Big Data

von Jan Schallaböck, iRights.Law für iRights.Lab

Themenpapier im Projekt „Braucht Deutschland einen Digitalen Kodex?“

1. Einleitung

Die Digitalisierung ermöglicht das Sammeln und Verarbeiten von Daten in einem neuen, bislang unbekannten Ausmaß. Jede Äußerung – etwa wem ich wann eine E-Mail schreibe – und jede Aktivität – wann ich welche Webseite an welchem Ort anschaue – kann gespeichert und ausgewertet werden. Durch die zunehmende Verdatung der Welt – Sensoren in Haushaltsgeräten, Überwachungskameras im öffentlichen Raum, Funkzellenabfrage in der mobilen Telefonie – entstehen immer mehr Daten, die von verschiedenen Stellen ausgewertet und genutzt werden. Das sind auf der einen Seite der Staat und seine Organe, aber auch zunehmend private Unternehmen. Die bekanntesten sind wohl die großen Internetunternehmen wie Facebook, Google und Amazon – sie sind aber nicht die einzigen.

Große Datensammlungen sind kein neues Phänomen. Von der Volkszählung 1987, die in Deutschland von einer kontroversen Debatte begleitet wurde, bis zum heutigen Internet der Dinge sind sie immer wieder von gesellschaftlichen Diskussionen begleitet worden. Wissenschaftler, Politiker, Kulturkritiker und Experten spekulieren über ihre gesellschaftlichen Auswirkungen – mit durchaus unterschiedlichen Ergebnissen.

Der Zeitpunkt und die begriffliche Ausrichtung der derzeitigen Debatte sind nicht zufällig. Die Analyse von umfangreichen Datenbeständen ist seit Jahrzehnten in Wirtschaft und Wissenschaft etabliert, jedoch haben neue Technologietrends zu einer erheblichen Beschleunigung geführt. In den letzten zehn Jahren hat sich sowohl die Verfügbarkeit als auch die technische Machbarkeit von Analysen erheblich weiterentwickelt. Damit wird eine Vielzahl bislang unerschlossener Auswertungen möglich.

Mit Big Data lassen sich erhebliche wirtschaftliche Potentiale realisieren: Das können effektivere Werbeschaltungen auf Webseiten sein, wo den Nutzern anhand ihres Nutzungsprofils relevante Werbung angeboten wird, oder intelligente Thermostate, die ihre Heizleistung der tatsächlichen Nutzung anpassen, die aufgrund von Big-Data-Analysen errechnet wurde. Es entstehen neue Geschäftsmodelle und Produkte, die effizienter und intelligenter arbeiten – und im Idealfall Gewinn bringen.

In der Wissenschaft kann Big Data erkenntnisbringend eingesetzt werden: In der Medizin werden über Big-Data-Analysen Krankheitsursachen und Heilungsmethoden entwickelt, in der Soziologie können Bevölkerungsbewegungen analysiert und in der Linguistik neue automatische Übersetzungstechnologien eingeführt werden, um nur einige Einsatzmöglichkeiten zu nennen.

Big Data ist aber nicht nur positiv besetzt. Der Begriff schürt Ängste und Befürchtungen, insbesondere vor umfassenden Persönlichkeitsprofilen und Verhaltensprognosen. Viele Kritiker sehen einen umfassenden Überwachungsstaat auf uns zukommen. Wenn private Anbieter Daten sammeln und sie verarbeiten, erwerben sie intime Kenntnisse über die einzelnen Nutzer. Sie können daraus Rückschlüsse über das Individuum, einzelne Personengruppen oder die gesamte Gesellschaft ziehen. Problematisch daran ist, dass meist nicht gesichert ist, wer was mit den Daten macht, ob sie weitergegeben und mit anderen

Daten in Beziehung gesetzt werden und wer welche Erkenntnisse aus ihnen ableitet. Die Einzelnen wissen in aller Regel nicht, welche Daten zu welchen Zwecken gesammelt werden, was mit ihnen geschieht und welche Chancen und Risiken die Sammlung und Verarbeitung für sie birgt. Selbst wenn Nutzer in einer Datenschutzerklärung der Verarbeitung zustimmen, treffen sie oft keine informierte Entscheidung, sondern akzeptieren einfach die Bedingungen, die vorgegeben werden. Bequemlichkeit geht vor informationelle Selbstbestimmung.

Dazu kommt ein weiteres Problem: Selbst wenn ich zustimme, dass meine Daten ausgewertet werden dürfen, kann ich diese Entscheidung nicht für Dritte treffen. Genau dies passiert aber. Der New Yorker Rechtswissenschaftler Eben Moglen spitzte dies in einem Vortrag an der Columbia Law School zu, indem er sagte: „Privacy is not transactional.“[1] Sammlungen von persönlichen Daten wirken auch jenseits des Sammlers und des von der Sammlung Betroffenen. Denn die Verallgemeinerungen, die sich aus statistischen Auswertungen ergeben, führen zu Zuschreibungen auf Dritte. So funktioniert jedes Kreditscoring. Vereinfacht gesagt: Wenn bei 28-jährigen Männern Kredite häufiger ausfallen, bekommt ein 28-jähriger Mann keinen Kredit. Diese Drittbetroffenheit könnte gravierende Konsequenzen für bestehende Regulierungsansätze etwa des Datenschutzrechtes haben, die regelmäßig auf die Übereinkunft zwischen Datenverarbeiter und dem unmittelbar von der Verarbeitung Betroffenen setzen, Dritte aber nicht einbeziehen.

Der österreichische Jurist und Professor am Oxford Internet Institute Victor Mayer-Schönberger schlägt vor, sich darauf zu konzentrieren, wie Daten verarbeitet werden und wozu sie genutzt werden, statt ausschließlich auf die Frage, ob sie überhaupt erhoben werden.[2] Vieles spricht dafür, die Verarbeitungsprozesse stärker ins Blickfeld zu rücken. Neue Methoden der Datenanalyse haben auch in der Vergangenheit dazu geführt, dass gesellschaftliche Übereinkünfte neu geregelt werden müssen. Stephan Noller, Geschäftsführer von nugg.ad, einer Plattform für zielgruppenorientierte Online-Werbung, forderte auf der Sommerakademie des Unabhängigen Landeszentrums für Datenschutz Schleswig-Holstein zum selben Thema sogar die Einführung von ethischen Grundsätzen für Algorithmen, eine „Algorithmen-Ethik“.

Solche Aushandlungsprozesse sollten auf einem soliden Wissensfundament und ethischen Grundlage stattfinden. Damit ist das Thema Big Data auf mehreren Ebenen prädestiniert, im Rahmen des Projektes „Braucht Deutschland einen Digitalen Kodex?“ näher betrachtet zu werden.

Unter dem Begriff „Big Data“ wird eine Vielzahl von – teilweise komplexen – technischen und gesellschaftlichen Verfahren aus unterschiedlichsten Anwendungsfeldern diskutiert

---

[1] Eben Moglen, Snowden and the future, Part III: The Union, May it be Preserved, Vortrag an der Columbia Law School am 13. November 2013. Online unter http://snowdenandthefuture.info/PartIII.html.

[2] Viktor Mayer-Schönberger und Kenneth Cukier, Big Data: Die Revolution, die unser Leben verändern wird. 2013.

und problematisiert. Um diese vielfältigen Sachverhalte greifbarer zu machen, umreißt dieser Text zunächst einige Beispiele, die veranschaulichen sollen, um welche Art der Datenverarbeitung es sich handelt, und welche gesellschaftlichen Diskurse daraus folgen. Ausgehend von diesen Beispielen identifizieren wir anschließend einige Grundlagen und Technologietrends, die für Big Data charakteristisch sind. Im Folgenden werden dann einige der Problemfelder und die Potentiale von Big Data verdeutlicht, um dann schließlich in der abschließenden Zusammenfassung in konkrete Fragestellungen für Aushandlungsprozesse zu münden, die einen Beitrag für einen weiteren Diskussionsprozess darstellen sollen.

2. Was ist Big Data? – exemplarische Nutzungsszenarien

Wie der Name schon sagt: Bei Big Data geht es um große Datenmengen. Die im Rahmen der Digitalisierung verfügbaren einzelnen Daten sind immens: Immer mehr Sensoren sammeln Daten. Das können Wetterdaten sein, Verkehrsdaten, das Mapping von Genen, große Sprachkorpora, Logdateien von Webservern und Nutzerinteraktionen, Überwachungskameras im öffentlichen Raum, RFID-Reader und so weiter. Big Data ist als Begriff weit gefasst und nicht fest definiert, sodass der Gebrauch des Begriffes weitere Aspekte umfassen kann.

Besonders wichtig für Big Data sind die neuen Möglichkeiten der Verarbeitung und Analyse. Durch parallelisierte Datenverarbeitung in großen Rechenzentren lassen sich diese Datenmengen teilweise in Echtzeit verarbeiten. Dabei ändert sich die Herangehensweise: Man sucht nicht mehr in den Daten nach Beweisen für vorher entworfene Theorien, sondern untersucht die Datensätze nach Korrelationen und lässt sie selbst „sprechen“, wie es Viktor Mayer-Schönberger in seinem Buch ausdrückt.

