422).
[282] vgl. Steinmüller, DuD 1984, 91 (94).
[283] vgl. BVerfG, *Volkszählungsurteil*, NJW 1984, 419 (422).
[284] Bergt, ZD 2015, 365 (365); s.a. Schefzig, DSRITB 2014, 103 (104) mwN.
[285] vgl. Agrawal/Srikant, 2000; vgl. BITKOM, 2012, S. 45.
[286] vgl. Teradata Perspectives, Forbes, 30. Januar 2015.
[287] vgl. Mayer-Schönberger/Cukier, 2013, S. 104ff.
[288] vgl. Mayer-Schönberger/Cukier, 2013, S. 107–109; vgl. Katko/Babaei-Beigi, MMR 2014, 360 (362); vgl. Schefzig, DSRITB 2014, 103 (110).
[289] vgl. Roßnagel, ZD 2013, 562 (564); vgl. Baeriswyl, digma 2013, (14).
[290] vgl. Dorner, CR 2014, 617; vgl. Brink, PinG 2014, 15.

Eingebundenheit eines jeden[291], insbesondere aber eine rechtsökonomische Betrachtung, wonach auch eine Abwägung zugunsten der Erhöhung des Gemeinwohls zulässig sein darf[292]. Dennoch wird zurecht seit Jahren unter dem Namen „Privacy by Design" eine Geisteshaltung verfochten, welche die Privatsphäre umfassend und grundsätzlich proaktiv, standardmäßig und sichtbar in Lösungen eingebaut geschützt möchte[293]. Nachdrücklich wird vertreten, dass es sich nicht um ein Nullsummenspiel[294] handelt, sondern beide Seiten profitieren können[295].

Ausnahmsweise kann die Verarbeitung von Daten außerdem zur Erfüllung eines Vertrages notwendig sein, was aber eng auszulegen ist[296]. Beim Kreditscoring muss die Bank im Rahmen der Interessenabwägung angeben, dass die Datenübermittlung für die Bonitätsprüfung z.B. durch die Schufa dazu dient, eigene Geschäftsrisiken abzuwenden[297].

Die Wirksamkeit einer Einwilligung wird sehr oft daran scheitern, dass nicht klar ist, von wem, wie und wozu die Daten alles verwendet werden könnten und welche Missbrauchsrisiken entstehen[298]. Zudem wird der Umfang der eingeräumten und Erlaubnis, die oftmals sehr lang und unverständlich verfasst ist, einem schlichten einfachen Klick nicht gerecht[299]. Das Problem der Freiwilligkeit der Einwilligung stellt sich zudem, wenn man aufgrund der Marktmacht eines Anbieters im Grunde genommen eigentlich keine andere Wahl hat, als einzuwilligen[300]. Auf die Spitze getrieben wird dies mit Googles Prinzip, die Verwendung der eigenen Daten jegliche Dienste unbeschränkt akzeptieren zu müssen; dies stellt eine nach deutschem Recht stets unwirksame Blankoeinwilligung dar[301].

Das Ganze soll aber nicht bedeuten, dass Einwilligungen heutzutage grundsätzlich unrealistisch sind. Es kommt eben auf die Frage an, ob man den Zweck, die

---

[291] vgl. Giesen, PinG 2013, 62.
[292] vgl. CR 2014, 617 (625f).
[293] vgl. Cavoukian, 2009.
[294] vgl. Baumol/Blinder, 2011, S. 226f.
[295] vgl. Cavoukian, 2009; vgl. Cavoukian, iapp.org Privacy Perspectives, 18. April 2014.
[296] vgl. Simitis, in: Simitis, § 28 Rn 98.
[297] vgl. Kamlah, MMR 1999, 395 (398).
[298] vgl. Becker/Schwab, ZD 2015, 151 (153); vgl. Baeriswyl, digma 2013, (16).
[299] vgl. Härting, CR 2014, 528 (533).
[300] vgl. Roßnagel, ZD 2013, 562 (562); vgl. Weichert, ZD 2013, 251 (256).
[301] vgl. Becker/Becker, MMR 2012, 351 (354).

Verwendung und die Risiken eindeutig eingrenzen kann, oder nicht und wie formell und verständlich auch der Akt der Einwilligung ist. was durchaus als Herausforderung bezeichnet werden kann[302]. Big Data kann aber auch umgekehrt helfen, Überzeugungsarbeit zu leisten, dass auch der Einwilligende z.B. durch genauer an ihn angepasste oder kostengünstigere Services profitiert[303].

Insgesamt sind wirksame Einwilligungen zur Datenerhebung und -Verarbeitung nur sehr eingeschränkt einholbar, wenn es um die Verarbeitung für Zwecke der Big-Data-Vorhersagbarkeit gruppenspezifischen Verhaltens geht. Oft wird es in einem ersten Schritt wesentlich einfacher sein, auf die Anonymisierung als Mittel zu setzen[304], um erst im Schritt der Anwendung der Erkenntnisse auf den Einzelnen eine spezifische Einwilligung zu setzen. Eine Anonymisierte Speicherung von Daten hat im Übrigen auch den Vorteil, weniger Löschzwängen durch das ab 2018 geltende Recht auf Vergessenwerden ausgesetzt zu sein[305].

Es herrscht jedoch Uneinigkeit, welche Anforderungen an die Anonymisierung bzw. an die Resistenz gegenüber der Aufhebung der Anonymität zu stellen sind. Dazu steht der potentielle Analysewert der Daten im Spannungsverhältnis[306]. Die aktuell dem EuGH vorliegende Frage des Personenbezugs von IP-Adressen[307], bei der keine Anonymisierung, etwa durch Streichen der letzten 8 Stellen vollzogen wurde[308], ist dafür ein konkretes Beispiel im Zusammenhang mit Geodaten. Die entscheidende Frage ist die der Bestimmbarkeit einer Person im Sinne von § 3 Abs. 1 BDSG[309]. Im Beispiel des Merkmals IP-Adresse kann der Standort durch Wegfall des letzten Oktetts nur noch auf mehrere Kilometer genau bestimmt werden[310]. Dies schränkt Analysemöglichkeiten offensichtlich ein, andererseits ist es der Privatsphäre von Konsumenten zuträglich. Zur Bestimmbarkeit im Allgemeinen existieren zwei Theorien mit diversen Abstufungen[311], die sich um

---

[302] vgl. BITKOM, 2012, S. 43.
[303] vgl. BITKOM, 2012, S. 44.
[304] vgl. Brisch/Pieper, CR 2015, 724 (727); vgl. Härting, NJW 2013, 2065 (2066).
[305] vgl. Malle et al., 2016.
[306] vgl. Li/Li, 2009.
[307] BGH, *Speicherung von IP--Adressen durch die Bundesrepublik*, MMR 2015, 131.
[308] vgl. Schweda, MMR-Aktuell 2011.
[309] vgl. Brisch/Pieper, CR 2015, 724 (725); vgl. Bergt, ZD 2015, 365 (365).
[310] vgl. Arning/Moos, ZD 2014, 126 (131).
[311] vgl. Bergt, ZD 2015, 365 (365ff).

den Aufwand der Reidentifizierung bzw. Deanonymisierung drehen[312]. Um Schutzlücken zu schließen[313], genügt gemäß der objektiven Theorie schon die rein hypothetische Möglichkeit, den Personenbezug herzustellen. Der Aufwand spielt keine Rolle. Demnach zählt nur die tatsächliche Unmöglichkeit der Bestimmung einer Einzelperson, um den Personenbezug auszuschließen.

Nach der relativen Theorie ist nur verhältnismäßiger Aufwand zu berücksichtigen[314]. Man nennt dies auch faktische Anonymität[315]. Argument ist ein Erwägungsgrund der europäischen Richtlinie 95/46/EG, die mittelbar auch in Deutschland Anwendung findet, die enthält eine Formulierung bezüglich „vernünftigerweise einsetzbaren Mittel“ enthält[316]. Dies entspricht der Idee von der „Sicherheit durch Unklarheit“ (Security through Obscurity), die keine Garantie für Sicherheit bietet, sondern lediglich den Aufwand in die Höhe treibt[317].

Problematisch ist insbesondere, dass das, was als verhältnis- und was unverhältnismäßig bzw. als anonym ist anzusehen ist, sich mit dem technologischen Fortschritt verändert. Einige Autoren der Rechtswissenschaften gehen davon aus, dass es anonyme Daten heutzutage nicht mehr gebe, auch da heutzutage große Mengen von Daten leicht kombinierbar seien[318]. Damit werden sozusagen Mosaiksteine zusammengesetzt (Mosaicking) – was die eine Quelle an Informationen streicht, wird von anderswo legal beschafft[319]. Als Indiz wird angeführt, dass schon wenige und vergleichsweise dezente Merkmale genügen, um einen Menschen eindeutig identifizierbar zu machen. So genügen die drei Attribute Geschlecht, Postleitzahl und Geburtsdatum, um mehr als drei Viertel aller US-Amerikaner eindeutig zu unterscheiden[320]. In Bezug auf gesammelte Geodaten von Smartphones genügen vier Orte, um über 95 % der Personen zu identifizieren und zwei Orte genügen immer noch, um eine Rate von über 50 % zu erreichen[321]. Verschlüsselung durch simple Pseudonymisierung[322] wie etwa die

---

[312] vgl. Boehme-Neßler, DuD 2016, 419 (420).
[313] vgl. Schefzig, DSRITB 2014, 103 (106).
[314] vgl. Schefzig, DSRITB 2014, 103 (106); vgl. Bergt, ZD 2015, 365 (365f).
[315] Härting, NJW 2013, 2065 (2065).
[316] vgl. Schefzig, DSRITB 2014, 103 (107).
[317] vgl. Hartzog/Selinger, Stanford Law Rev. 2013, (83ff).
[318] vgl. Boehme-Neßler, DuD 2016, 419 (422); Ohm, 2009.
[319] vgl. Leetaru, Forbes.
[320] Sweeney, 2000, S. 2.
[321] Montjoye, de et al., Sci. Rep. 2013.
[322] vgl. Kroschwald, ZD 2014, 75 (80); vgl. Marnau, DuD 2016, 428 (431ff).

schlichte Verwendung von Decknamen[323] oder Kennnummern[324], wird daher heutzutage oft kaum mehr ausreichend sein.

Die oben erwähnten Angaben über PLZ, Geschlecht und Geburtsdatum wurden Mitte der 90er dazu benutzt, veröffentlichte pseudonymisierte Gesundheitsdaten zum damaligen Gouverneurs von Massachusetts zu verbinden, der deren Veröffentlichung zuvor als sicher bezeichnet hatte[325]. Netflix stoppte seinen zweiten Wettbewerb mit veröffentlichten pseudonymisierten Daten seiner Benutzer zur Verbesserung von Algorithmen seines Systems zur Filmempfehlung, nachdem es gelang, Personen mit ihren sexuellen Vorlieben zu identifizieren. Dazu wurden die Filmbewertungen von Netflix mit den öffentlich zugänglichen Benutzernamen und Bewertungen des Netzwerks für Filme namens IMDB verglichen, womit u.a. schon aufgrund einer Übereinstimmung von nur zwei Filmen mit ungefährem Datum (+- 3 Tage) über zwei Drittel der Benutzer identifiziert werden konnten[326]. Pseudonymisierte Suchhistorien von AOL über mehr als einer halben Million US-Amerikaner führten zu einem ähnlichen Desaster[327]. Die oben erwähnten Taxidaten von New York City waren zum großen Teil nur durch sehr oberflächliche Verschlüsselung pseudonymisiert und ermöglichten dadurch Reidentifizierung[328], was ebenfalls zum einen großen Imageschaden für die Regierung der Stadt mit sich brachte. Zum anderen wurde noch viel Erschreckenderes möglich: über die Kombination mit öffentlich zugänglichen Fotodaten anderer Herkunft die Feststellung, welche prominenten Taxikunden sich wie durch die Stadt bewegten und wieviel an Trinkgeld sie gaben.