Weitgehend akzeptiert ist die Analyse großer Datenmengen wohl im Rahmen klassischer, akademischer Forschung. Sie funktioniert regelmäßig einwilligungsbasiert und ist sowohl in Hinblick auf den Zweck, die Verfahren und Ziele weitgehend transparent. Zudem werden nach akademischer Tradition die Ergebnisse veröffentlicht und dienen damit der Allgemeinheit.[3] Grundrechtlich schafft die Wissenschaftsfreiheit des Grundgesetzes eine befürwortende Wertungslage.

Ebenfalls weitgehend unumstritten sind jene Bereiche, in denen nicht die Beobachtung von Personen Gegenstand der Erhebung ist, etwa beim Large Hadron Collider in Genf, der Sternen- und der Wetterbeobachtung. Problematisch und diskursrelevant scheinen damit vor allen Dingen jene großen Datensammlungen zu sein, die menschliche Verhaltensweisen zum Gegenstand haben. Dies gilt insbesondere dann, wenn sie jenseits wissenschaftlicher Tätigkeit in der Privatwirtschaft und im öffentlichen Sektor stattfinden.

[3] Dennoch beginnt jüngst auch hier eine Debatte, vgl. die Gründung eines „Council for Big Data, Ethics, and Society” in den USA, Meldung online unter: http://www.whitehouse.gov//sites/default/files/microsites/ostp/Data2Action%20Announcements.pdf.

Erhebung und Speicherung von großen Datenbeständen sind vielfach von gesellschaftlichen Aushandlungsprozessen begleitet. Diese können sich um das Ob der Erhebung, aber auch um das Wie der Verarbeitung drehen. Im Folgenden sollen exemplarisch verschiedene Bereiche nachgezeichnet werden, in denen Big Data eine Rolle spielt, und die in der Gesellschaft mehr oder weniger breit diskutiert werden.

2.1. Der erste Big-Data-Diskurs in Deutschland: Das Volkszählungsurteil

Der Datenbestand der Volkszählung von 1987 nimmt mit einen Speichervolumen von geschätzt 80 Megabyte einen Umfang ein, der allenfalls Festplatten aus den 1980er Jahren an ihre Grenzen bringt. Dennoch wird man bei diesem Datenbestand von Big Data sprechen können. Das Ziel war, aktuelle Zahlen über Bevölkerung, Versorgung, Verkehrsmittel und so weiter zu erhalten, um notwendige infrastrukturelle Maßnahmen einzuleiten.

Die Volkszählung ist ein gutes Beispiel dafür, wie die Erhebung von großen Datenbeständen öffentliche Diskurse auslöst. Die für 1983 festgesetzte Volkszählung musste ausgesetzt werden und wurde schließlich im gleichen Jahr vom Bundesverfassungsgericht untersagt. Dieses Urteil des Bundesverfassungsgerichts setzte gewichtige Eckpfeiler für unser heutiges Datenschutzrecht, indem es den Grundsatz der „informationellen Selbstbestimmung“ definierte. Es machte zum einen Vorgaben für die Erhebung von Daten und etablierte zum anderen Prinzipien wie Datensicherheit und Zweckbindung der Daten, die für die anschließende Verarbeitung relevant sind.

2.2. Internetnutzungsdaten: Rezeptions- und Kommunikationsverhalten

Das jährliche Internetnutzungsverhalten einer Person inklusive des Inhalts aller besuchten Internetseiten beträgt bei einer durchschnittlichen Nutzung für drei Stunden pro Tag etwa 65 Gigabyte und könnte gut lokal auf einer handelsüblichen Festplatte abgelegt werden.

Der US-Mathematiker und Unternehmer Stephen Wolfram protokolliert seit vielen Jahren die Interaktion mit seinen Rechnern. Zusätzlich zu der Analyse seiner E-Mail-Daten loggt er mit einem sogenannten Keylogger jede seiner Tastatureingaben mit. Wie er zeigt, lassen sich daraus interessante Ableitungen gewinnen. Er kann ablesen, welche Anwendungen er in den vergangenen Jahren intensiv genutzt hat und dass seine Fehlerraten beim Tippen relativ konstant geblieben sind.[4] Hierbei entstehen auch bei intensiver Nutzung pro Tag Datenmengen im Bereich von lediglich hundert Kilobyte, mithin pro Jahr eine Datenmenge, die sich noch auf einem längst überholten Speichermedium wie einer Diskette speichern ließe. Wolfram reiht sich mit seinen Selbstprotokollen ein in eine Bewegung, die unter dem Begriff „Quantified Self“ bekannt geworden ist – das „quantifizierte Ich“. Dabei tracken die Anwender etwa Gesundheitsdaten, wie oft sie Sport machen, zählen ihre Schritte oder andere persönliche Daten. Erste Geschäftsideen und Applikationen, die eine solche Selbstbeobachtung unterstützen, existieren bereits: etwa Fitnessarmbänder, die

[4] Vgl. http://blog.stephenwolfram.com/2012/03/the-personal-analytics-of-my-life/.

aufzeichnen, wie oft man sich bewegt, ob man ausreichend schläft und ähnliches. Bei der neuen iOS-Version von Apple ist die „Health“-App schon vorinstalliert, mit der Nutzer ihre Blutwerte, Gewichtsdaten, Schlaf oder Vitalzeichen aufzeichnen können. Solange diese Daten beim Nutzer selbst bleiben und er Auswertungen seines Nutzungsverhaltens selbst vornimmt, hat dies keine problematischen Auswirkungen. Zum Problem wird es erst, wenn diese Daten zentral gesammelt und von externen Firmen ausgewertet werden. Auch anonymisiert können daraus Profile erzeugt werden.

Für die Online-Werbevermarktung sind das interessante Daten. Große Teile des globalen Nutzungsverhaltens werden bereits an zentralen Stellen erhoben und ausgewertet. Dieses ist begleitet von einer kontroversen öffentlichen Debatte. So gab es kürzlich eine intensive Diskussion um die Einführung einer „Do-not-track“-Funktion für Browser, die es den Nutzern erlauben sollte, selbst zu steuern, welche Webseiten ihn tracken dürfen und welche nicht. Die damit verbundenen Implikationen für bestehende Geschäftsmodelle haben zu erheblichen Kontroversen in technischen Standardisierungsgremien geführt. Auch der deutsche und europäische Gesetzgeber hat sich wiederholt kontrovers mit der Regulierung der Onlinewerbung befasst und die Grundlagen für eine Erhebung differenziert (und kompliziert) ausgestaltet.[5]

Neben dem Tracking des Rezeptionsverhaltens entstehen im Internet umfassende Datenbestände über das Kommunikationsverhalten. Diese Datenbestände bestehen auf der einen Seite aus den eigentlichen Inhalten der Kommunikation (zum Bespiel dem Wortlaut von E-Mails, Nachrichten oder Chats), die im Rahmen gezielter Werbeschaltung analysiert werden, auf der anderen vor allen Dingen aus den sogenannten Metadaten der Kommunikation, also wer, wann mit wem kommuniziert hat. Diese Daten sind – insbesondere unter Anwendung von Big Data-Methoden – ausgesprochen aufschlussreich.

Ein Beispiel dafür ist die Analyse der Kommunikationsmetadaten von Arbeitnehmern, aus der sich Kündigungswahrscheinlichkeiten ableiten lassen.[6] Der Arbeitgeber kann so schon vor dem Arbeitnehmer abschätzen, ob er dem Betrieb erhalten bleibt oder eigentlich schon auf dem Absprung ist. Ob eine solche Analyse in Deutschland rechtlich zulässig ist, ist zweifelhaft, aber nicht in allen Fällen eindeutig. Eine Überarbeitung des Arbeitnehmerdatenschutzes ist seit Jahren auch in Deutschland in der Diskussion, aber ein entsprechender Kabinettsentwurf wurde Anfang 2013 wieder von der Agenda genommen. Sollten sich solche Verfahren etablieren, ist wohl mit einer erheblichen Dynamisierung des bisher eher schleppenden parlamentarischen Beratungsprozesses bei der Novellierung des Arbeitnehmerdatenschutzes zu rechnen.

---

[5] Insbesondere in den Absätzen 3, 3a und 3b des § 28 Bundesdatenschutzgesetzes (BDSG).

[6] Vgl. etwa Frank Rieger: Der Mensch wird zum Datensatz. Veröffentlicht unter anderem auf FAZ.NET, dem Onlineauftritt der Frankfurter Allgemeinen Zeitung am 16. Januar 2010, online: http://www.faz.net/aktuell/feuilleton/ein-echtzeit-experiment-der-mensch-wird-zum-datensatz-1591336.html.

2.3. Sammlung von Nutzerdaten durch Geheimdienste

Neben der Sammlung von Nutzerdaten von Seiten privater Firmen ist seit letztem Jahr – mit den Enthüllungen von Edward Snowden – die Datensammlung durch Geheimdienste, insbesondere durch die NSA, für eine erhebliche öffentliche Auseinandersetzung gesorgt.