Tatsächlich gibt es auch unter Informatikern Zweifel, wie sicher Verfahren zur Anonymisierung gegen die Angriffe mittels Hintergrundwissen sind[329]. Damit ist das oben erwähnte Verbinden von Datenquellen gemeint. Dennoch sind die genannten Beispiele kein Beweis für die Unmöglichkeit von Anonymisierung. In allen Fällen wurden meist offensichtliche Fehler gemacht, die die Reidentifizierung ermöglichten: etwa die Veröffentlichung von zu vielen Details

---

[323] vgl. Kühling/Klar, NJW 2013, 3611 (3613).
[324] vgl. Härting, NJW 2013, 2065 (2066).
[325] vgl. Ohm, 2009, S. 18; vgl. Sweeney, 2000.
[326] vgl. Narayanan/Shmatikov, 2008, S. 11.
[327] vgl. Barbaro/Zeller, The New York Times, 9. August 2006.
[328] vgl. Pandurangan, 21. Juni 2014.
[329] vgl. Pinto, 2012, S. 214; vgl. Fung et al., CSUR 2010, 1 (46).

und damit verbundene naive Pseudonymisierung, obwohl es um höchstsensible Daten ging[330]. Vielfach war auch keine Sensibilisierung dafür zu erkennen, dass auch Metadaten, also Daten, aus denen man über andere Daten schließen kann, für vielfältige Erkenntnisse gut sein können. Als Beispiel sei etwa die Supermarktkette Target genannt, die eine Kombination von Waren herausfand, deren Kauf mit hoher Wahrscheinlichkeit eine Schwangerschaft impliziert[331]. Übertragen auf das Taxi-Beispiel könnte man z.B. die Theorie aufstellen, über das regelmäßige fünfmalige Beten am gleichen Ort könnte man die Zugehörigkeit zum Islam herausfinden[332].

Es gibt ganz im Gegensatz zur erwähnten pessimistischen Überzeugung durchaus leistungsfähige Anonymisierungstechniken[333], wobei es stets um einen Kompromiss zwischen Sicherheitsgewinn und Informationsverlust geht[334]. Vielleicht werden sich Kritiker, die Anonymisierungsfähigkeit generell ausschließen, darauf einigen können, dass Anonymität nicht nur ein statisches Idealziel ist, dass man erreichen muss, sondern dass es um eine dynamische Betrachtung geht, den Prozess der Anonymisierung so gut es geht zu optimieren[335]. Anonymisierung ist jedenfalls mit Aufwand verbunden[336]. Das bedeutet z.B. sich damit auseinanderzusetzen, wie sensibel die gewinnbaren Informationen sind und inwiefern Datensätze daher verkürzt und Personen zu Gruppen zusammengefasst, d.h. aggregiert werden sollen. Nur dann ist eine Verarbeitung akzeptabel.

Unter dem Stichwort „Differential Privacy“ werden außerdem vielversprechende anonymisierte Verfahren behandelt, bei denen nicht die Daten, sondern der Zugang dazu anonymisiert wird[337]. Diese Verfahren machen im Besonderen ausdrücklich keine Einschränkungen gegenüber Hintergrundwissen, was ein großer Vorteil ist[338]. Die damit verbundenen Fragen sind sehr ähnlich zur Deanonymisierung in statistischen Datenbanken über sogenannte Tracker[339]. Dabei geht es um das Einfügen von Rauschen in die Suche, sodass das Risiko im

---

[330] vgl. Privacy Analytics, Privacy Analytics Blog, 29. August 2016.
[331] vgl. Hill, Forbes, 16. Februar 2012.
[332] vgl. Berlee, The Interdisciplinary Internet Institute, 21. Januar 2015.
[333] vgl. Information Commissioner's Office, 2012, S. 1.
[334] vgl. Qi/Zong, Procedia Env. Sci. 2012, 1341 (1345f).
[335] vgl. Roßnagel et al., 2001, S. 272.
[336] vgl. Schonschek, Experton Group ICT-News Dach, 2. Mai 2014.
[337] Dwork, 2006.
[338] vgl. Narayanan/Shmatikov, CACM 2010, 24 (26).
[339] vgl. Möncke, 1999, S. 45ff; vgl. Denning/Denning, ACM TODS 1979, 76.

Sinne einer Wahrscheinlichkeit auf einen Grenzwert ε hin minimiert werden kann[340]. Um Missbrauch vorzubeugen, können verdächtige Suchanfragen auch beobachtet und verboten werden[341]. Aber auch ohne derartige Techniken können wie gesagt auf vielfältige Weise die Daten selbst anonymisiert werden. Auch in Geodaten kann Rauschen eingefügt werden[342], was die Verletzung von Persönlichkeitsrechten im Taxi-Beispiel jedenfalls erheblich aufwändiger gemacht hätte.

Die Anwendung der Erkenntnisse auf den Einzelnen wiederum ist in jedem Fall personenbezogen[343]. Damit ist zwingend eine Einwilligung einzuholen. Es ist allerdings zu beachten, dass gemäß § 6a BDSG eine komplett automatisierte Einzelentscheidung unzulässig ist, d.h. ein Mensch muss die letzte Gewalt über Entscheidungen haben, die für den Betroffenen rechtliche Relevanz haben[344]. Gemäß §28b BDSG muss es sich beim Kreditscoring zudem um anerkannte mathematische Verfahren handeln. Dieser Konflikt drückt den Anspruch auf mathematische Objektivität aus, der jedoch mit der fehlenden Beweisbarkeit von Wahrscheinlichkeiten kollidiert[345].

Nachdem Daten gewonnen, zu Gruppen zusammengefasst, interpretiert und schließlich herausgefunden wurde, zu welcher Gruppe das betrachtete Individuum zu zählen ist, lässt sich dessen Verhalten mit einer gewissen Wahrscheinlichkeit vorhersagen – so jedenfalls die Idee. Im Übrigen sind die technischen Mittel in den meisten Fällen weit entfernt davon, exakte örtliche Vorhersagen bieten zu können[346]. Die bis hier erläuterten Datenschutzfragen erfassen allerdings nicht alle Risiken.

---

[340] vgl. Tockar, neustar. Research, 8. September 2014.
[341] vgl. Schonschek, Experton Group ICT-News Dach, 2. Mai 2014.
[342] vgl. Zandbergen, Advances in Medicine 2014, (11).
[343] vgl. Schefzig, DSRITB 2014, 103 (116).
[344] vgl. Wolber, CR 2003, 623 (625).
[345] vgl. Türpe et al., DuD 2014, 31 (34).
[346] vgl. Song et al., Science 2010, 1018.

4.2.2 Wertende Anwendung der Vorhersagbarkeit

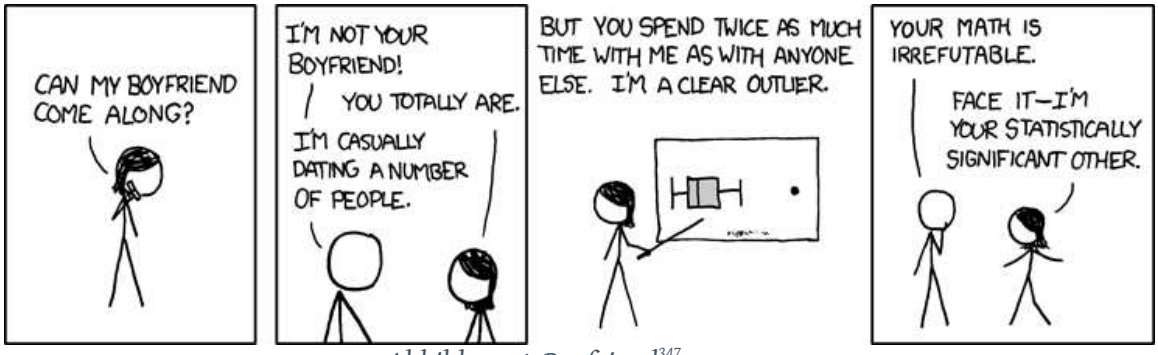

Abbildung 4: Boyfriend[347]

Nach der Informationsgewinnung geht es als nächstes zum einen um unbewusste wirtschaftliche/hoheitliche Fehlentscheidungen, zum anderen um mehr bewusstes Agieren, das dennoch mit Nachteilen für die Gesamtwirtschaft verbunden ist oder sogar mit unseren Vorstellungen von Fairness, Gerechtigkeit und sozialer demokratischer Gemeinschaft in Widerspruch steht.

4.2.2.1 Unbewusste Fehlentscheidungen

Wie in Kapitel 3 Technische Vorgehensweise gezeigt, sind mit dem Prozess der Gruppenspezifischen Verhaltensvorhersage verschiedene Unsicherheiten verbunden. Durch Menschen zu treffende, damit subjektive und für Fehler sowie Irrationalität anfällige Entscheidungen sind dem Prozess der Big-Data-Vorhersagbarkeit immanent. Das betrifft zunächst einmal den komplexen Prozess der Gewinnung der Vorhersagbarkeit, der faktisch nie zu perfekten Vorhersagen führen kann. Dies gilt umso mehr, wenn man nach der Repräsentativität von Prognosen fragt, denn Menschen sind schließlich Individuen mit einzigartigen Merkmalen. Auch wenn man davon ausgeht, dass das durchschnittliche Verhalten von Menschen leichter vorherzusagen ist, bleiben in jedem Fall Unsicherheiten dieses Versuchs der Abbildung von Realität, die nicht in aller Gänze nachvollzogen werden kann. Damit ist aber nicht gemeint, dass durch das mathematische Verfahren nicht mehr nachvollzogen werden kann, wie eine Vorhersage im konkreten Fall entstanden ist[348]. Tatsächlich wird es mehr auf den

[347] xkcd.
[348] vgl. Mayer-Schönberger/Cukier, 2013, S. 179.

Willen ankommen, Nachvollziehbarkeit zu ermöglichen[349] und Transparenz zu schaffen[350], indem man z.B. durch Visualisierung Licht in „Black Boxes“ wirft[351] und die Intentionen hinter der algorithmischen Ausgestaltung erklärt[352]. Manchmal müssen hier auch Gerichte einschreiten, wie das Schufa-Urteil zeigt, bei dem die Transparenz der Einflussfaktoren der Bonitätsbewertung, aber nicht des Algorithmus selbst erzwungen wurde[353].

Im Übrigen ist auch die Entscheidung, was man mit statistischen Ergebnissen macht, subjektiv und für menschliche Voreingenommenheit anfällig[354]. Fahrlässiges oder vorsätzliches Vertrauen in die Gültigkeit von Korrelationen und Vorhersagen kann Fehlentscheidungen zur Folge haben, egal ob der Akteur aus der Wirtschaft oder dem hoheitlichen Umfeld kommt. Die Frage und Problematik der Aussagekraft von Korrelationen ist allerdings nichts Neues, wurde doch schon vor Jahrzehnten prägnant formuliert, dass sämtliche statistischen Modelle falsch, aber manche nützlich seien[355]. Welcher Grad an Abstraktion und Vereinfachung sinnvoll sind, hängt vom jeweiligen analytischen Zweck ab[356]. Auch die Tatsache, dass Diskriminierung auch durch indirekte Merkmale möglich ist[357], wurde schon früh erkannt, wie das Zitat Anatole Frances zeigt, wonach vor dem Gesetz alle gleich sind, indem es Reichen wie Armen verboten ist, unter Brücken zu schlafen, zu betten und Brot zu stehlen[358]. Daneben ist aber auch die Datenqualität selbst oft schlecht, da Software und Hardware ebenfalls bedeutsame Fehler erzeugen können, die z.B. im medizinischen Bereich kritisch sind[359].

Ein Risiko besteht insbesondere in unbeabsichtigter Diskriminierung. Beispiel dafür ist die App für Smartphones der Stadt Boston von oben für die Verbesserung der Straßen, bei der man z.B. Schlaglöcher melden kann. Dort musste z.B.