Über die Speicherkapazitäten der NSA existieren unterschiedliche und bisweilen absurd erscheinende Spekulationen. Selbst wenn man ersten Einschätzungen der Magazine Der Spiegel[7] und Wired[8] nicht trauen mag, die Kapazitäten seien im Bereich von „Yottabytes“ (ein Yottabyte sind 10 hoch 24 Bytes oder etwa das 1000-fache der von der Firma Cisco antizipierten globalen Internetkommunikation für 2015[9]), erscheint es nicht ganz abwegig, dass eine Speicherung aller digital vermittelten menschlichen Kommunikations- und Rezeptionssachverhalte in Kürze möglich sein wird. Eine detaillierte Analyse, was und wie viel die Geheimdienste sammeln, steht trotz des großen öffentlichen Interesses noch aus. Ein sachlicher öffentlicher Diskurs ist jedoch schwierig, weil es um staatliche Geheimnisse geht. Der gesellschaftliche Aushandlungsprozess findet trotzdem statt, wenngleich unter erschwerten Bedingungen und in einer frühen Phase.

2.4. Sensordaten: das Internet der Dinge

Ebenfalls noch nicht abgeschlossen sein dürften die Aushandlungsprozesse über die Gefahren und Potentiale des sogenannten Internets der Dinge, also etwa moderne Hausmesstechnik und -steuerung („smart home“, „smart metering“, „smart grid“), Smart Cars (bis hin zu selbstfahrenden Fahrzeugen) und Smart Watches (die als Fitnesstracker und PDAs dienen).

Bei der privaten Heizungstechnik, besonders bei Geräten mit starkem Elektrizitätsverbrauch, sind derzeit erhebliche digitale Produktinnovationen zu beobachten. Über die in diesen Produkten enthaltene Messtechnik werden detaillierte Daten erhoben: Wann die Bewohner da sind, welches ihre bevorzugte Raumtemperatur ist und vieles mehr. Die darauf basierenden Innovationen versprechen erhebliche Einsparungen für den Energieverbrauch.

Im vergangenen Jahr hat die Markteinführung von „Nest“, einem „smarten“ Thermostat, öffentliche Aufmerksamkeit erfahren. Das Thermostat sammelt neben den Angaben über den Installationsort Sensordaten und Nutzungsdaten des Geräts und sendet sie an das Unternehmen Nest. Laut Datenschutzerklärung misst das Thermostat Raumtemperatur, Luftfeuchtigkeit und Lichteinwirkung. Welche Sensoren das Gerät hierfür verwendet, teilt die Firma nicht mit, genauso wenig, wie häufig gemessen wird. Darüber hinaus meldet ein

---

[7] http://www.spiegel.de/netzwelt/netzpolitik/bluffdale-das-datensammel-zentrum-der-nsa-a-904355.html

[8] http://www.wired.com/2012/03/ff_nsadatacenter/all/

[9] Cisco (Hrsg.), Cisco Visual Networking Index: Forecast and Methodology, 2013–2018, online: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.pdf

Bewegungsmelder, ob sich etwas im Raum bewegt.[10] Das Nest-Thermostat lernt aus dem Verhalten der Nutzer, die das Gerät über ihr Smartphone steuern können. Auf diesem Wege sollen der Verbrauch für Heizung und Klimatechnik um bis zu 26 Prozent reduziert werden können. Anfang 2014 wurde das Unternehmen von Google zum Preis von 3,2 Milliarden US-Dollar übernommen. Allerdings ist bisher weitgehend unbekannt, wie Nest die entstehenden Daten verarbeitet.

Die Verbreitung von verbesserter Mess- und Steuerungstechnik bei der Elektrizität durch „intelligente Zähler" soll nicht nur den Stromverbrauch in privaten Haushalten senken, sondern auch Lastunterschiede im Stromnetz besser nivellieren können.[11] Weitgehend unbemerkt von der Öffentlichkeit hat hierzu ein umfangreicher Fachdiskurs stattgefunden, der zur Entwicklung zahlreicher technischer Normen geführt hat, die dem Datenschutz- und IT-Sicherheitsaspekten einen Stellenwert zumessen.[12] Neben wirtschaftswissenschaftlichen Untersuchungen wird die Einführung intelligenter Stromzähler in Deutschland auch durch Forschung über die datenschutzrechtlichen Implikationen begleitet.[13] Ihr Gegenstand sind die Mechanismen, die eine Datenerhebung und Weitergabe im Sinne der intendierten Analyseverfahren sicherstellen sollen. Der Diskurs nimmt hier also sowohl die Prozesse als auch die Verfahren ins Blickfeld.

Fitnesstracker – Geräte zur Messung der eigenen – insbesondere körperlichen – Aktivität – finden in den letzten Jahren zunehmend einen Markt. Die Geräte erfassen neben Schritten und zurückgelegter Strecke zumeist die Pulsfrequenz. Krankenversicherungen beginnen nun Interesse an den entstehenden Datenbeständen zu entwickeln. Je nach gemessener sportlicher Betätigung gewähren einzelne Versicherungen Prämien, so zum Beispiel beim Versicherer Generali.[14]

Diesem Produktbereich kommt eine besondere Relevanz zu, weil völlig neue, bisher unerschlossene Datenquellen zugänglich gemacht werden, die gleichzeitig einen hochgradig

---

[10] Vgl. http://nest.com/legal/privacy-statement/.

[11] Hauptsächlich auf indirekte Effekte abstellend und ohne konkrete Zahlen zu den tatsächlich erwarteten Einsparungen beim Energieverbrauch: http://www.bmwi.de/BMWi/Redaktion/PDF/Publikationen/Studien/kosten-nutzen-analyse-fuer-flaechendeckenden-einsatz-intelligenterzaehler,property=pdf,bereich=bmwi2012,sprache=de,rwb=true.pdf. Für den Bereich privater Haushalte ist die Sinnhaftigkeit des Einsatzes unterdessen allerdings wohl fraglich, vgl. https://www.bdew.de/internet.nsf/id/eine-frage-der-perspektive-de dort unter Verweis auf eine im Auftrag der RWE in Mühlheim durchgeführte Breitenstudie, nach der nur relativ geringe Effizienzsteigerungen möglich sein sollen.

[12] Bundesamt für Sicherheit in der Informationstechnologie (Hrsg.), Schutzprofil für die Kommunikationseinheit eines intelligenten Messsystems für Stoff- und Energiemengen (BSI-CC-PP-0073), online: https://www.bsi.bund.de/DE/Themen/SmartMeter/Schutzprofil_Gateway/schutzprofil_smart_meter_gateway_node.html, sowie International Standardisation Organisation (Hrsg.), ISO/IEC TR 27019:2013 Information technology – Security techniques – Information security management guidelines based on ISO/IEC 27002 for process control systems specific to the energy utility industry.

[13] Oliver Raabe et al., Datenschutz in Smart Grids: Anmerkungen und Anregungen, 2011.

[14] Ralf Grötker, Der Gläserne Patient, in Handelsblatt Online vom 26.11.2014, online: http://www.handelsblatt.com/technologie/forschung-medizin/medizin/der-glaeserne-patient-du-musst-dein-leben-aendern/11030186.html.

persönlichen Charakter aufweisen und Rückschlüsse auf den sensiblen Bereich der Gesundheitsdaten ermöglichen. Gleichzeitig stehen eine Vielzahl neuer Produkte in den Startlöchern, so zum Beispiel die „Apple Watch“, mit der man die Zeit ablesen, telefonieren, SMS schreiben, aber auch seinen Herzschlag und sein Bewegungsprofil aufzeichnen kann. Die Mitbewerber von Samsung, Sony und LG sind schon mit Android-basierten Smartwatches auf dem Markt, Motorola und Apple folgen 2015. Welche Anwendungen möglich sind, ist noch nicht absehbar. Über Potenziale und Risiken muss noch diskutiert werden. Ebenso wenig ist absehbar, welche neuen Verfahren zur Auswertung der neuen Datenmassen entstehen werden.

Neben Smartwatches und Fitnesstrackern beschäftigten sich die Medien in letzter Zeit vermehrt mit Geräten zur Protokollierung des Fahrverhaltens von Autofahrern. Auch hier stellt sich die Frage, ob die umfassenden Datensammlungen, die dadurch entstehen, nicht zu sehr in das Persönlichkeitsrecht eingreifen. Die Geräte können detaillierte Bewegungsprofile herstellen, zeichnen Geschwindigkeit und Fahrverhalten auf, optimieren den Kraftstoffverbrauch und so weiter. Einige Versicherungen erwägen, auf der Grundlage solcher Daten teilnehmenden Autofahrern abhängig von ihrem Fahrverhalten günstigere Versicherungsprämien einzuräumen. Den datenschutzrechtlichen Bedenken wollen die Hersteller durch Privacy-by-Design begegnen. Durch technische Modellierung der Verfahren soll verhindert werden, dass detailliertere persönliche Daten entstehen, als für die Auswertung nötig sind. Die Daten sind aber auch für andere Analysen interessant: Die Diskussion über geschlechtsspezifische Unterschiede im Fahrverhalten könnten auf der Basis konkreter Daten geführt werden, eine Verknüpfung mit der Pulsmessung durch Fitnesstracker könnte Hinweise geben, wann der Fahrer oder die Fahrerin sich eine Pause gönnen sollte oder ähnliches.