---

[349] vgl. Goodman/Flaxman, arXiv:1606.08813 [cs, stat] 2016, (6ff).
[350] vgl. MIT Management Sloan School, Massachusetts Institute of Technology Sloan Experts Blog, 23. April 2015.
[351] vgl. Cortez/Embrechts, Inf. Sci. 2013, 1.
[352] vgl. Diakopoulos, S. 18ff.
[353] BGH, *Schufa*, NJW 2014, 1235.
[354] vgl. Mayer-Schönberger/Cukier, 2013, S. 63ff.
[355] George Edward Pelham Box, zitiert in: Velickovic, American Scientist 2015, 26 (26).
[356] vgl. Baumol/Blinder, 2011, S. 10.
[357] vgl. Federal Trade Commission, 2016, S. ii.
[358] „La loi, dans un grand souci d' égalité, interdit aux riches comme aux pauvres de coucher sous les ponts, de mendier dans les rues et de voler du pain.“, France, 1906, S. 118.
[359] vgl. Hoffman, 2014, S. 305f.

der Tatsache Rechnung getragen werden, dass alte und arme Leute seltener Smartphones mit sich herumtragen bzw. die App herunterladen und daher auch seltener Straßenschäden melden, wodurch die Gefahr bestand, gewisse Bürger und Stadtteile zu diskriminieren[360].

Empfehlungssysteme für Filme, Bücher und ähnliches können zu unerwünschten Vorschlägen führen, wenn man die Wahl des Filmes einem seltenen Gast zuhause überlässt oder etwas als Geschenk für jemand anderen kauft[361] - umso mehr, wenn bei gemeinschaftlich benutzten Geräte in den vorgeschlagenen Artikeln auf einmal auch die Geburtstagsüberraschung auftaucht[362]. Vielfach führen Versuche der Profilbildung zu Werbezwecken auch zu Abbildern, die aufgrund ihrer Fehlerhaftigkeit grotesk erscheinen, was in anderen Kontexten wie der Kreditvergabe wesentlich gravierendere Auswirkungen haben kann[363]. Suchmaschinenvorschläge werden zur sich selbst erfüllenden Prophezeiung, indem aufgrund von vielen Klicks vorgeschlagene Links wiederum angeklickt werden[364]. Dies kann rufschädigend wirken[365] und zudem das Vertrauen in die Validität der Suchergebnisse mindern. Ähnlichen Probleme hat die Idee, bestimmte Stadtteile als „Problembezirke“ unter besondere Beobachtung zu stellen, worauf hin mehr Straftaten erfasst werden, was schließlich die Einstufung als Problembezirk bestätigt[366].

Mehr als die Hälfte der Unternehmen in den USA benutzte 2013 die Kredithistorie eines Bewerbers bei Einstellungen, wobei es für viele auf den ersten Blick unverständlich erscheint, was dies mit der Fähigkeit für einen Job zu tun haben soll[367]. Seit 2007 hatten daher elf US-amerikanische Staaten derartige Praktiken verboten, um Chancengleichheit zu gewährleisten (Equal Opportunity[368])[369]. Tatsächlich wurden die Bewerbungsverfahren im Anschluss darauf aber diskriminierender, als je zuvor, da die einstellenden Personen unbewusst oder bewusst nun auf andere Merkmale auswichen, wie etwa die Hautfarbe oder das Alter;

---

[360] vgl. Executive Office of the President, 2014, S. 51f; vgl. Federal Trade Commission, 2016, S. 27.
[361] vgl. Cranor, 2003, S. 5.
[362] vgl. Stein, Time 2011.
[363] vgl. Stein, Time 2011; vgl. Federal Trade Commission, 2016, S. 30.
[364] vgl. Pasquale, 2011, S. 237; vgl. Federal Trade Commission, 2016, S. 28.
[365] vgl. Kastl, DSRITB 2014, 203 (210).
[366] vgl. Piltz, DSRITB 2014, 149 (150).
[367] vgl. The Editorial Board, The New York Times. The Opinion Pages. Editorial, 22. April 2013.
[368] vgl. Executive Office of the President, 2016, S. 9f.
[369] Guo, Washington Post Wonkblog, 23. März 2016.

paradoxerweise hatten also vor allem genau die Gruppe der Minderheiten und oberflächlich betrachtet Schwächeren, der man durch die Abschaffung der Einbeziehung der Kredithistorie helfen wollte, zuvor davon profitiert[370].

Besonders spektakulär ist auch das Scheitern des Projekts „Google Flu", das in geradezu hysterischer Weise mit Hoffnungen und Erwartungen erfüllt war[371]. So behauptete Google beim Start des Dienstes 2009, man könne genau vorhersagen, wo eine Grippewelle tagesgenau auftreten werde[372]. Tatsächlich spätestens 2013 klar, dass die Vorhersagen bis zur Unbrauchbarkeit hin ungenau entweder unterschätzt (Nichtvorhersage der Schweinegruppe 2009)[373] oder überschätzt wurden[374]. Dies ist wohl vor allem der Tatsache zu schulden, dass 80 bis 90% derjenigen, die Symptome von Grippe vorbringen, tatsächlich an einer anderen Erkrankung leiden[375].

Auch die immer zielgruppenspezifischere Werbung im Sinne von Targeted Advertising führt trotz immenser Datenauswertungen nicht unbedingt zu besseren Ergebnissen. Der Konsumgüterhersteller Procter & Gamble erzielte bessere Ergebnisse, als darauf verzichtet und schlicht für jeden einer gewissen Altersgruppe Werbung bei Facebook angezeigt wurde[376].

Auch für andere Unternehmen oder den Staat kann es mitunter besser sein, sich nicht auf nur einige wenige spezielle Gruppen, Regionen oder Services zu konzentrieren. Zwar kann Big Data mitunter innovative Hinweise geben, welche Investitionen überdurchschnittliche Rendite versprechen. Doch können bei großer Einseitigkeit Probleme sogenannter Risikokonzentration entstehen[377]. Auf Diversifikation zu achten kann hier helfen, gegebenenfalls von Regulierungsbehörden überwacht, welche wiederum durch neue Big Data-Werkzeuge unterstützt werden könnten[378]. Im Übrigen können selbst die besten Algorith-

---

[370] Guo, Washington Post Wonkblog, 23. März 2016.
[371] vgl. Mayer-Schönberger/Cukier, 2013, S. 55, 169.
[372] "...we can accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day.", Ginsberg et al., Nature 2009, 1012.
[373] Salzberg, Forbes, 23. März 2014.
[374] vgl. Butler, Nature 2013, 155.
[375] vgl. Salzberg, Forbes, 23. März 2014.
[376] vgl. Werner, Sueddeutsche Zeitung, 10. August 2016.
[377] vgl. Deutsche Bundesbank, 2006, S. 35.
[378] vgl. Gittleman, Reuters Blogs, 25. Februar 2014; vgl. Federal Trade Commission, 2016, S. 31f.

men nicht vor Fehlentscheidungen wie vor der globalen Finanzkrise helfen, wenn ihre Ergebnisse schlicht nicht mehr berücksichtigt werden[379].

Allerdings kann auch das vorbildlichste Bewusstsein über die Risiken und Schwächen von Big-Data-Analysen nicht alle Unzulänglichkeiten ausräumen, die die freie Marktwirtschaft zwangsläufig mit sich bringt; selbst wenn die Datenverarbeitung und Vorhersagbarkeit nicht auf eine individuelle Person bezogen sind. Damit sind unter Anderem Monopole angesprochen.

4.2.2.2 Gruppen, Demokratie und Gemeinwohl

Schon das Volkszählungsurteil[380] erwähnte die Gefahren für die demokratische Gesellschaft, wenn die informationelle Selbstbestimmung gestört ist[381]. Es geht also mittelbar um den Schutz von Gruppenverhalten[382].

Wenn die Marktmacht eines Unternehmens derart groß ist, dass von einer beherrschenden Stellung, in Bezug auf Google etwa von einem Informationsmonopol[383] gesprochen werden kann, dann besteht die Gefahr des Gebrauchs dieser Position zuungunsten der Mehrheit. Viele Unternehmen des Internets und damit auch von Big Data profitieren vom sogenannten Netzwerk-Effekt, der durch sich selbstverstärkende Rückkopplung die Bildung „natürlicher" Monopole begünstigt, indem ein Service umso besser und wertvoller wird, je mehr Menschen ihn benutzen[384]. Dazu kommen andere Effekte der Monopolbildung wie etwa Kostenvorteile durch schiere Größe (Economies of Scale) und versenkten Kosten[385] bzw. Austrittsbarrieren (Barriers to Exit) der Anwender in Form von Bequemlichkeit, die davon abhalten, den Anbieter zu wechseln (Vendor Lock-In[386]).

---

[379] „In the years leading up to the Great Recession that began in 2007, companies were giving mortgages to people with lower and lower credit scores, income, and wealth, and higher and higher debt levels. In other words, they either rewrote or ignored their previous mortgage approval algorithms. It wasn't that the old mortgage algorithms stopped working; it was that they stopped being used.", Brynjolfsson/McAfee, 2014, S. 17 (Fußnote).
[380] vgl. BVerfG, *Volkszählungsurteil*, NJW 1984, 419 (422).
[381] Simitis, in: Simitis, §1 Rn 31-41.
[382] vgl. Steinmüller, DuD 1984, 91 (94).
[383] vgl. Danckert/Mayer, MMR 2010, 219.
[384] vgl. Baumol/Blinder, 2011, S. 197.
[385] vgl. Baumol/Blinder, 2011, S. 195.
[386] „Vendor lock-in, or just lock-in, is the situation in which customers are dependent on a single manufacturer or supplier for some product (i.e., a good or service), or products, and cannot move to another vendor without substantial costs and/or inconvenience. This dependency is typically a result of standards that are controlled by the vendor (i.e., manufacturer or supplier).

Eine Austrittsbarriere kann aber auch die Angst vor Korrelation sein, wenn ein häufiger Wechsel von Telefonanbietern potentiell mit einem schlechteren Rating einhergeht. Monopolisten sitzen in einer für sie vorteilhaften Situation von Angebot und Nachfrage, da sie von Konkurrenz und Substituten weitgehend verschont bleiben[387]. Dadurch wird zu Lasten der Kunden der Preis festgesetzt, auch was die Zubilligung des Rechts auf informationelle Selbstbestimmung angeht (Nachfrage nach Datenschutz). Diese Preisbildung ist gesamtwirtschaftlich gesehen ineffizient[388] und steht unserem Verständnis fairer wirtschaftlicher Entfaltungsmöglichkeiten und mündiger Konsumenten entgegen.

Zu warnen ist aber vor zu strenger Regulierung, denn die Aussicht auf eine gewisse monopolistische Preisbildung ist ein bedeutsamer Anreiz für Innovation[389]. Diese könnte z.B. in unerreichten Vorhersagefähigkeiten für gruppenspezifisches Verhalten liegen. Hier kann jedoch wieder der Aspekt des sich selbst verstärkenden, natürlichen Monopols auftauchen, sodass gerade die Unternehmen, die sowieso schon aufgrund ihrer Marktmacht besonders viele Daten haben, auch noch bessere Vorhersagen machen können, was ihre Marktmacht weiter steigert[390]

Im Übrigen geht es um ein für Fragen des Verbraucherschutzes typisches Informationsgefälle[391]. Auch Probleme der Irreführung und Diskriminierung gehören dazu. Demokratie basiert aber gerade auf dem Schutz von Minderheiten[392]. Wenn die Zugehörigkeit zu einer Gruppe als solche diskriminierend wirkt, so wird sie zur Stigmatisierung[393], die die Menschenwürde und den Gleichbehandlungsgrundsatz berührt[394]. Dabei geht es um die Gefahr der Steuerung von Verhalten[395], vor denen andere Stimmen mit Begriffen wie „statistisches Kastenwesen“ oder „Ghettoisierung“ warnen[396]. Allein die Tatsache selbst anonymi-

---

It can grant the vendor some extent of monopoly power and can thus be much more profitable than would be the absence of such dependency.“, The Linux Information Project, 2006.

[387] vgl. Baumol/Blinder, 2011, S. 196ff.

[388] vgl. Baumol/Blinder, 2011, S. 199.

[389] vgl. Baumol/Blinder, 2011, S. 202.

[390] vgl. Mayer-Schönberger/Cukier, 2013, S. 115.

[391] vgl. Härting, CR 2014, 528 (529).

[392] vgl. Steinmüller, 1993, S. 656; vgl. Determann, NVwZ 2016, S. 565; vgl. Calliess, in: Calliess/Ruffert, Rn 22.

[393] vgl. Steinmüller, 1993, S. 656.

[394] vgl. Karg, ZD 2012, 255 (255).

[395] vgl. Karg, ZD 2012, 255 (255).

[396] vgl. Korczak/Wilken, 2008, S. 24.

sierter Registrierung und Katalogisierung kann psychischen Druck aufbauen, wie schon 1969 im Mikrozensus-Urteil erwähnt wurde[397]. Dabei kann die Angst vor Kontrolle zu einer wesentlich größeren Steuerung von Verhalten führen, als es die tatsächlichen Möglichkeiten der Kontrolle eigentlich erlauben[398].

In Bereichen, wo es um die Bonität, Gesundheit oder sonstige Leistungsfähigkeit von Personen geht, trifft ebenfalls häufig der Effekt der sogenannten asymmetrischen Information auf[399], hier in Form von „adverser Selektion" (Adverse Selection)[400] auf. Dabei ist genau ein gegenteiliges Informationsgefälle zwischen Verbraucher und Unternehmen im Vergleich zu oben gemeint. Die Personen, die z.B. einen Kredit wollen, wissen mehr über sich selbst, als das Gegenüber, und wollen diese Tatsache auch ausnutzen. Umgekehrt ist der Beweis der Leistungsfähigkeit (Fehlerfreiheit) mitunter schwierig[401]. So gehen Kreditgeber, Versicherungen etc. tendenziell vom Schlimmsten aus, ehrliche und leistungsfähige Personen werden diskriminiert. Dies schränkt den Wettbewerb ein und führt daher wieder zu gesamtgesellschaftlich unerwünschten Wohlstandsverlusten und Fehlanreizen. So werden Banken möglicherweise strenger auswählen und lieber einen guten Kunden weniger, als einen schlechten mehr akzeptieren wollen, d.h. lieber Fehler vom Typ II bzw. β statt I bzw. α. Diese Art von Ineffizienzen bezeichnet man auch als hohe Transaktionskosten, auf die man mit staatlichen Interventionen reagieren kann[402].

Neben monopolistischen Situationen kann aber auch das Gegenteil, nämlich starke Konkurrenz dazu führen, das Niveau des Datenschutzes abzusenken. Kein Unternehmen möchte nämlich von sich aus auf Möglichkeiten verzichten und sich gegenüber der Konkurrenz schlechter stellen[403]. Es handelt sich um ein Problem, das mit spieltheoretischen Mitteln[404] betrachtet werden kann. Die Durchsetzbarkeit eines umfassenden Datenschutzes durch kollektiven Rechtsschutz sollte indes durch die Schaffung einer Verbandsklagebefugnis durch die Änderung des Unterlassungsklagegesetzes vom Februar 2016 sicher verbessert

---

[397] vgl. BVerfG, *Mikrozensus-Entscheidung*, NJW 1969, 1707 (1707); vgl. Richter, DuD 2016, 581 (586).
[398] vgl. Coing/Horn, 1982, S. 511.
[399] vgl. Baumol/Blinder, 2011, S. 293.
[400] vgl. Ausubel, 1999, S. 1; vgl. Akerlof, Q. J. Econ. 1970, 488.
[401] vgl. Baumol/Blinder, 2011, S. 293.
[402] vgl. North, Econ. Inq. 1987, 419 (6).
[403] vgl. Brookman, Ars Technica, 18. Juli 2011.
[404] Baumol/Blinder, 2011, S. 223ff.

werden[405]. Häufig wird Diskriminierung allerdings versteckt erfolgen und man deshalb gar nicht auf die Idee kommen, sich dagegen zur Wehr zu setzen[406]. Es wäre zu überlegen, ob nicht Big-Data-Kommissionen, die Algorithmen geschäftsgeheimniswahrend[407] überprüfen, sinnvoll sein könnten, analog zu bereits existierenden Gendiagnostik-Kommissionen[408].

Auch auf höherer Ebene sind Manipulationen von monopolistischen Suchmaschinenbetreibern denkbar, indem etwa Wahlen entscheidend beeinflusst werden könnten[409]. Tatsächlich wurde ja die Wahl von Präsident Barack Obama 2012 angeblich durch die Unterstützung durch Big-Data-Analysen mitentschieden[410]. Es geht dabei auch um den Schutz der öffentlichen Meinungsbildung, deren Beeinträchtigung nicht immer offensichtlich ist[411]: „Denn auch wer nur hört, was er schon weiß, fühlt sich dadurch nicht unbedingt schlecht informiert.“.

[405] vgl. Spindler, ZD 2016, 114 (114ff); vgl. Zinke, DSRITB 2014, 161 (166ff); vgl. Robak, GRUR-Praxis 2016, 139.
[406] vgl. Federal Trade Commission, 2016, S. 14f.
[407] vgl. Richter, DuD 2016, 581 (92f).
[408] vgl. Piltz, DSRITB 2014, 149 (152ff).
[409] vgl. Epstein/Robertson, PNAS 2015, E4512.
[410] vgl. Lynch, Computerworld, 13. November 2012.
[411] vgl. Bunge, ZD-Aktuell 2015, 4635.

5 Fazit

Technische und rechtliche Begriffe, die im Zusammenhang mit der Big-Data-Vorhersagbarkeit gruppenspezifischen Verhaltens von Belang sind, wurden mithilfe eines auch historischen Vergleichs von Quellen unterschiedlicher fachlichen Herkunft verständlich gemacht. Besondere Erläuterung erfuhr die Relevanz verschiedener Entwicklungen der Technik für Big Data wie etwa zu den Fragen Verteilung, Skalierung und Datenvielfalt. Ein Überblick über die einzelnen Schritte der Big Data-Vorhersage mittels Knowledge Discovery in Databases wurde gegeben, auch im Vergleich zur klassischen Statistik. Dabei wurden Algorithmen, aber auch künstlerisch-visuelle Aufgaben besprochen. Überdies wurde die These widerlegt, Automatisierung sei mit absoluter Objektivität gleichzusetzen. Ganz im Gegenteil wurde auf zahlreiche Stellen hingewiesen, bei denen mit Unsicherheit umgegangen werden muss und menschliche und damit subjektive Entscheidungen notwendig sind. Danach wurden die mannigfaltigen Chancen der Analyse insbesondere von Geodaten aufgezeigt, auch zur anonymisierten Verhaltensvorhersage. Subjektivität kann allerdings zu Fehlern führen; diese Risiken wurden untergliedert und anhand von vielen Beispielen erläutert. Dabei wurde sich zunächst Risiken des Prozesses der Big-Data-Gewinnung von Vorhersagbarkeit auseinandergesetzt, die vor Allem den Schutz personenbezogener Daten betreffen. Dabei wurde insbesondere auf das Volkszählungsurteil Bezug genommen und die Problematik wirksamer Einwilligung und Anonymisierung veranschaulicht. Anschließend wurden die Risiken der wertenden Anwendung der Vorhersagbarkeit diskutiert. Zunächst wurden unbewusste Fehlentscheidungen, schließlich Gefahren intendierten Handelns betrachtet. Dabei wurde jeweils auf paradoxe Effekte hingewiesen und allgemein herausgearbeitet, wie ökonomische Prinzipien für ein tieferes Verständnis für auftretende Phänomene geben können. Es wurde auch von neuen Entwicklungen des Datenschutzrechts wie der Verbandsklage berichtet. Schließlich wurde erklärt, wie auch die scheinbar harmlose Verarbeitung nicht personenbezogener, d.h. anonymisierter Daten Gefahren für die Demokratie haben kann.

Trotz Schwierigkeiten lassen sich Big-Data-Analysen mit ihren Charakteristika definieren und damit begreifbar machen. Nichts ist grundlegend neu, dennoch werden vielfältige Entwicklungen aus Statistik und Informatik systematisch,

kreativ und in besonderer Konsequenz angewandt. Die Darstellung der Chancen und Risiken von Big Data kann von verschiedenartigen, lange erprobten Methoden und Erkenntnissen aus Statistik, Recht und Volkswirtschaft profitieren. Wenigstens ein grundlegendes Verständnis des technischen Ablaufs inklusive der Algorithmen ist dabei für eine seriöse und angemessene Bewertung der Vorhersagbarkeit gruppenspezifischen Verhaltens unabdingbar. Dabei ist vieles von der klassischen Statistik übertragbar. Die gegebene Subjektivität und Ungewissheit sind sich bei Entscheidungen immer bewusst zu halten; es handelt sich um ein Wechselspiel zwischen Vertrauen und Skepsis in neue Techniken. Die Gefahren einer anonymisierten Datenverarbeitung für Individuen, Gruppen und schließlich die Gesellschaft an sich können transparent gemacht und durch eine wirtschaftliche Perspektive zielgerichtet Anreize geschaffen werden, die dem Grundrecht auf freie Entfaltung der Persönlichkeit und dem Gemeinwohl insgesamt förderlich sind.

6 Quellenverzeichnis

Aggarwal, Charu, *Data Mining*, Springer International Publishing (New York, NY), 2015,

Agrawal, Rakesh / Srikant, Ramakrishnan, *Privacy-preserving Data Mining* in: *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, ACM (New York, NY), 2000, 439–450,

Akerlof, George Arthur, *The Market for „Lemons": Quality Uncertainty and the Market Mechanism*, The Quarterly Journal of Economics, 1970, 488–500,

Alghuraybi, Bayan / Marvaniya, Krishna / Xia, Guojun / Woo, Jongwook, *Analyze NYC Taxi Data Using Hive and Machine Learning*, International Journal of Database Theory and Application, 30. Juni 2016, 191–198,

Angles, Renzo / Gutierrez, Claudio, *Survey of Graph Database Models*, ACM Computing Surveys, Februar 2008, 1:1–1:39,

Arning, Marian / Moos, Flemming, *Location Based Advertising. Datenschutzkonforme Verwendung von Ortsdaten bei verhaltensbezogener Online-Werbung*, Zeitschrift für Datenschutz, 2014, 126–133,

Ashton, Kevin, *That „Internet of Things" Thing*, RFID Journal, 22. Juni 2009, verfügbar unter http://www.rfidjournal.com/articles/view?4986 (zugegriffen 20. September 2016),

Auer-Reinsdorff, Astrid / Conrad, Isabell / Baldus, Oliver / Bierekoven, Christiane / Bischof, Elke / Dressler, Maximilian / Eckhardt, Jens / Fischl, Thomas / Förster, Romy / Grapentin, Sabine / Hassemer, Ines M. / Hausen, Dominik / Hertneck, Danielle / Huppertz, Peter / Kast, Christian R. / Kociok, Carsten / Kremer, Sascha / Lapp, Thomas / Luckhaus, Ulrich / Marberth-Kubicki, Annette / Maties, Martin / Mayer, Georg S. / Müller, Wolfgang / Picot, Henriette / Pohle, Jan / Pruß, Michael / Redeker, Helmut / Roth-Neuschild, Birgit / Sarre, Frank / Schmidt, Markus / Schneider, Jochen / Schöttle, Hendrik / Schrader, Paul Tobias / Schuster, Fabian / Sobola, Sabine / Stadler, Andreas / Streitz, Siegfried / Strittmatter, Marc / Thalhofer, Thomas / Venetis, Frank / Widmer, Ursula / Wiesemann, Hans Peter / Witte, Andreas / Witzel, Michaela, *Handbuch IT- und Datenschutzrecht*, C. H. Beck (München), 2. Auflage 2015 (Zitiert: *Bearbeiter,* in: Auer-Reinsdorff/Conrad),