3. Trends und Grundlagen

„Big Data“ ist vielschichtig, es gibt keine einheitliche, feste Definition für diesen Begriff. Um ihn und die Implikationen zu (be-)greifen, ist es notwendig, die mit dem Begriff verknüpften technischen Trends im Blick zu haben. Neben der reinen Datenmenge bezieht sich der Begriff typischerweise auf die „drei Vs“: Volumen, Velocity (Geschwindigkeit) und Varietät, der hier um die Eigenschaft der Verfügbarkeit ergänzt wird. Schließlich spielt regelmäßig Wahrscheinlichkeitsrechnung in verschiedenen Formen, zum Beispiel als selbstlernende Verfahren („machine learning“), eine große Rolle.

Datenanalysen kommen in vielen Anwendungsfeldern zum Einsatz. Nicht minder vielfältig sind die Verfahren, die dabei verwendet werden: „Query and Reporting“,[15] „Data-Mining“,[16]

[15] Deutsch: „Suchen und Berichten“.

[16] Der Begriff wird uneinheitlich verwendet, vgl. auch Fn. 21. Man kann hierunter das gezielte Auffinden einzelner Informationen in größeren Datenbeständen verstehen, wobei hier analytische Verfahren zum Einsatz kommen können, die bestehende Informationen verknüpfen.

Datenvisualisierung, Vorhersagemodelle und Prognosen, Optimierung (vor allem von Prozessen), Simulation, Integration verschiedener Datenformate (etwa bei Sprach- und Bilderkennung), Geodaten- und raumbezogene Analysen sind nur einige.

3.1. „Volume“ und Parallelisierung

Grundsätzlich sind auch sehr große Datenmengen mit aktuellen technischen Architekturen in überschaubarer Zeit verarbeitbar. Es stehen inzwischen alternative Ansätze zu traditionellen Datenbankmodellen zur Verfügung, die es ermöglichen, die Verarbeitung von Daten auf parallel arbeitenden Computern zu koordinieren. Damit können Datenmengen jenseits von mehreren Gigabyte verarbeitet werden – zu erschwinglichen Tarifen.

Der Suchindex von Google – weltweit die meistgenutzte Suchmaschine – hat derzeit einen Umfang von 100 Petabyte (oder 100 Millionen Gigabyte).[17] Er wird kontinuierlich fortgeschrieben, indem immer neue Webseiten indiziert und gespeichert werden. Die Milliarden täglichen Suchanfragen[18] zu beantworten, kann nur funktionieren, wenn sie gleichzeitig auf parallelgeschalteten Computern bearbeitet werden. Diese Parallelisierung ist weder technisch trivial, noch war sie immer selbstverständlich. Sie hat erst in den letzten Jahren an Popularität gewonnen und ist inzwischen im Rahmen von Cloud-basierten Infrastrukturen leicht nutzbar.

3.2. „Velocity“ – Just-in-time-Verarbeitung

Die Zeit, die für die Verarbeitung großer Datenbestände notwendig ist, stellt nach wie vor einen limitierenden Faktor dar, weil ein Ergebnis nur dann hilfreich ist, wenn es sehr schnell verfügbar ist („just in time“). Es kommt also auf den Einzelfall an, ob bestimmte Verarbeitungen möglich und sinnvoll sind.

Soll neben der Beantwortung einer Suchanfrage gleichzeitig die passende Werbung eingeblendet werden und auf das aktuelle Surfverhalten abgestimmt sein, stellt dies hohe Anforderungen an die Geschwindigkeit der Verarbeitung. Erfolgt die Auswahl der richtigen Werbung nicht rechtzeitig für das Sucherergebnis, kommt sie zu spät.[19] Verzögert sich die Beantwortung der Suchanfrage jenseits gewisser hinnehmbarer Grenzen, besteht die Gefahr, dass die Suchmaschine Marktanteile verliert. Mit der Verbreitung von Real-Time-Bidding-Systemen, die es ermöglichen, automatisiert im Augenblick der getätigten Suchanfrage für die Platzierung einer Anzeige ein Gebot abzugeben, gibt es einen erheblichen

---

[17] Google publiziert hierzu nicht regelmäßig. Die genannte Zahl wird in 2010 und 2012 erwähnt: http://googleblog.blogspot.de/2010/06/our-new-search-index-caffeine.html und hier: http://googleforwork.blogspot.de/2012/07/introducing-google-cloud-platform.html.

[18] Derzeit sind es ca. 3,5 Milliarden pro Tag, vgl. http://www.internetlivestats.com/google-search-statistics/.

[19] Anschaulich illustriert in dem Video von Matt Cutts, Software-Ingenieur bei Google, How Search Works, https://www.youtube.com/watch?v=BNHR6IQJGZs.

Bedarf für Systeme, die den Wert eines Gebots schnell auf einer großen Datenbasis errechnen können.

3.3. „Variety“ – Unstrukturierte Daten

Eine weitere Herausforderung für die Verarbeitung stellt nach wie vor die Struktur der vorhandenen Daten dar. Eine Datenbank enthält neben dem Inhalt der Daten stets Felder, die durch einen Feldtyp näher beschrieben sind, beispielsweise Vorname, Nachname, Geburtsdatum.

Den Großteil aller vorhandenen Datenbestände wird man als unstrukturiert bezeichnen können. Ein Prosatext beispielsweise weist in der Regel keine annähernd exakten Strukturen auf; dazu ist die Grammatik und Semantik natürlicher Sprachen zu komplex, flexibel und mehrdeutig. Schwierigkeiten für die Analyse können entstehen, wenn Daten aus unterschiedlichen Quellen mit zwar vorhandenen, aber nicht übereinstimmenden Strukturmerkmalen auftauchen. Die Übergänge sind jedoch fließend, weshalb die Verwendung des Begriffs „unstrukturierte Daten“ gelegentlich als unpräzise bezeichnet wird. Für den Umgang mit unstrukturierten Daten existiert eine Vielzahl von Verfahren, von denen einige im weiteren Verlauf exemplarisch näher beleuchtet werden sollen.

3.4. Verfügbarkeitsanforderungen als Treiber für Innovation

Die Verfügbarkeit von Datenbeständen spielt für die möglichen Verarbeitungen eine zentrale Rolle. Diese Frage gerät in den typischen kommerziellen Einsatzszenarien leicht aus dem Blickfeld, da Unternehmen meist vom eigenen Datenbestand ausgehen und gegebenenfalls zukaufbare Datenbestände heranziehen und nach Auswertungswegen und Erkenntnismöglichkeiten suchen, wenn sie überlegen, wo ihnen Big Data helfen kann.

Es ist aber auch der umgekehrte Weg denkbar, indem man sich der Frage nähert, in welchen Bereichen man Erkenntnisse aus Datenbeständen gewinnen möchte und anschließend nach Möglichkeiten sucht, diese Datenbestände aufzubauen.

Neben Startups, deren Geschäftsmodelle direkt auf Datenanalyse begründet sind, ist dieser Blickwinkel auch für große Unternehmen relevant, die ihre strategische Geschäftsentwicklung und Akquise zunehmend daran ausrichten, welche Datenbestände hierdurch erschließbar werden. Die starken Bewertungen und Kaufpreise von Firmen mit solchen Datenbeständen oder Potenzialen in diesen Bereichen sprechen eine deutliche Sprache. Beispiele hierfür sind offenkundig die Unternehmen Facebook und Google selber, aber auch die schon genannte Firma Nest. Nicht abwegig ist wohl die Überlegung, dass Google sich nach der Entwicklung seiner Kernkompetenz im Bereich der Internetsuche gezielt auf weitere Anwendungsfelder wie E-Mail, Soziale Netzwerke und Kartographie bewegt hat, um damit Zugang zu Datenbeständen zu gewinnen, die mit den bestehenden korreliert werden können. Über die genauen Verknüpfungen dieser Datenbestände und die ange-

wendeten Verfahren ist indes wenig bekannt. Dies unterfällt dem Schutz der Geschäftsgeheimnisse des Unternehmens.[20]

Innovationstreiber ist demnach zunehmend, aus der Big-Data-Analyse Erkenntnisse zu gewinnen, die dann als Geschäftsgeheimnis einen mitunter relevanten Teil des Marktwertes des Unternehmens ausmachen. Es stellt sich allerdings die Frage, ob dies ein guter Innovationsanreiz ist, oder ob er zu intransparenten und mitunter sogar marktverzerrenden Geschäftsmodellen führt.

3.5. Wahrscheinlichkeiten und Machine Learning

Eine wichtige Grundlage zum Verständnis von Datenanalysen ist in vielen Fällen die Wahrscheinlichkeitsrechnung. Nahezu allen Verfahren ist eine gewisse Verwurzelung im Methodenkanon der Stochastik gemein, wobei der Wahrscheinlichkeitsrechnung besondere Bedeutung zukommt.

In manchen Bereichen sind Wahrscheinlichkeitserwägungen allerdings nur am Rande relevant oder spielen gar keine Rolle. Zu denken ist etwa an Datenvisualisierungen oder auch konventionelle Anfragen, in deren Rahmen es nur darum geht, eine bestimmte Information in einem großen Datenbestand auffindbar zu machen („Query-Verfahren“).[21] Allerdings können hier auch stochastische Methoden integriert werden. Für die Suche nach einzelnen Informationen in unstrukturierten Datenbeständen kann man über Wahrscheinlichkeiten etwa die Anzahl der näher zu untersuchenden Datenbestände minimieren.