Ausubel, Lawrence, *Adverse selection in the credit card market*, University of Maryland (College Park, MD), 1999,

Author, Markus Ziener, *Google and the German Angst*, The Huffington Post, 17. Juli 2014, verfügbar unter http://www.huffingtonpost.com/markus-ziener/google-and-the-german-angst_b_5592889.html (zugegriffen 18. September 2016),

Baeriswyl, Bruno, *„Big Data" ohne Datenschutz-Leitplanken*, digma, 2013,

Ballsun-Stanton, Brian, *Asking About Data: Exploring Different Realities of Data via the Social Data Flow Network Methodology*, University of New South Wales, 2012,

verfügbar unter http://philpapers.org/rec/BALAAD-3 (zugegriffen 14. September 2016),

Banko, Michele / Brill, Eric, *Scaling to Very Very Large Corpora for Natural Language Disambiguation* in: *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics (Stroudsburg, PA), 2001, 26–33,

Barbaro, Michael / Zeller, Tom Jr, *A Face Is Exposed for AOL Searcher No. 4417749*, The New York Times, 9. August 2006, verfügbar unter http://www.nytimes.com/2006/08/09/technology/09aol.html?_r=0 (zugegriffen 16. September 2016),

Batty, Michael, *Big data, smart cities and city planning*, Dialogues in Human Geography, 1. November 2013, 274–279,

Bäumler, Roland / Gutsche, Jörg, *Auswirkungen von Scoring und risikoorientiertem Pricing für Privatkunden*, Verbraucher und Recht, 2008, 81–84,

Baumol, William Jack / Blinder, Alan Stuart, *Economics: Principles and Policy by William J. Baumol*, South-Western (Boston, MA), International edition of the 12. revised edition 2011,

Beaver, Doug / Kumar, Sanjeev / Li, Harry C. / Sobel, Jason / Vajgel, Peter, *Finding a Needle in Haystack: Facebook's Photo Storage* in: *Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation*, USENIX Association (Berkeley, CA), 2010, 47–60,

Becker, Eva-Maria / Schwab, David, *Big Data im Gesundheitswesen. Datenschutzrechtliche Zulässigkeit und Lösungsansätze*, Zeitschrift für Datenschutz, 2015, 151–155,

Becker, Maximilian / Becker, Felix, *Die neue Google-Datenschutzerklärung und das Nutzer-MetaprofilVereinbarkeit mit nationalen und gemeinschaftsrechtlichen Vorgaben*, Multimedia und Recht. Zeitschrift für Informations-, Telekommunikations- und Medienrecht, 2012, 351–355,

Bergt, Matthias, *Die Bestimmbarkeit als Grundproblem des DatenschutzrechtsÜberblick über den Theorienstreit und Lösungsvorschlag*, Zeitschrift für Datenschutz, 2015, 365–371,

Berlee, *Using NYC Taxi Data to identify Muslim taxi drivers*, The Interdisciplinary Internet Institute, 21. Januar 2015, verfügbar unter http://www.theiii.org/index.php/997/using-nyc-taxi-data-to-identify-muslim-taxi-drivers/ (zugegriffen 17. September 2016),

BGH, *28.1.2014 - VI ZR 156/13 - Umfang des Auskunftsanspruchs gegen die Schufa – Scorewerte*, NJW 2014, 1235,

BGH, *28.10.2014 - VI ZR 135/13 - Speicherung von IP-Adressen durch die Bundesrepublik*, MMR 2015, 131,

BITKOM, *Leitfaden Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte*, 2012, verfügbar unter https://www.bitkom.org/Bitkom/Publikationen/Leitfaden-Big-Data-im-Praxiseinsatz-Szenarien-Beispiele-Effekte.html (zugegriffen 15. August 2016),

BITKOM, *Big-Data-Technologien - Wissen für Entscheider*, 2014, verfügbar unter https://www.bitkom.org/Bitkom/Publikationen/Big-Data-Technologien-Wissen-fuer-Entscheider.html (zugegriffen 15. August 2016),

BITKOM, *Big Data und Geschäftsmodell - Innovationen in der Praxis: 40+ Beispiele*, 2015, verfügbar unter https://www.bitkom.org/Bitkom/Publikationen/Big-Data-und-Geschaeftsmodell-Innovationen-in-der-Praxis-40-Beispiele.html (zugegriffen 15. August 2016),

Bock, Hans-Hermann, *Origins and extensions of the k-means algorithm in cluster analysis*, Journ@l Electronique d'Histoire des Probabilités et de la Statistique, Dezember 2008,

Boehme-Neßler, Volker, *Das Ende der Anonymität: Wie Big Data das Datenschutzrecht verändert*, Datenschutz und Datensicherheit, Juli 2016, 419–423,

Bolthausen, Erwin / Wüthrich, Mario Valentin, *Bernoulli's Law Of Large Numbers*, ASTIN Bulletin: The Journal of the IAA, Mai 2013, 73–79,

Bosc, Patric / Prade, Henri *An Introduction to the Fuzzy Set and Possibility Theory-Based Treatment of Flexible Queries and Uncertain or Imprecise Databases* in: Motro, Amihai / Smets, Philippe (Hrsg.), *Uncertainty Management in Information Systems*, Springer US (Boston, MA), 1997, 285–324,

Brink, Stefan, *Kurzes Plädoyer für unser „Supergrundrecht" auf informationelle Selbstbestimmung*, Privacy in Germany, 2014, 15–17,

Brisch, Klaus / Pieper, Fritz, *Das Kriterium der „Bestimmbarkeit" bei Big Data-Analyseverfahren*, Computer und Recht, 2015, 724–729,

Brookman, Justin, *Why the US needs a data privacy law - and why it might finally get one*, Ars Technica, 18. Juli 2011, verfügbar unter http://arstechnica.com/tech-policy/news/2011/07/why-the-us-needs-a-data-privacy-lawand-why-it-might-actually-happen.ars (zugegriffen 18. September 2016),

Brynjolfsson, Erik / McAfee, Andrew, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, W. W. Norton & Company, 2014,

Bunge, Felix, *Über die kollektive Schutzrichtung des Rechts auf informationelle Selbstbestimmung*, Newsdienst ZD-Aktuell. Zeitschrift für Datenschutz, 2015, 4635,

Burd, Greg, *NoSQL*, ;login:, Oktober 2011,

Butler, Declan, *When Google got flu wrong*, Nature, 13. Februar 2013, 155–156,

BVerfG, *16. 7. 1969 - 1 BvL 19/63 - Zur Verfassungsmäßigkeit einer Repräsentativstatistik (Mikrozensus)*, NJW 1969, 1707,

BVerfG, *15-12-1983 - 1 BvR 209/83 u. a. - Verfassungsrechtliche Überprüfung des Volkszählungsgesetzes 1983*, NJW 1984, 419,

Calliess, *EU-Vertrag (Lissabon) Art. 2 [Die Werte der Union]*, 5. Auflage 2016 (Zitiert: *Bearbeiter,* in: Calliess/Ruffert),

Cattell, Rick, *Scalable SQL and NoSQL Data Stores*, ACM SIGMOD Record, Mai 2011, 12–27,

Cavoukian, Ann, *Privacy by Design. The 7 Foundational Principles*, Information and Privacy Commissioner of Ontario (Ontario, Canada), 2009, verfügbar unter http://www. privacybydesign. ca/index. php/about-pbd/7-foundamental-principles. (zugegriffen 25. Oktober 2014),

Cavoukian, Ann, *More Privacy Paternalism: "We Know What's Best for You"*, iapp.org Privacy Perspectives, 18. April 2014, verfügbar unter https://iapp.org/news/a/more-privacy-paternalism-we-know-whats-best-for-you/ (zugegriffen 18. September 2016),

Chandy, Mani K. / Etzion, Opher / Ammon, Rainer von *The Event Processing Manifesto* in: Chandy, K. Mani / Etzion, Opher / Ammon, Rainer von (Hrsg.), *Dagstuhl Seminar Proceedings. Event Processing*, Leibniz-Zentrum fuer Informatik (Schloss Dagstuhl, Germany), 2011,

Choudari, Akhil / Joshi, Sankalp / Bembalkar, Akshay / Marathe, Nainesh / Sankpal, L. J., *Book Tracking Application in Android for Library Using GPS*, International Journal of Innovative Research in Computer and Communication Engineering, März 2013, 30–34,

City of Boston, *Boston Taxi Data*, verfügbar unter https://data.cityofboston.gov/Transportation/Boston-Taxi-Data/ypqb-henq (zugegriffen 15. September 2016),

Coenen, Frans, *Data mining: past, present and future*, The Knowledge Engineering Review, März 2011, 25–29,

Coing, Helmut / Horn, Norbert, *Europäisches Rechtsdenken in Geschichte und Gegenwart: Festschrift für Helmut Coing zum 70. Geburtstag Band 2*, C. H. Beck (München), 1982,

Cortez, Paulo / Embrechts, Mark J., *Using Sensitivity Analysis and Visualization Techniques to Open Black Box Data Mining Models*, Information Sciences, März 2013, 1–17,

Crampton, Charlotte, *Germany: The Rise and Growth of App Usage*, comScore Insights, 8. Juli 2016, verfügbar unter https://www.comscore.com/Insights/Data-Mine/Germany-The-Rise-and-Growth-of-App-Usage (zugegriffen 20. September 2016),

Cranor, Lorrie Faith, *«I Didn't Buy it for Myself». Privacy and Ecommerce Personalization* in: *Proceedings of the 2003 ACM Workshop on Privacy in the Electronic Society*, ACM (New York, NY), 2003, 111–117,

Daimler AG, *Annual Report 2015*, 2016, verfügbar unter https://www.daimler.com/investors/reports/annual-reports/2015/ (zugegriffen 20. September 2016),

Danckert, Burkhard / Mayer, Frank Joachim, *Die vorherrschende Meinungsmacht von Google. Bedrohung durch einen Informationsmonopolisten?*, Multimedia und Recht. Zeitschrift für Informations-, Telekommunikations- und Medienrecht, 2010, 219–222,

Das, Raghu / Harrop, Peter, *RFID Forecasts, Players and Opportunities 2016-2026. The complete analysis of the global RFID industry*, IDTechEx, Incorporated, 2015, verfügbar unter http://www.idtechex.com/research/reports/rfid-forecasts-players-and-opportunities-2016-2026-000451.asp (zugegriffen 24. August 2016),

Davis, Ben, *TeradataVoice: What Is A Creative Data Scientist Worth?*, Forbes, 30. Juni 2015, verfügbar unter http://www.forbes.com/sites/teradata/2015/06/30/what-is-a-creative-data-scientist-worth/print/ (zugegriffen 17. September 2016),

Dean, Jeffrey / Ghemawat, Sanjay, *MapReduce: Simplified Data Processing on Large Clusters*, Communications of the ACM, Januar 2008, 107–113,

Denning, Dorothy E. / Denning, Peter J., *The Tracker: A Threat to Statistical Database Security*, ACM Transactions on Database Systems, März 1979, 76–96,

Determann, Lothar, *Datenschutz in den USA - Dichtung und Wahrheit*, Neue Zeitschrift für Verwaltungsrecht, 2016, 561–567,

Deutsche Bundesbank, *Concentration risk in credit portfolios*, 2006, verfügbar unter http://www.bundesbank.de/Redaktion/EN/Downloads/Publications/Monthly_Report_Articles/2006/2006_06_concentration_risk.pdf (zugegriffen 9. September 2016),

Deutsche Vereinigung für Datenschutz e.V. (DVD) (Hrsg.), *TomTom verkauft ano-nymisierte Bewegungs-profile an Polizei*, Datenschutznachrichten, 2011, 86,