Ein Grundprinzip aller Wahrscheinlichkeitsangaben ist, dass sie keine Aussagen über den Einzelfall zulassen, sondern nur Prognosen im Rahmen von (bei korrekter Anwendung klar definierten) Wahrscheinlichkeiten ermöglichen. Dabei ist der Begriff der Korrelation zentral. Eine Korrelation beschreibt die Abhängigkeit einer Größe von einer anderen. Ist die Korrelation hoch, kann man (wiederum im Rahmen der Wahrscheinlichkeiten) vom einen Wert eine Aussage für den anderen ableiten.

Ein klassisches Beispiel für einfache Datenanalysen ist die Prozessoptimierung. In wiederkehrenden Prozessen können Daten über Produktionsfehler erfasst werden. Findet sich in den Daten eine Korrelation der Produktionsfehler mit der Tageszeit, ergibt das einen Anhaltspunkt für eine Optimierung. Um ein ganz einfaches Beispiel zu nennen: So kann eine Korrelation darauf hindeuten, dass Mitarbeiter am Fließband im Idealfall nach drei Stunden eine Pause machen müssen oder dass ein Industrieroboter nach 48 Stunden

---

[20] Insbesondere im Fall des schon 2007 erfolgten Zukaufs des Werbedienstleisters DoubleClick ist die Nähe zum Kerngeschäft offensichtlich, eine Verknüpfung der Datenbestände liegt nahe. Google hat zwar versichert, dass eine Verknüpfung auf der Ebene der personenbezogenen Profile im Rahmen der Übernahme nicht erfolgt. Das heißt aber nicht unbedingt, dass abstrahierte Datensätze, die wegen einer erfolgten Aggregation keinen Personenbezug aufweisen, nicht miteinander kombiniert wurden.

[21] Derartige Suchen nach der sprichwörtlichen Nadel im Heuhaufen könnte man begrifflich auch dem Data-Mining zuordnen. Tatsächlich werden die Begriffe Data-Mining und Big Data beide nicht streng definiert und nicht selten nahezu synonym verwendet.

gewartet und mit neuem Schmieröl versorgt werden muss. Korrelationen über die Zeit spielen auch in der Verkehrsplanung eine Rolle. Belastungssituationen ergeben sich zu bestimmten Stoßzeiten. Um den Verkehrsfluss zu optimieren, werden Ampelschaltungen und Fahrpläne, aber auch die Verkehrsplanung und der Infrastrukturausbau auf der Basis von Verkehrszählungen verbessert.

Zunehmend werden bei der Datenanalyse auch selbstlernende Verfahren eingesetzt, die aus der Künstliche-Intelligenz-Forschung stammen. Mustererkennungsverfahren oder maschinelles Lernen mittels neuronaler Netze oder vereinfachter Ansätze ermöglichen das Auffinden von regelhaften Strukturen in Datenbeständen. Auch diese Verfahren fußen auf statistischen Berechnungen.

Ein beliebtes Beispiel für den Einsatz von Big-Data-Analysen sind automatisierte Übersetzungen. Für diesen Bereich haben sich lernende Mechanismen den Ansätzen als überlegen gezeigt, deren Regeln vorher determiniert sind. Nahezu alle automatischen Übersetzungsdienste bauen darauf auf, durch die Analyse eines größeren Textkorpus Regeln zu gewinnen, die sie dann auf neue Texte, deren Übersetzung nicht bekannt ist, anwenden. Das Faszinierende an diesen Verfahren ist, dass es für das System weitgehend egal ist, für welche Sprachen es trainiert wird. Entscheidender ist die Masse (und Qualität) des Datenbestandes, aus dem „gelernt“ wird. Frühere Ansätze, die auf bestehenden grammatikalischen Regeln und Wörterbüchern aufgebaut waren, sind weniger erfolgreich gewesen. Ähnliche Verfahren werden nun zur Optimierung der Spracherkennung eingesetzt.[22]

4. Potentiale und Herausforderungen

Big Data bietet enorme Potentiale, weist aber auch erheblichen gesellschaftlichen und – daraus folgend – politischen und rechtlichen Klärungsbedarf auf. Bestimmend für die wachsende Bedeutung von Datenanalysen ist die der Digitalisierung innewohnende Verdatung nahezu aller Lebensbereiche.

Sowohl Potentiale als auch Probleme stellen sich den verschiedenen Akteursgruppen mit jeweils spezifischen Handlungsoptionen. Die Potentiale für die einzelnen Akteure sind vielfältig, ebenso wie die Herausforderungen, aus denen im Nachfolgenden fünf zentrale herausgehoben werden sollen.

Durch die Verfügbarkeit umfassender Datenbestände ergeben sich so beispielsweise für staatliche Akteure neue Möglichkeiten, ihre Entscheidungen präziser an den gesellschaftlichen Bedürfnissen auszurichten. Diese Potentiale werden bisher nur zögerlich genutzt. In der Stadtplanung wird zwar in nahezu jedem Politikbereich auf Empirie zurückgegriffen. Die Frage ist aber, ob durch eine stärkere Verdatung nicht ganz neue Qualitäten politischer Gestaltung in einer Vielzahl von Feldern erreicht werden können.

---

[22] Zum Einsatz von neuronalen Netzen bei der Spracherkennung siehe etwa: http://www.forbes.com/sites/roberthof/2013/05/01/meet-the-guy-who-helped-google-beat-apples-siri/.

Die Nutzung von Big Data gehört in vielen Wissenschaftsdisziplinen zum grundlegenden Handwerk. Wissenschaftler sind – wenn man so will – die natürlichen Early Adopter. Allerdings ist die Bandbreite erheblich: Während in der experimentellen Elementarteilchenphysik, wie sie am Genfer Institut für Elementarteilchenphysik CERN praktiziert wird, Big Data zum Kerngeschäft gehört, ist in der Rechtswissenschaft die Datenanalyse jenseits der Kriminalistik bisher kaum ein Thema. Es ist aber davon auszugehen, dass auch in datenfernen Disziplinen über Datenanalysen interessante Erkenntnisse gewonnen werden können. Entsprechend werden in vielen Bereichen neue Wege beschritten. So versucht das vom Bildungsministerium geförderte Projekt „Argumentum“ eine automatisierte Analyse von Argumentationsstrukturen in Gerichtsurteilen zu entwickeln.[23] Aber schon eine – weitaus einfachere – Analyse von Gesetzesverweisen in Urteilen und anderen Rechtstexten könnte Aufschluss darüber geben, welche Normen praktisch relevant sind und so effektiv zu einer Verschlankung des Rechtsbestandes führen.

Erhebliche Potentiale bieten sich in der empirischen Sozialforschung. Dieses Gebiet, dessen Primärdaten bisher in nicht unerheblichem Maße aus Umfragen gespeist waren, kann unter Zugriff auf Beobachtungsdaten, wie sie im Internetzeitalter entstehen, völlig neuen Wirkzusammenhängen nachgehen. Zentrales Problem in diesem, wie auch in anliegenden Feldern wie der Psychologie und sicherlich auch der Medizin, ist jedoch die Verfügbarkeit des Datenmaterials für wissenschaftliche Forschungszwecke.

Die Möglichkeiten der wirtschaftlichen Nutzung von Big Data werden zunehmend erschlossen. Nach einer repräsentativen Studie des Branchenverbands BITKOM werten 9 von 10 Unternehmen grundsätzlich IT-gestützt Daten für ihre Entscheidungsprozesse aus; 46 Prozent der Unternehmen setzten 2013 dafür bereits spezielle Analysetools ein; weitere 36 Prozent beabsichtigten 2014 Maßnahmen in diesem Bereich einzuführen.[24] Nachteile ergeben sich mitunter für kleinere Unternehmen, die nicht über geeignete Datenbestände verfügen. Große Anbieter haben daher eine bessere Ausgangsposition für Optimierungen ihrer Geschäftsmodelle.

Hier sind allerdings in einigen Bereichen auch Disruptionen bestehender Märkte denkbar. So verfügen Landwirte beispielsweise nicht über hinreichende Informationen über ihre Endkunden, da sie ihre Produkte meist über Zwischenhändler vertreiben. Diese Zwischenhändler profitieren von ihrem besseren Marktwissen und können so mitunter erhebliche Preisaufschläge rechtfertigen. Denkbar ist, dass sich Landwirte – zumindest in einigen spezialisierten Märkten – über Vertriebsplattformen zusammenschließen, um diese Informationen zu sammeln und den Vertrieb weitgehend selbst zu organisieren. Die Datenor-

[23] Vgl. http://argumentum.eear.eu.