Deutscher Bundestag, Entwurf eines Gesetzes zur Änderung des Bundesdatenschutzgesetzes. Drucksache 16/10529, 10. Oktober 2008,

Diakopoulos, Nicholas, *Algorithmic accountability: On the investigation of black boxes*, Knight Foundation and the Tow Center on Digital Journalism Columbia Journalism School, verfügbar unter http://towcenter.org/research/algorithmic-accountability-on-the-investigation-of-black-boxes-2 (zugegriffen 20. September 2016),

Dorner, Michael, *Big Data und „Dateneigentum“*, Computer und Recht, 2014, 617,

Dwork, Cynthia, *Differential Privacy* in: *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*, Springer (Berlin; Heidelberg), 2006, 1–12,

Edelstein, Herbert A., *Introduction to Data Mining and Knowledge Discovery*, Two Crows Corporation (Potomac, MD), 3rd Edition 1999,

Edwards, Jim, *Yes, Your Credit Card Company Is Selling Your Purchase Data To Online Advertisers*, Business Insider, 16. April 2013, verfügbar unter http://www.businessinsider.com/credit-cards-sell-purchase-data-to-advertisers-2013-4 (zugegriffen 15. September 2016),

Eifrem, Ian Robinson, Jim Webber, Emil, *Graph Databases*, O'Reilly Media (Beijing; Cambridge, MA; Farnham, UK; Köln; Sebastopol, CA; Tokyo), 2. Edition 2015,

Epstein, Robert / Robertson, Ronald E., *The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections*, Proceedings of the National Academy of Sciences, 18. August 2015, E4512–E4521,

Estivill-Castro, Vladimir, *Why so many clustering algorithms: a position paper*, ACM SIGKDD Explorations Newsletter, 1. Juni 2002, 65–75,

Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values*, Executive Office of the President, 2014,

Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values*, The White House (Washington, DC), 2014, verfügbar unter https://www.hsdl.org/?abstract&did=752636 (zugegriffen 17. August 2016),

Executive Office of the President, *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*, Executive Office of the President, 2016,

Facebook, *Second Quarter 2016 Results Conference Call*, Facebook, Incorporated, 2016, verfügbar unter https://investor.fb.com/investor-news/press-release-details/2016/Facebook-Reports-Second-Quarter-2016-Results/default.aspx (zugegriffen 20. September 2016),

Falk, Michael / Hain, Johannes / Marohn, Frank / Fischer, Hans / Michel, René, *Statistik in Theorie und Praxis*, Springer (Berlin; Heidelberg), 2014,

Fayyad, Usama / Piatetsky-Shapiro, Gregory / Smyth, Padhraic, *The KDD process for extracting useful knowledge from volumes of data*, Communications of the ACM, 1. November 1996, 27–34,

Federal Trade Commission, *Big Data. A Tool for Inclusion or Exclusion? Understanding the Issues*, Federal Trade Commission, 2016, verfügbar unter https://www.ftc.gov/reports/big-data-tool-inclusion-or-exclusion-understanding-issues-ftc-report (zugegriffen 20. September 2016),

Ferreira, Nivan / Poco, Jorge / Vo, Huy T. / Freire, Juliana / Silva, Cláudio T., *Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips*, IEEE Transactions on Visualization and Computer Graphics, Dezember 2013, 2149–2158,

Flouris, Ioannis / Giatrakos, Nikos / Deligiannakis, Antonios / Garofalakis, Minos / Kamp, Michael / Mock, Michael, *Issues in complex event processing: Status and prospects in the Big Data era*, Journal of Systems and Software, 16. Juni 2016,

Fraley, Chris / Raftery, Adrian Elmes, *How many clusters? Which clustering method? Answers via model-based cluster analysis*, Computer Journal, 1998,

France, Anatole, *Le lys rouge*, Calmann-Lévy (Paris), 1906,

Freiknecht, Jonas, *Big Data in der Praxis: Lösungen mit Hadoop, HBase und Hive. Daten speichern, aufbereiten, visualisieren*, Carl Hanser (München), 2014,

Fung, Benjamin C. M. / Wang, Ke / Chen, Rui / Yu, Philip S., *Privacy-preserving data publishing: A survey of recent developments*, ACM Computing Surveys, 1. Juni 2010, 1–53,

Gantz, John / Reinsel, David, *The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East*, IDC Analyze the Future, 2012, verfügbar unter https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf (zugegriffen 20. September 2016),

Gewirtz, David, *We've come a long way, baby: the iPhone 4 vs. the IBM PC (Infographic)*, ZDNet, 13. April 2012, verfügbar unter http://www.zdnet.com/article/weve-come-a-long-way-baby-the-iphone-4-vs-the-ibm-pc-infographic/ (zugegriffen 15. September 2016),

Giesen, Thomas, *Kurzes Plädoyer gegen unser Totalverbot: Deine Daten gehören Dir keineswegs!*, Privacy in Germany, 2013, 62–64,

Ginsberg, Jeremy / Mohebbi, Matthew H. / Patel, Rajan S. / Brammer, Lynnette / Smolinski, Mark S. / Brilliant, Larry, *Detecting influenza epidemics using search engine query data*, Nature, 19. Februar 2009, 1012–1014,

Gittleman, Stuart, *"Big data" tools will improve regulatory oversight, FINRA's di Florio says*, Reuters Blogs, 25. Februar 2014, verfügbar unter http://blogs.reuters.com/financial-regulatory-forum/2014/02/25/big-data-tools-will-improve-regulatory-oversight-finras-di-florio-says/ (zugegriffen 9. September 2016),

Goebel, Michael / Gruenwald, Le, *A survey of data mining and knowledge discovery software tools*, ACM SIGKDD Explorations Newsletter, 1. Juni 1999, 20–33,

Goodman, Bryce / Flaxman, Seth, *European Union regulations on algorithmic decision-making and a „right to explanation"*, arXiv:1606.08813 [cs, stat], 28. Juni 2016,

Google Incorporated, *Google Trends for „Big Data" 1/2004-9/2016*, 18. September 2016, verfügbar unter https://www.google.com/trends/explore?q=big%20data (zugegriffen 18. September 2016),

Gormley, Clinton / Tong, Zachary, *Elasticsearch. The Definitive Guide. A Distributed Real-time Search And Analytics Engine*, O'Reilly Media (Beijing; Cambridge, MA; Farnham, UK; Köln; Sebastopol, CA; Tokyo), 2015,

Gregorius, Thierry, *My first cartoon in ages. (Arrivals. „Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here")*, Wikimedia Commons, 16. Juni 2011, verfügbar unter https://commons.wikimedia.org/wiki/File:Big_data_cartoon_t_gregorius.jpg (zugegriffen 16. August 2016),

Guo, Jeff, *The law was supposed to reduce discrimination. But it made hiring more racially biased*, Washington Post Wonkblog, 23. März 2016, verfügbar unter https://www.washingtonpost.com/news/wonk/wp/2016/03/23/the-law-was-supposed-to-reduce-discrimination-but-it-made-hiring-more-racially-biased/ (zugegriffen 28. August 2016),

Halevy, Alon / Norvig, Peter / Pereira, Fernando, *The Unreasonable Effectiveness of Data*, IEEE Intelligent Systems, 2009, 8–12,

Hamel, Lutz, *Database Queries, Data Mining, and OLAP*, Idea Group Publishing (Amsterdam; Boston, MA), 2005,

Han, Jiawei / Kamber, Micheline / Pei, Jian, *Data Mining: Concepts and Techniques*, Morgan Kaufmann (Haryana, IN; Burlington, MA), 3. Edition 2011,

Härting, Niko, *Anonymität und Pseudonymität im Datenschutzrecht*, Neue Juristische Wochenschrift, 2013, 2065–2071,

Härting, Niko, *Profiling*, Computer und Recht, 2014, 528,

Hartzog, Woodrow / Selinger, Evan, *Big Data in Small Hands*, Stanford Law Review, September 2013,

Helbing, Dirk, *Draußen eigene Wege gehen - Natur*, Sueddeutsche Zeitung Magazin, 2010,

Helbing, Dirk, *Thinking Ahead - Essays on Big Data, Digital Revolution, and Participatory Market Society*, Springer (Heidelberg; New York etc), 2015,

Hilbert, Martin / López, Priscila, *The World's Technological Capacity to Store, Communicate, and Compute Information*, Science, 1. April 2011, 60–65,

Hildebrandt, Mireille, *Profiling: From data to knowledge*, Datenschutz und Datensicherheit, September 2006, 548–552,

Hildebrandt, Mireille / Gutwirth, Serge (Hrsg.), *Profiling the European Citizen*, Springer Netherlands (Dordrecht, NL), 2008,

Hill, Kashmir, *How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did*, Forbes, 16. Februar 2012, verfügbar unter http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/ (zugegriffen 17. September 2016),

Hirson, Ron, *Uber: The Big Data Company*, Forbes, 23. März 2015, verfügbar unter http://www.forbes.com/sites/ronhirson/2015/03/23/uber-the-big-data-company/ (zugegriffen 14. September 2016),

Hof, Christian van't, *RFID and Identity Management in Everyday Life - Striking the Balance between Convenience, Choice and Control - Think Tank*, European Technology Assessment Group, 2006, verfügbar unter http://www.europarl.europa.eu/thinktank/en/document.html?reference=IPOL-JOIN_ET(2007)383219 (zugegriffen 9. September 2016),

Hoffman, Sharona, *Medical Big Data and Big Data Quality Problems*, Social Science Research Network (Rochester, NY), 2014, verfügbar unter http://papers.ssrn.com/abstract=2464299 (zugegriffen 17. September 2016),

Huth, Alexa / Cebula, James, *The Basics of Cloud Computing*, Carnegie Mellon University, produced for United States Computer Emergency Readiness Team (Pittsburgh, PA), 2011,

IBM Cloudant, *Why NoSQL? Your Database Options In The New Non-Relational World*, IBM Corporation, 2010, verfügbar unter https://cloudant.com/wp-content/uploads/Why_NoSQL_IBM_Cloudant.pdf (zugegriffen 20. September 2016),

Information Commissioner's Office, *Anonymisation: managing data protection risk code of practice summary*, Information Commissioner's Office, 2012, verfügbar unter https://ico.org.uk/media/1042731/anonymisation_code_summary.pdf (zugegriffen 20. September 2016),

Internet of Things Global Standards Initiative, *Overview of the Internet of things - ITU-T Y.4000/Y.2060 (06/2012)*, International Telecommunication Union, 2012, verfügbar unter http://www.itu.int/ITU-T/recommendations/rec.aspx?rec=y.2060 (zugegriffen 24. August 2016),

Jain, Anil, *Data Clustering: 50 Years Beyond K-means*, Pattern Recognition Letters, Juni 2010, 651–666,

Junqué de Fortuny, Enric / Martens, David / Provost, Foster, *Predictive Modeling With Big Data: Is Bigger Really Better?*, Big Data, Dezember 2013, 215–226,

Kamlah, Wulf, *Das SCHUFA-Verfahren und seine datenschutzrechtliche Zulässigkeit*, Multimedia und Recht. Zeitschrift für Informations-, Telekommunikations- und Medienrecht, 1999, 395–404,

Kamp, Meike / Weichert, Thilo, *Scoringsysteme zur Beurteilung der Kreditwürdigkeit, Gutachten erstellt im Auftrag des Bundesverbraucherministeriums*, Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein, 2005, verfügbar unter https://www.datenschutzzentrum.de/scoring/2005-studie-scoringsysteme-uld-bmvel.pdf (zugegriffen 20. September 2016),

Karg, Moritz, *Die Rechtsfigur des personenbezogenen Datums. Ein Anachronismus des Datenschutzes?*, Zeitschrift für Datenschutz, 2012, 255–260,

Kastl, Graziana, *Algorithmen – Fluch oder Segen? Eine Analyse der Autocomplete-Funktion der Google-Suchmaschine*, Deutsche Stiftung für Recht und Informatik. Tagungsband Herbstakademie 2012, 2014, 203–218,

Katko, Peter / Babaei-Beigi, Ayda, *Accountability statt Einwilligung? Führt Big Data zum Paradigmenwechsel im Datenschutz?*, Multimedia und Recht. Zeitschrift für Informations-, Telekommunikations- und Medienrecht, 2014, 360–364,

King, Gary, *Computational Social Science: Discovery and Prediction*, Cambridge University Press (Cambridge, UK), 2016,

Korczak, Dieter / Wilken, Michael, *Scoring im Praxistest - Aussagekraft und Anwendung in der Kreditvergabe. Studie im Auftrag der Verbraucherzentrale Bundesverband e.V.*, GP Forschungsgruppe, 2008, verfügbar unter http://www.vzbv.de/dokument/studie-zum-scoring-aussagekraft-und-anwendung-der-kreditvergabe (zugegriffen 9. September 2016),

Kroschwald, Steffen, *Verschlüsseltes Cloud Computing. Auswirkung der Kryptografie auf den Personenbezug in der Cloud*, Zeitschrift für Datenschutz, 2014, 75–80,

Kubat, Miroslav, *An Introduction to Machine Learning*, Springer (Heidelberg; New York etc), 2015,

Kühling, Jürgen / Klar, Manuel, *Unsicherheitsfaktor Datenschutzrecht – Das Beispiel des Personenbezugs und der Anonymität*, Neue Juristische Wochenschrift, 2013, 3611–3617,

Laney, Doug, *3D Data Management*, Meta Group (Gartner), 2001, verfügbar unter https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (zugegriffen 20. September 2016),

Laplace, Pierre Simon : de, *Essai Philosophique sur les Probabilites; Par M. Le Comte Laplace*, Mme. Ve. Courcier (Paris), 1814,

Larson, Jeffrey / Bradlow, Eric / Fader, Peter, *An exploratory look at supermarket shopping paths*, International Journal of Research in Marketing, Dezember 2005, 395–414,

Leetaru, Kalev, *The Big Data Era of Mosaicked Deidentification: Can We Anonymize Data Anymore?*, Forbes, verfügbar unter http://www.forbes.com/sites/kalevleetaru/2016/08/24/the-big-data-era-of-mosaicked-deidentification-can-we-anonymize-data-anymore/ (zugegriffen 16. September 2016),

Li, Tiancheng / Li, Ninghui, *On the Tradeoff Between Privacy and Utility in Data Publishing* in: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2009, 517–526,

Lindemuth, Michael / Lembke, Chad, *Sparse glider datasets: A case study for NoSQL databases* in: *2013 OCEANS - San Diego*, MTS Publications, 2013, 1–6,

Lohr, Steve, *Big Data's Impact in the World*, The New York Times, 11. Februar 2012, verfügbar unter http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html (zugegriffen 17. September 2016),

Lohr, Steve, *The Origins of „Big Data": An Etymological Detective Story*, The New York Times, 1. Februar 2013, verfügbar unter http://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/ (zugegriffen 17. August 2016),

Luckham, David, *The instantly responsive enterprise*, Information Age, 2004, 20–24,

Lynch, Mike, *Barack Obama's Big Data won the US election*, Computerworld, 13. November 2012, verfügbar unter http://www.computerworld.com/article/2492877/government-it/barack-obama-s-big-data-won-the-us-election.html (zugegriffen 17. September 2016),

Malle, Bernd / Kieseberg, Peter / Weippl, Edgar / Holzinger, Andreas *The Right to Be Forgotten: Towards Machine Learning on Perturbed Knowledge Bases* in: Buccafurri, Francesco / Holzinger, Andreas / Kieseberg, Peter / Tjoa, A Min / Weippl, Edgar (Hrsg.), *Availability, Reliability, and Security in Information Systems*, Springer International (New York, NY), 2016, 251–266,

Manyika, James / Chui, Michael / Brown, Brad / Bughin, Jacques / Dobbs, Richard / Roxburgh, Charles / Byers, Angela Hung, *Big data: The next frontier for innovation, competition, and productivity*, McKinsey Global Institute, 2011, verfügbar unter http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation (zugegriffen 18. August 2016),

Marnau, Ninja, *Anonymisierung, Pseudonymisierung und Transparenz für Big Data: Technische Herausforderungen und Regelungen in der Datenschutz-Grundverordnung*, Datenschutz und Datensicherheit, Juli 2016, 428–433,

Martin, Scott, *Demystifying cloud computing for consumers*, USA TODAY, 6. November 2011, verfügbar unter http://www.usatoday.com/tech/news/2011-06-22-cloud-consumer-apple-google_n.htm (zugegriffen 21. September 2016),

Massachusetts Institute of Technology, *The MIT Big Data Challenge*, bigdata@CSAIL, 12. November 2013, verfügbar unter http://bigdata.csail.mit.edu/challenge (zugegriffen 15. September 2016),

Mayer-Schönberger, Viktor / Cukier, Kenneth, *Big Data: A Revolution That Will Transform How We Live, Work and Think*, Hodder and Stoughton (Boston, MA), 2013,

Meeker, Mary, *Internet Trends 2016 - Code Conference*, Kleiner Perkins Caufield & Byers, 2016, verfügbar unter kpcb.com/InternetTrends (zugegriffen 20. September 2016),

Meinicke, Dirk, *Big Data und Data Mining: Automatisierte Strafverfolgung als neue Wunderwaffe der Verbrechensbekämpfung?*, Deutsche Stiftung für Recht und Informatik. Tagungsband Herbstakademie 2012, 2014, 183–202,

Mell, Peter / Grance, Timothy, *SP 800-145. The NIST Definition of Cloud Computing*, National Institute of Standards & Technology (Gaithersburg, MD, United States), 2011, verfügbar unter http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf (zugegriffen 20. September 2016),

MIT Management Sloan School, *MIT Sloan professor uses machine learning to design crime prediction models - MIT Sloan School of Management*, Massachusetts Institute of Technology Sloan Experts Blog, 23. April 2015, verfügbar unter http://mitsloan.mit.edu/newsroom/press-releases/mit-sloan-professor-uses-machine-learning-to-design-crime-prediction-models/ (zugegriffen 17. September 2016),

Möller, Jan / Florax, Björn-Christoph, *Kreditwirtschaftliche Scoring-Verfahren. Verbot automatisierter Einzelentscheidungen gem. § 6a BDSG*, Multimedia und Recht. Zeitschrift für Informations-, Telekommunikations- und Medienrecht, 2002, 806–810,

Möncke, Ulrich *Sicherheit im Data Warehouse Profilbildung und Anonymität* in: Horster, Patrick / Fox, Dirk (Hrsg.), *Datenschutz und Datensicherheit*, Vieweg+Teubner (Wiesbaden), 1999, 30–59,

Montjoye, Yves-Alexandre de / Hidalgo, César A. / Verleysen, Michel / Blondel, Vincent D., *Unique in the Crowd: The privacy bounds of human mobility*, Scientific Reports, 25. März 2013,

Murphy, Kevin, *Machine Learning: A Probabilistic Perspective*, MIT Press (Cambridge, MA), 2012,

Narayanan, Arvind / Shmatikov, Vitaly, *Robust De-anonymization of Large Sparse Datasets* in: *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, IEEE Computer Society (Washington, DC, USA), 2008, 111–125,

Narayanan, Arvind / Shmatikov, Vitaly, *Myths and Fallacies of „Personally Identifiable Information“*, Communications of the ACM, Juni 2010, 24–26,

National Institute of Standards and Technology, *NIST/SEMATECH Engineering Statistics Handbook*, U.S. Department of Commerce, 2003, verfügbar unter http://www.itl.nist.gov/div898/handbook/ (zugegriffen 15. August 2016),

National Institute of Standards and Technology, *Big Data Interoperability Framework: Volume 7, Standards Roadmap*, U.S. Department of Commerce, 2015, verfügbar unter https://www.nist.gov/node/790346 (zugegriffen 20. September 2016),

neo4J, *Powering Recommendations with a Graph Database*, Neo Technology, 2015, verfügbar unter http://info.neotechnology.com/rs/neotechnology/images/GraphDB_Recommendations_EN.pdf (zugegriffen 20. September 2016),

neo4J, *Overcoming SQL Strain and SQL Pain*, Neo Technology, 2015, verfügbar unter 9/20/2016 (zugegriffen 20. September 2016),

Noll, Patrick, *Statistisches Matching mit Fuzzy Logic*, Vieweg+Teubner (Wiesbaden), 2009,

North, Douglass C., *Institutions, Transaction Costs and Economic Growth*, Economic Inquiry, 1. Juli 1987, 419–428,

Ohm, Paul, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, Social Science Research Network (Rochester, NY), 2009,

ORACLE, *An Enterprise Architect's Guide to Big Data. Reference Architecture Overview*, 2016, verfügbar unter http://www.oracle.com/technetwork/topics/entarch/articles/oea-big-data-guide-1522052.pdf (zugegriffen 20. September 2016),

Pandurangan, Vijay, *On Taxis and Rainbows*, 21. Juni 2014, verfügbar unter https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1 (zugegriffen 15. September 2016),

Pasquale, Frank A., *Restoring Transparency to Automated Authority*, Social Science Research Network (Rochester, NY), 2011, verfügbar unter http://papers.ssrn.com/abstract=1762766 (zugegriffen 17. September 2016),

Piltz, Carlo, *Benötigen wir Big Data-Kommissionen?*, Deutsche Stiftung für Recht und Informatik. Tagungsband Herbstakademie 2012, 2014, 149–160,

Pinto, Alexandre M., *A Comparison of anonymization protection principles* in: IEEE (Las Vegas, NV), 2012, 207–214,

Plummer, Andrew, *The Chicago Area Transportation Study: Creating the First Plan (1955-1962) - Travel Forecasting Resource*, 2010, verfügbar unter http://tfresource.org/The_Chicago_Area_Transportation_Study:_Creating_the_First_Plan_(1955-1962) (zugegriffen 15. September 2016),

Privacy Analytics, *Real Anonymization vs Data Masking*, Privacy Analytics Blog, 29. August 2016, verfügbar unter http://www.privacy-analytics.com/de-id-university/blog/real-anonymization-vs-data-masking/ (zugegriffen 16. September 2016),

Qi, Xinjun / Zong, Mingkui, *An Overview of Privacy Preserving Data Mining*, Procedia Environmental Sciences, 2012, 1341–1347,

Rauner, Max / Stockrahm, Sven, *Unglück in Duisburg: „Auf der Loveparade wurden Warnzeichen nicht erkannt“*, Zeit Online, 25. Juni 2012, verfügbar unter http://www.zeit.de/wissen/2012-06/loveparade-massenpanik-analyse/komplettansicht (zugegriffen 18. September 2016),

rbb-online, *Ordnungsamt-App kommt bei den Berlinern an*, 12. Juni 2016, verfügbar unter http://www.rbb-online.de/politik/beitrag/2016/07/app-ordnungsamt-online-erfolgreich.html (zugegriffen 15. September 2016),

Richter, Philipp, *Big Data, Statistik und die Datenschutz-Grundverordnung*, Datenschutz und Datensicherheit, September 2016, 581–586,

Robak, Markus, *Neue Abmahnrisiken im Datenschutzrecht*, Gewerblicher Rechtsschutz und Urheberrecht, Praxis im Immaterialgüter und Wettbewerbsrecht, 2016, 139–141,

Roßnagel, Alexander, *Big Data - Small Privacy? Konzeptionelle Herausforderungen für das Datenschutzrecht*, Zeitschrift für Datenschutz, 2013, 562–567,