[24] Vgl. BITKOM (Hrsg.): Potenziale und Einsatz von Big Data – Ergebnisse einer repräsentativen Befragung von Unternehmen in Deutschland, Berlin, 5.5.2014, online: http://www.bitkom.org/files/documents/Studienbericht_Big_Data_in_deutschen_Unternehmen.pdf, wobei sämtliche Angaben sich auf Unternehmen mit einer Größe von mehr als 50 Mitarbeitern beschränken.

ganisation, die früher eine Domäne des Vertriebs war, fällt damit zurück in die Hände der Erzeuger.[25]

Auch aus individueller Perspektive ist Big Data zunehmend ein Thema. Die bereits erwähnte „Quantified Self"-Bewegung, die sich der quantifizierbaren Selbsterfassung verschrieben hat, findet zunehmend in die Alltagstechnologie Einzug. Über Pulsmessgeräte mit digitaler Schnittstelle und Schrittzähler wird für Hobbysportler eine Trainingsoptimierung möglich, wie sie früher nur Leistungssportlern zugänglich war. GPS-Systeme in Smartphones machen eine detaillierte Auswertung der eigenen Wege möglich. Es ist davon auszugehen, dass zukünftig eine Vielzahl von Informationen des Alltags sensorisch erfasst wird und damit detailliert Aufschluss etwa über Ernährungsgewohnheiten, Gesundheitsentwicklung oder über Lernfortschritte in der Schule geben kann.

Viele dieser Geräte sind jedoch an spezialisierte Dienste gebunden und übermitteln die erfassten Daten umfassend an den Hersteller oder Diensteanbieter. Hierdurch wird es dem Nutzer erleichtert, den Austausch und Vergleich mit anderen zu suchen. Welche Daten übermittelt werden, ist allerdings nicht immer transparent. Bei manchen Daten, zum Beispiel gesundheitsrelevanten, ist eine Übermittlung aufgrund ihrer Sensitivität nicht wünschenswert. Darüber hinaus bestehen Probleme bezüglich der Profilbildung und Manipulation.

4.1. Erkenntnisgewinn, aber: Korrelation ist nicht Kausalität

Eine gute Datenanalyse birgt die Chance auf Erkenntnisse, die sich für Wirtschaft, Wissenschaft und politische Gestaltung nutzbar machen lassen. Doch Korrelation bedeutet nicht Kausalität: Nur weil man Abhängigkeiten zwischen zwei Größen erkennt, heißt das nicht, dass die eine die andere verursacht. Es kann auch genau umgekehrt sein oder beide Größen hängen gleichermaßen von einer Dritten ab. Und schließlich kann es auch reiner Zufall sein: Die Größen sind unabhängig voneinander. Das klassische Beispiel hierfür ist die Korrelation zwischen dem Bestand an Störchen und der Geburtenrate in einem Dorf.[26] Selbst wenn jedes Mal, wenn ein Storch ins Dorf kommt, auch ein Kind geboren wird, heißt das noch nicht zwingend, dass die Störche die Kinder bringen.

Allerdings gilt auch: ohne Korrelation keine Kausalität. Kommt ein Kind zur Welt, ohne dass ein Storch im Dorf war, ist damit jedenfalls belegt, dass nicht (nur) Störche Kinder bringen, sondern Kinder auch ohne Störche zur Welt kommen können. Eine nicht ganz unwichtige Erkenntnis.

---

[25] Diesen Weg geht etwa das kanadische Startup „Provender", vgl. http://www.theglobeandmail.com/report-on-business/small-business/sb-growth/day-to-day/farms-factories-and-film-sets-startups-bring-big-data-to-inefficient-industries/article20574000/.

[26] Siehe z.B. „Storch und Mensch", http://www.zeit.de/2006/25/Stimmt-s_P-25_xml.

Die Korrelation bietet durchaus eine gute Grundlage für die Suche nach tatsächlichen Sachzusammenhängen, insbesondere aber zur Falsifikation von angenommen Sachzusammenhängen.

Das Problem aktueller Big-Data-Analysen ist jedoch, dass sie so viele verschiedener Faktoren gleichzeitig in eine Korrelationsanalyse einbeziehen, dass einzelne, mögliche Ursachenzusammenhänge nicht mehr erkennbar sind. Ergebnis: Man weiß, dass unter bestimmten Bedingungen mit einer hohen Wahrscheinlichkeit ein bestimmtes Ereignis eintritt, hat aber keine Ahnung wieso.

4.2. Vermeintliche Vollerhebungen und blinde Flecken

War man vor der Digitalisierung auf Umfragen und Zählungen angewiesen, um Datenmaterial für Analysen zu erhalten, entstehen heute diese Datenbestände in ungleich höherer Quantität und zumeist auch Qualität quasi nebenbei.

Ein relevanter Teil traditioneller empirischer Markforschung erfolgt auf der Basis von Umfragen. Auf diese Weise versucht man zum Beispiel, die Effektivität einer Werbemaßnahme zu erfassen und gegebenenfalls anzupassen. Die gesammelten Aussagen sind allerdings nur dann aussagekräftig, wenn es gelingt, eine repräsentative Auswahl von Befragten zu bestimmen und zur Teilnahme zu bewegen. Selbst dann bleiben die Aussagen mit einer gewissen Ungenauigkeit behaftet. Verzerrungen ergeben sich dadurch, dass Befragte im Einzelfall absichtlich oder unabsichtlich falsche Angaben machen. Bezieht jedoch ein Datenbestand nicht nur einen Teil, sondern alle (potentiellen) Kunden ein und baut er zudem auf der direkten Verhaltensbeobachtung auf, ist er von erheblich höherer Qualität. Dies geschieht beispielweise beim Online-Marketing. Hierbei wird das Nutzungsverhalten detailliert aufgezeichnet. Firmen, die auf Online-Marketing spezialisiert sind, können zuordnen, welche Internetseiten ein Nutzer besucht hat. Die Spuren, die ein Nutzer hinterlässt, geben Aufschluss darüber, wer welchen Werbemitteln ausgesetzt war und wo diese in eine Kaufentscheidung münden.

Eine gute Datenanalyse muss sich aber der blinden Flecken in den Datenbeständen gegenwärtig sein. Prognoseaussagen können immer *nur* bezogen auf die Grundgesamtheit getroffen werden. Eine Prognose, die auf der Verhaltensbeobachtung von Nutzern bestimmter sozialer Netzwerke aufbaut, ist – selbst wenn das soziale Netzwerk erhebliche Größe hat – nicht auf den Rest einer Gesellschaft übertragbar. Große Datenbestände laden mitunter zu unsachgemäßen Verallgemeinerungen ein. Es ist sehr verlockend, auf die Datenbestände Verfahren anzuwenden, die eigentlich nicht zu richtigen Ergebnissen führen. Gerade aufgrund der Komplexität der Verfahren sind solche Fehler nicht immer leicht nachzuvollziehen. Baut man Entscheidungen gesamtgesellschaftlicher Bedeutung auf derartigen Analysen auf, läuft man Gefahr, großen Bevölkerungsgruppen nicht gerecht zu werden.

4.3. Exklusives Wissen und Marktversagen

Erkenntnisgewinn grundlegender Art war bislang die Domäne der Wissenschaft. Aufgrund ihres Charakters orientiert sich Wissenschaft an dem Grundsatz, diese Erkenntnisse nicht exklusiv zu behandeln, sondern zu publizieren. Dies ändert sich, wenn die Grundlagen für den Erkenntnisgewinn exklusiv in privater Hand sind.

Für Marktakteure bieten Datenanalysen eine Vielzahl von Optimierungsmöglichkeiten. Gleichzeitig können sie – sofern die Daten nur exklusiv verfügbar sind – zu erheblichen Verzerrungen führen und wünschenswerte Disruptionen und Innovationen aufhalten. Die Markteintrittsbarriere für Suchmaschinen ist nicht deswegen so hoch, weil es technisch aufwendig ist, den Bestand an Internetseiten zu erfassen und zu indizieren, sondern weil es an der Datenbasis fehlt, die Seiten danach zu gewichten, welche besonders stark frequentiert werden. Die Überlegung mag in verschärfter Weise für die Entwicklung eines nicht nur inhaltlich guten, sondern auch rentablen Geschäftsmodells gelten, da der Markt bereits stark auf – wiederum exklusive – Datenbestände der Werbevermarktung optimiert ist.

4.4. Umfassende Profilbildung

Für Individuen bieten die genannten Möglichkeiten oft positive Effekte. Nutzer profitieren von Erkenntnissen und Produktoptimierungen, weil die Produkte dadurch preiswerter angeboten werden können. Jedoch gibt es auch Gefahren: Offenkundig ist das Problem der Profilbildung, die sich aus der umfassenden Digitalisierung ergibt. Datensammler können ihre Datenbestände, etwa über das Surfverhalten, auf einen individuellen Nutzer zurückführen. Über Korrelationen können sie dann Wahrscheinlichkeitsaussagen zu Fragen treffen, die den Nutzer in sensiblen Bereichen betreffen, ohne dass er dazu jemals direkt Auskunft gegeben hat. Es ist sogar denkbar, dass aus den Daten auf Informationen geschlossen wird, die darüber hinausgehen, was das Individuums über sich selbst weiß, etwa wenn auf Grundlage des Verhaltens in sozialen Netzwerken Aussagen über das Vorliegen von unerkannten Krankheiten oder von sexuellen Vorlieben extrapoliert werden. Studien deuten darauf hin, dass im Einzelfall durchaus Aussagen von erheblicher Genauigkeit möglich sind. So wurde nachgewiesen, dass in sozialen Netzwerken signifikante Aussagen über die sexuelle Orientierung einzelner Mitglieder gemacht werden können. Dabei wurden die Kontakte der Mitglieder ausgewertet (eine sogenannte Netzwerkanalyse). Auch wenn das betreffende Mitglied keine Auskunft über seine eigene sexuelle Orientierung abgegeben hat, konnte gestützt auf die Angaben der jeweiligen Kontakte ein Rückschluss darauf gezogen werden.[27]

Eine ähnliche Untersuchung gibt es auch zur Depressionsneigung. Hierbei wurde das Kommunikationsverhalten von Personen, deren Depressionserkrankung bekannt ist, mit

[27] Jernigan, Carter/Mistree, Behram FT: Gaydar: Facebook friendships expose sexual orientation, First Monday 14.10 (2009), online: http://firstmonday.org/article/view/2611/2302.

dem von Menschen korreliert, bei denen die Krankheit nicht diagnostiziert worden war. Wenn bestimmte Charakteristiken, etwa die Verringerung sozialer Aktivität, eine negative Grundstimmung in den Inhalten der Kommunikation oder stärkere Religiosität auftauchen, lieferte dies Anhaltspunkte für die Depressionswahrscheinlichkeit.[28] Die Untersuchung zeigt, dass eine Aussagegenauigkeit von etwa 70 Prozent möglich war, wenn die Eigenschaften an Personen getestet wurden, bei denen die Erkrankung bekannt war, aber diese Tatsache erst nach der Überprüfung berücksichtigt wurde.

Hinzu tritt eine weitere Gefahr: Aussagen sind nicht nur über die Mitglieder eines sozialen Netzwerkes möglich, sondern auch über Dritte, die nicht Mitglied sind oder je waren. Eine Studie von Wissenschaftlern an der ETH Zürich zeigt, dass Netzwerkanalysen über sexuelle Vorlieben auch Aussagen über Externe zulassen.[29]

4.5. Manipulation der Meinungsbildung

Die Gefahr der Manipulation auf der Basis von Verhaltensprognosemöglichkeiten, die einzelnen Organisationen exklusiv vorliegen, betrifft nicht nur das Individuum. Der Präsidentschaftswahlkampf von Barack Obama im Jahr 2012 war stark auf die Analyse von Daten sozialer Netzwerke in den entscheidenden Swing-States gestützt. Es spricht einiges dafür, dass diese Herangehensweise einen entscheidenden Beitrag zu seinem späteren (knappen) Wahlerfolg geleistet hat. Aufgrund tiefgehender Datenanalysen konnte sich das Team von Barack Obama effektiv auf besonders relevante Wählerschichten konzentrieren und die Themen identifizieren, mit denen diese Gruppen überzeugt werden konnten.[30]

Steht derartiges Wissen nur einzelnen Kandidaten zur Verfügung, gefährdet das die Chancengleichheit verschiedener politischer Gruppen. Das wiederum gefährdet die Demokratie und damit die Grundlage des gesellschaftlichen Zusammenhalts. Haben die Bürger Grund für den Verdacht, dass Meinungsbildung nicht frei stattfindet, sondern gezielt gesteuert wird, schwindet das Vertrauen in Demokratie. Damit droht ihre Bereitschaft zu sinken, Einschränkungen individueller Freiheit und Vorlieben zugunsten einer (aus ihrer Sicht bis dato noch unterstellten) gesellschaftlichen Mehrheit zu akzeptieren.

5. Zu diskutierende Fragestellungen

Überall dort, wo Probleme und Interessenkonflikte zutage treten, besteht die Notwendigkeit von Aushandlungsprozessen. Diese können entweder auf der Basis konkreter Szena-

---

[28] De Choudhury, Munmun, et al.: Predicting Depression via Social Media, ICWSM, 2013, online: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewFile/6124/6351.

[29] Sarigöl, Emre/Garcia, David/Schweitzer, Frank: Online Privacy as a Collective Phenomenon, CoRR, arXiv:1409.6197, online: http://arxiv.org/abs/1409.6197.

[30] Issenberg, Sasha: How President Obama's campaign used big data to rally individual voters. MIT Technology Review, 19.12.2012, online: http://www.technologyreview.com/featuredstory/509026/how-obamas-team-used-big-data-to-rally-voters/.

rien geschehen, die dann abstrahiert auf ähnliche Fragestellungen angewendet werden, oder die Gesellschaft einigt sich auf abstrakter Ebene über die Grundsätze, die wichtig genug sind, um in die Aushandlungsprozesse einzufließen.

Da aber die Einsatzgebiete von Big Data so unterschiedlich sind, ist es sinnvoll, im Augenblick auf der Ebene der konkreten Einsatzgebiete zu bleiben. Es lassen sich aber zumindest strukturelle Überlegungen anstellen, die aufzeigen, wie Aushandlungsprozesse gestaltet werden können. Derzeit herrschen allerdings noch grundlegende Aufgabenstellungen vor, die Räume für individuelle Aushandlungsprozesse stark einschränken. Hier bedarf es zunächst einer grundsätzlichen gesellschaftlichen Einigung.

Im Rahmen des Projekts „Braucht Deutschland einen Digitalen Kodex?“ ergeben sich somit mehrere mögliche Ansatzpunkte. Dabei empfiehlt sich eine Verknüpfung abstrakter und konkreter Fragen. Letztere sind notwendig, weil – wie dargestellt wurde – sich viele Fragen nur und erst im konkreten Einsatzszenario stellen. Dennoch ist zu vermuten, dass sich auch auf der Metaebene grundsätzlichere Erkenntnisse gewinnen lassen. Hierbei könnte es hilfreich sein, neben der grundsätzlichen Diskussion auch einem „Bottom-up“-Zugriff Raum zu gewähren und aus der Diskussion einzelner Big-Data-Einsatzszenarien Abstraktionen zu ermöglichen.

Bezüglich der konkreten Einsatzszenarien, die zu erörtern sind, sind diejenigen vielversprechend, die bisher nicht Gegenstand einer umfassenden Diskussion waren, weil die technischen und marktbetreffenden Entwicklungen noch recht jung sind, also etwa der Bereich der „Smart Watches“ und Selbsttracker. Daneben könnte man erwägen, die Möglichkeiten umfassender Datensammlungen bei staatlichen Diensten oder den großen wirtschaftlichen Akteuren ins Blickfeld zu nehmen. Eine ganze Reihe gesellschaftlicher Herausforderungen stellen sich nur bei diesen umfassenden Sammlungen, die insoweit eine eigene Qualität haben.

Für die Diskussion von verallgemeinerbaren Überlegungen sollen die folgenden Abschnitte einen ersten Ansatzpunkt liefern. Sie stellen die rechtlichen Instrumentarien dar inklusive der Beschränktheit derzeitiger Regelungsansätze im Rahmen des Datenschutzrechtes (Abschnitt 5.1), und beleuchten schließlich allgemeine, prozedurale und steuernde Aspekte für einzelne Verfahren (Abschnitt 5.2). Diese beiden Abschnitte beleuchten das Problem gleichsam aus zwei Perspektiven, zunächst „top down“ und dann „bottom up“. Es ist davon auszugehen, dass über die Betrachtung einzelner Verfahren abstrakte Verallgemeinerungen gefunden werden können. Das Projekt könnte diese Erwägungen zum Ausgangspunkt für eine Diskussion nehmen und im Laufe des Prozesses weiterentwickeln.

Neben der Frage nach einem Digitalen Kodex für den privaten Bereich stellt sich naturgemäß auch die Frage der Regulierung bei staatlicher Nutzung von Big Data. Dieser Bereich wird jedoch für den folgenden Prozess zunächst ausgeklammert.

5.1. Selbstregulierung durch den Markt?

In einem Rechtsstaat wie Deutschland, der auf allgemeiner Handlungsfreiheit und Privatautonomie aufgebaut ist, ist bei Interessenkonflikten zunächst einer direkten Aushandlung zwischen den beteiligten Parteien der Vorzug zu geben. Staatliche Eingriffe in diese Aushandlungen gehen stets mit einem Eingriff in die Handlungsfreiheit einher und bedürfen daher einer Rechtfertigung.

Dieses Primat privater Aushandlungsprozesse kann sich nicht nur auf diese grundrechtlichen Erwägungen stützen. Die oben beschriebene Vielzahl von einzelnen Anwendungsfällen würde zu einer extrem kleinteiligen Regulierung führen, die aus Sicht der rechtsschöpfenden Politik nicht oder nur mit hohem Aufwand kompetent zu bewerkstelligen ist.

Staatliches Handeln ist jedoch dort geboten, wo die privaten Aushandlungsprozesse versagen, etwa weil sich ungleiche Parteien begegnen oder weil schutzwürdige Belange Dritter bzw. der Allgemeinheit betroffen sind.

Das Datenschutzrecht trägt der allgemeinen Handlungsfreiheit – insoweit es um den Schutz der informationellen Selbstbestimmung geht – dadurch Rechnung, dass es Datenverarbeitungen personenbezogener Daten dann gestattet, wenn diese für Vertragszwecke der beteiligten Parteien erforderlich sind. Dort, wo es nicht um Verträge geht, können die von der Verarbeitung Betroffenen explizit einwilligen. Das Ungleichgewicht zwischen unter Umständen sehr großen Diensteanbietern und Konsumenten berücksichtigt das Datenschutzrecht dadurch, dass es den Diensteanbietern Transparenzverpflichtungen auferlegt, etwa dass sie die Daten nur für die Zwecke verarbeiten dürfen, die beide Parteien miteinander vereinbaren.