Roßnagel, Alexander / Garstka, Hansjürgen / Pfitzmann, Andreas, *Modernisierung des Datenschutzrechts. Gutachten im Auftrag des Bundesministeriums des Innern*, Bundesministerium des Innern, 2001, verfügbar unter http://www.bfdi.bund.de/SharedDocs/VortraegeUndArbeitspapiere/2001GutachtenModernisierungDSRecht.pdf?__blob=publicationFile (zugegriffen 20. September 2016),

Rusch, Thomas / Lee, Ilro / Hornik, Kurt / Jank, Wolfgang / Zeileis, Achim, *Influencing Elections with Statistics: Targeting Voters with Logistic Regression Trees*, Social Science Research Network (Rochester, NY), 2012, verfügbar unter http://papers.ssrn.com/abstract=2016956 (zugegriffen 8. September 2016),

Salzberg, Steven, *Why Google Flu Is A Failure*, Forbes, 23. März 2014, verfügbar unter http://www.forbes.com/sites/stevensalzberg/2014/03/23/why-google-flu-is-a-failure/print/ (zugegriffen 17. September 2016),

Schefzig, Jens, *Big Data = Personal Data? Der Personenbezug von Daten bei Big Data-Analysen*, Deutsche Stiftung für Recht und Informatik. Tagungsband Herbstakademie 2012, 2014, 103–118,

Schonschek, Oliver, *Anonymisierung für Big Data auf die andere Art*, Experton Group ICT-News Dach, 2. Mai 2014, verfügbar unter http://www.experton-group.de/research/ict-news-dach/news/anonymisierung-fuer-big-data-auf-die-andere-art.html (zugegriffen 16. September 2016),

Schroeck, Michael / Shockley, Rebecca / Smart, Janet / Romero-Morales, Dolores / Tufano, Peter, *Analytics – Big Data in der Praxis*, IBM Institute for Business Value, 2012, verfügbar unter http://www-935.ibm.com/services/de/gbs/thoughtleadership/studie-bigdata.html (zugegriffen 17. August 2016),

Schweda, Sebastian, *HmbBfDI: Datenschützer billigt Einsatz von Google Analytics in Deutschland*, MMR-Aktuell. Newsdienst. Multimedia und Recht Zeitschrift für Informations-, Telekommunikations- und Medienrecht, 2011,

Sharma, Sugam, *An Extended Classification and Comparison of NoSQL Big Data Models*, arXiv:1509.08035 [cs], 26. September 2015,

Shashid, Syed Mohammad, *Use of RFID Technology in Libraries: a New Approach to Circulation, Tracking, Inventorying, and Security of Library Materials*, Library Philosophy and Practice, 2005,

Shoshani, Arie, *OLAP and Statistical Databases: Similarities and Differences* in: *Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, ACM (New York, NY), 1997, 185–196,

Siddle, James, *I Know Where You Were Last Summer: London's public bike data is telling everyone where you've been*, 10. April 2014, verfügbar unter https://vartree.blogspot.com/2014/04/i-know-where-you-were-last-summer.html (zugegriffen 22. September 2016),

Siebes *Data Mining: Exploratory Data Analysis on Very Large Databases* in: Krzysztof, Abt / Schrijver, Lex / Temme, Nico (Hrsg.), *From universal morphisms to megabytes : a Baayen space Odyssey*, CWI (Amsterdam), 1994, 535–557,

Simitis, Spiros, *Bundesdatenschutzgesetz*, Nomos (Baden-Baden), 8. Auflage 2014 (Zitiert: *Bearbeiter,* in: Simitis),

Simonite, Tom, *The foundation of the computing industry's innovation is faltering. What can replace it?*, Massachusetts Institute of Technology. Technology Review, 13. Mai 2016, verfügbar unter https://www.technologyreview.com/s/601441/moores-law-is-dead-now-what/ (zugegriffen 11. September 2016),

Siraj, Fadzilah / Ali, Mansour *Mining Enrollment Data Using Descriptive and Predictive Approaches* in: Funatsu, Kimito (Hrsg.), *Knowledge-Oriented Applications in Data Mining*, InTech (Rijeka, Croatia), 2011,

Song, Chaoming / Qu, Zehui / Blumm, Nicholas / Barabási, Albert-László, *Limits of Predictability in Human Mobility*, Science, 19. Februar 2010, 1018–1021,

Sorensen, Herb, *The Science of Shopping*, Marketing Research, 1. September 2003, 30–35,

Spindler, Gerald, *Verbandsklagen und Datenschutz – das neue Verbandsklagerecht-Neuregelungen und Probleme*, Zeitschrift für Datenschutz, 2016, 114–119,

Stein, Joel, *Data Mining: How Companies Now Know Everything About You*, Time, 10. März 2011,

Steinmüller, Wilhelm, *Das Volkszälungsurteil des Bundesverfassungsgerichts*, Datenschutz und Datensicherheit, 1984, 91–96,

Steinmüller, Wilhelm, *Informationstechnologie und Gesellschaft*, Wissenschaftliche Buchgesellschaft (Darmstadt), 1993,

Steinmüller, Wilhelm / Lutterbeck, Bernd / Mallmann, Christoph / Harbort, Uwe / Kolb, Gerhard / Schneider, Jochen, *Schutz der Privatsphäre Bezug: Kleine Anfrage der Fraktionen der SPD, FDP. Gutachten im Auftrag des Bundesministeriums des Innern - Drucksache VI/3826*, Deutscher Bundestag 6. Wahlperiode, 1971, verfügbar unter http://dipbt.bundestag.de/doc/btd/06/038/0603826.pdf (zugegriffen 20. September 2016),

Sweeney, Latanya, *Simple Demographics Often Identify People Uniquely*, Carnegie Mellon University (Pittsburgh), Data Privacy Working Paper 3rd Edition 2000,

Tan, Pang-Ning / Steinbach, Michael / Kumar, Vipin, *Introduction to Data Mining*, Pearson Addison-Wesley (Boston, MA; San Francisco, CA etc.), 2006,

Tasse, Dan / Hong, Jason, *Using Social Media Data to Understand Cities*, Proceedings of NSF Workshop on Big Data and Urban Informatics, 2014, verfügbar unter http://repository.cmu.edu/hcii/272 (zugegriffen 20. September 2016),

Tencent, *Tencent Announces 2016 Second Quarter And Interim Results*, Tencent Holdings Limited, 2016, verfügbar unter http://www.tencent.com/en-us/ir/news/2016.shtml (zugegriffen 20. September 2016),

Teradata Perspectives, Teradata, *TeradataVoice: Big Data Needs More „Creative Types"*, Forbes, 30. Januar 2015, verfügbar unter http://www.forbes.com/sites/teradata/2015/01/30/big-data-needs-more-creative-types/print/ (zugegriffen 17. September 2016),

Teschl, Gerald / Teschl, Susanne, *Mathematik für Informatiker. Band 2. Analysis und Statistik*, Springer (Berlin: Heidelberg), 3. Auflage 2014,

The City of New York, *NYC OpenData - Taxi and Limousine Commission Official Data*, NYC Open Data, 2015 2009, verfügbar unter https://data.cityofnewyork.us/data?browseSearch=&scope=&type=new_view&agency=Taxi%20and%20Limousine%20Commission%20(TLC) (zugegriffen 15. September 2016),

The Editorial Board, *Credit History Discrimination*, The New York Times. The Opinion Pages. Editorial, 22. April 2013, verfügbar unter http://www.nytimes.com/2013/04/23/opinion/credit-history-discrimination.html (zugegriffen 28. August 2016),

The Linux Information Project, *Vendor Lock-in Definition*, 2006, verfügbar unter http://www.linfo.org/vendor_lockin.html,

Tockar, Anthony, *Differential Privacy: The Basics*, neustar. Research, 8. September 2014, verfügbar unter https://research.neustar.biz/2014/09/08/differential-privacy-the-basics/ (zugegriffen 16. September 2016),

Troester, Mark, *Big Data Meets Big Data Analytics.*, SAS Institute (Cary, NC), 2011, verfügbar unter http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/big-data-meets-big-data-analytics-105777.pdf (zugegriffen 20. September 2016),

Tukey, John W., *Exploratory Data Analysis*, Pearson (Reading, MA), 1977,

Türpe, Sven / Selzer, Annika / Poller, Andreas / Bedner, Mark, *Denkverbote für Star-Trek-Computer?*, Datenschutz Datensicherheit, 22. Januar 2014, 31–35,

ULD/GP Forschungsgruppe, *Scoring nach der Datenschutz-Novelle 2009 und neue Entwicklungen*, Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein; GP Forschungsgruppe (Kiel; München), 2014, verfügbar unter https://www.datenschutzzentrum.de/artikel/902-Studie-Scoring-nach-der-Datenschutz-Novelle-2009-und-neue-Entwicklungen.html (zugegriffen 20. September 2016),

Varia, Jinesh, *Cloud Architectures*, Amazon Web Services, 2008, verfügbar unter https://jineshvaria.s3.amazonaws.com/public/cloudarchitectures-varia.pdf (zugegriffen 20. September 2016),

Velickovic, Vladica, *What Everyone Should Know about Statistical Correlation*, American Scientist, 2015, 26,

Vidgen, Richard / Wood-Harper, Trevor / Wood, Robert, *A Soft Systems Approach To Information Systems Quality*, Scandinavian Journal of Information Systems, 1993, 97–112,

Voß, Jakob, *Was sind eigentlich Daten?*, LIBREAS. Library Ideas, 2013,

Ward, Jonathan Stuart / Barker, Adam, *Undefined By Data: A Survey of Big Data Definitions*, arXiv:1309.5821 [cs], 20. September 2013,

Wasserman, Larry, *All of Statistics: A Concise Course in Statistical Inference*, Springer (New York), 20 2004,

Weichert, Thilo, *Big Data und Datenschutz. Chancen und Risiken einer neuen Form der Datenanalyse*, Zeitschrift für Datenschutz, 2013, 251–259,

Wernecke, Klaus D., *Angewandte Statistik für die Praxis*, Addison-Wesley (Bonn; Paris etc.), 1995,

Werner, Kathrin, *Werbung: Es ist egal, wer ein Haustier hat*, Sueddeutsche Zeitung, 10. August 2016, verfügbar unter http://www.sueddeutsche.de/wirtschaft/werbung-es-ist-egal-wer-ein-haustier-hat-1.3115538 (zugegriffen 22. August 2016),

Wetzel, Daniel, *Amprion baut erste Stromtrasse unter der Erde*, Welt Online, 22. September 2015, verfügbar unter https://www.welt.de/wirtschaft/energie/article132723021/Hier-entsteht-die-unsichtbare-Stromautobahn.html (zugegriffen 18. September 2016),

Willis, John, *Docker and the Three Ways of DevOps*, Docker Incorporated, 2015, verfügbar unter https://www.docker.com/sites/default/files/WP_Docker%20and%20the%203%20ways%20devops_07.31.2015%20(1).pdf (zugegriffen 20. September 2016),

Wolber, Tanja, *Datenschutzrechtliche Zulässigkeit automatisierter Kreditentscheidungen.*, Computer und Recht, 2003, 623,

xkcd, *Statistically significant other*, verfügbar unter http://xkcd.com/539/ (zugegriffen 18. September 2016),

Yue, Yisong / Lucey, Patrick / Carr, Peter / Bialkowski, Alina / Matthews, Iain, *Learning Fine-Grained Spatial Models for Dynamic Sports Play Prediction* in: *2014 IEEE International Conference on Data Mining*, IEEE, 2014, 670–679,

Zandbergen, Paul A., *Ensuring Confidentiality of Geocoded Health Data: Assessing Geographic Masking Strategies for Individual-Level Data*, Advances in Medicine, 29. April 2014,

Zikopoulos, IBM Paul / Eaton, Chris / Zikopoulos, Paul, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw Hill Professional (New York, NY; Chicago, IL etc.), 2011,

Zinke, Michaela, *Chancen und Risiken neuer Daten-basierter Dienste für die Industrie*, Deutsche Stiftung für Recht und Informatik. Tagungsband Herbstakademie 2012, 2014, 161–170,