Big Data-Ansätze zeigen aber deutlich die Beschränktheit dieses Modells auf, das auf den übereinstimmenden Willen der beteiligten Parteien abstellt. Wie wir schon dargelegt haben, sind von der Verarbeitung gegebenenfalls Personen betroffen, die ihre Daten gar nicht oder nur im geringeren Umfang zur Verfügung gestellt haben. So gelangt der eingangs zitierte Eben Moglen zu der Erkenntnis, die Privatsphäre als „nicht transaktional“ zu charakterisieren, was heißt, dass sie keine abgeschlossene Austauschbeziehung zwischen zwei Parteien ist. In Fällen der Betroffenheit Dritter scheitert die Privatautonomie. Hier besteht ein Problem, an das sich der Gesetzgeber in Deutschland bisher noch nicht herangewagt hat.

Hinzu tritt hier ein weiteres Problem: Der Anwendungsbereich des traditionellen Datenschutzrechts bezieht sich auf die Verarbeitung von personenbezogenen Daten. In der Datenbasis muss aber Personenbezug im Sinne herkömmlicher Definitionen gar nicht bestehen, um statistisch untermauerte Aussagen über Menschengruppen mit bestimmten Eigenschaften zu machen. Allenfalls wird er im Augenblick der Aussage offenkundig, was das Recht wohl nach bisher vorherrschender Ansicht nicht erfasst.

Verschiedentlich wurde gezeigt, dass asymmetrische Machtverhältnisse zwischen einzelnen Parteien im Kontext von Big Data eine neue qualitative Dimension erreicht haben könnten. Soweit solche Wissensbestände zu Marktversagen führen, wird dies kartellrecht-

lich relevant. Vereinzelt ist bereits die Forderung nach stärkeren Offenlegungspflichten beispielsweise für Suchalgorithmen zu hören. Im staatlichen Bereich weisen Informationsfreiheitsgesetze in eine ähnliche Richtung.

Welche Folgen die Analysemöglichkeiten haben, die aus einer umfassenden Verdatung des Alltags resultieren, ist noch nicht abschätzbar. Insbesondere stellt sich die Frage nach gesamtgesellschaftlichen Steuerungserfordernissen und -möglichkeiten. Hierbei werden vielfach grundlegende Fragen aufgeworfen. Eine Klärung zentraler Herausforderungen, wie die der Drittbetroffenheit und neuen Machtasymmetrien, ist vordringlich. Auf dieser Basis können dann Modelle für kleinteiligere Aushandlungsprozesse entwickelt werden.

Problemlagen wie diesen kommt das Recht regelmäßig über staatliche Verbote mit Genehmigungsverfahren bei. Doch verbieten und genehmigen kann man nur, wenn man weiß, wann. Im ersteren Fall muss das mit der Handlung verbundene Risiko hinreichend hoch sein. Im zweiten Fall muss deutlich sein, dass im konkreten Fall das Risiko anders zu bewerten ist. Doch was sind die Risiken? Und wann sind sie handhabbar? Wir wissen es noch nicht – jedenfalls nicht in allen Fällen.

Hieraus ergeben sich folgende übergeordnete Fragen:

A.1. Gibt es Fälle, in denen die Datenerhebung nicht (ausschließlich) auf eine Einwilligung des Betroffenen gestützt werden kann? Fraglich ist dies vor allem in Fällen, in denen aus diesen Daten Rückschlüsse auf Dritte, die nicht eingewilligt haben, möglich sind.

Wenn dem so ist:

A.2. In welchen Fällen gilt das? Wo also bedarf es ergänzender Mechanismen? Wenn man auf den Begriff des Risikos abstellt, welche Risiken sind das konkret? Gibt es hierfür definierbare, abstrakte Risikoschwellen? Ist beispielsweise die Anzahl der Datensätze ein Kriterium?

A.3. Welche Regulierungsinstrumente kommen hierfür in Betracht? Genehmigungsverfahren? Generelle Verbote? Weitere? Wie kommen ergänzende Regelungen zustande? Durch staatliche Maßnahmen oder alternative Regulierungsansätze?[31]

[31] Siehe hierzu auch im Folgenden 5.2, insbesondere dort B.2.

5.2. Eckpunkte für die Einordnung einzelner Verfahren

Die Etablierung von neuen Datenanalysen bringt auch historisch regelmäßig auf den Einzelfall bezogene gesellschaftliche Aushandlungsprozesse mit sich. Diese finden typischerweise zu zwei Zeitpunkten statt, die gelegentlich auch zusammen fallen. Eine erste Aushandlung findet dann statt, wenn deutlich wird, dass bestimmte Datenbestände anfallen und damit der Auswertung zugänglich werden, ein weiterer dann, wenn deutlich wird, welche Auswertungen vorgenommen werden können, und welche Erkenntnisse daraus gewonnen werden können. Ein dritter Fall – vom zweiten in der Praxis freilich nur graduell zu unterscheidender Prozess – kann dann eintreten, wenn deutlich wird, welche Handlungen aus den gewonnen Erkenntnissen folgen. Ausgehend von den obigen, abstrakten Fragestellungen, lassen sich auf der Ebene der konkreten Verfahren folgende Fragen aufwerfen und ableiten:

B.1. Wann sind Aushandlungsprozesse für konkrete Analyseverfahren erforderlich? Welches sind die Indikatoren für deren Erforderlichkeit? Wer prüft, ob dies der Fall ist?

Bereits bekannt ist, dass die Analysen insbesondere dann problematisch werden, wenn Informationen über Personen gesammelt werden. Dass diese im Weiteren aber nicht als personenbezogene Daten im Sinne des Datenschutzrechts verarbeitet werden müssen, wurde gezeigt. Das Datenschutzrecht greift hier (zumindest derzeit) als Instrumentarium wohl zu kurz.

B.2. Gibt es verallgemeinerbare prozedurale Elemente für derartige Aushandlungsprozesse? Welche Möglichkeiten gibt es, und wann sind sie einschlägig?

Zu denken ist etwa an

- den politisch-parlamentarischen Prozess,
- behördliches Ermessen,
- technische Standards,
- industrielle Selbstverpflichtungen,
- daneben aber auch an neue Multi-Stakeholder-Prozesse,

die hier Anregungen liefern können. Insbesondere die drei letztgenannten Punkte bieten dabei jedoch noch erheblichen Raum, prozedurale Fragen auch vor dem Hintergrund der konkreten Hausforderungen durch Big Data weiter zu erörtern.

B.3. Welches sind die Aushandlungslinien? Was kann Gegenstand der Aushandlung sein? Mit anderen Worten: Welche Anforderungen sind an die Durchführung von Big-Data-Analysen im konkreten Einzelfall zu stellen?

Erkennbar ist hier etwa das Kriterium Transparenz, und zwar entlang der drei Achsen,

- der Erhebung (welche Daten werden erfasst?),
- der Verarbeitung (in welchen Verfahren werden die Daten verarbeitet, welche Erkenntnisarten können daraus entstehen und welche Erkenntnisse werden gewonnen?) und
- der Nutzung (der Erkenntnisse) inklusive der Rückwirkungen (also etwa Preisdiskriminierungen).

Entsprechend können für jede relevante Big-Data-Analyse im Sinne der Frage B.1. (siehe oben) die folgenden Fragen gestellt werden:

B.4. Ist in dem konkreten Prozess die Erhebung der Daten offenzulegen? Ist sie etwa in einem öffentlichen Register zu hinterlegen?[32]

B.5. Ist der konkrete Verarbeitungsprozess öffentlich zu machen? Welche Erkenntnisse können gewonnen werden? Welcher Algorithmus wird verwendet? Wie ist dieser implementiert?

B.6. Ist zu veröffentlichen, wie die gewonnen Erkenntnisse tatsächlich genutzt werden?

B.7. Gibt es Fälle, in denen die Erkenntnisse selbst zugänglich gemacht werden müssen?

Grundlegender, aber entlang derselben Achsen, sind auch Einschränkungen möglich: bestimmte Daten nicht einzubeziehen, bestimmte Verfahren nicht anzuwenden, bestimmte Erkenntnisse nicht wirken zu lassen. Ein Beispiel für letzteres sind Antidiskriminierungsregeln.

Diese Kriterien stehen jedoch oft im Konflikt mit den Interessen der Verarbeiter, genau diese Informationen für sich zu behalten, etwa um Marktvorteile zu erzielen.

[32] Öffentliche Register könnten in einem erheblichen Maß zur Transparenz beitragen. Zwar werden sie gerne als „Bürokratiemonster“ gegeißelt, aber in Zeiten, in denen Datenverarbeitung omnipräsent ist, erscheint es nicht einsichtig, warum die Führung von Registern einen erheblichen Aufwand mit sich bringen muss. Neben einer Übersicht für den individuell Betroffenen, wo Daten über ihn oder sie gespeichert sein könnten oder wer über Erkenntnisse verfügen könnte, eröffnet ein Register zudem den reizvollen Pfad von Big-Data-Analysen über Big Data.

B.8. Gibt es Verfahren, deren Anwendung schon grundsätzlich auszuschließen ist? Gibt es Erkenntnisse, die nicht gewonnen oder zumindest nicht verwertet werden sollten?