Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein

Big Data und Datenschutz

Thilo Weichert

1 Was ist Big Data?

„Big Data“ steht für *große Datenmengen*, die über das Internet oder anderweitig gesammelt, verfügbar gemacht und ausgewertet werden. Viele der Daten sind personenbezogen. Sie lassen sich, herausgelöst aus den ursprünglichen Erhebungskontexten, zu beliebigen Zwecken nutzen, z. B. um statistische Trends zu erkennen.

„Big Data“ steht für *neue Chancen* – für neue soziale, ökonomische, wissenschaftliche Erkenntnisse, die dazu beitragen können, die Lebensverhältnisnisse in unserer komplexen Welt zu verbessern.

„Big Data“ steht auch für *neue Risiken* – die Möglichkeiten des informationellen Machtmissbrauchs durch Manipulation, Diskriminierung und Unterdrückung: Werden große Mengen von Daten durch private oder öffentliche Stellen zusammengeführt, so kann deren informationelle Ausbeutung zu massiver Verletzung informationeller Grundrechte der Menschen und damit zur Gefährdung ihrer Freiheitsrechte führen. Alles wäre mit allem kombinier- und dann auswertbar: Angaben über Finanztransaktionen, Bonität, medizinische Behandlung, privaten Konsum, Berufstätigkeit, aus der Internetnutzung, von elektronischen Karten und Smartphones, aus der Video- oder der Kommunikationsüberwachung.

Ob Google, Apple oder Facebook, ob Polizei, Finanzbehörden oder Geheimdienste, ob Krankenkassen, Arbeitgeber, Versicherungen, Banken oder Versandhandelsunternehmen – Zugriff und Nutzung auf die von den Menschen alltäglich generierten Daten müssen nach *legitimierten Regeln* erfolgen, insbesondere wenn dies in ein individuelles Scoring, Tracking oder Profiling einfließt. Zugleich müssen Regeln der Datentransparenz abgeleitet, festgelegt und umgesetzt werden.[1]

2 Das Problem

Der unter 1 wiedergegebene Text stammt aus der Ankündigung der Sommerakademie 2013 der Datenschutzakademie Schleswig-Holstein und des Unabhängigen Landeszentrums für Datenschutz, die am 26.08.2013 in Kiel stattfinden wird. Er thematisiert, dass Big Data derzeit mit großer Begeisterung von der Politik und von IT-Unternehmen propagiert wird, verbunden mit großen Erkenntnis-, ja *Heilsversprechen*

[1] Datenschutzakademie Schleswig-Holstein, Jahresprogramm 2013, S. 41.

für eine bessere Informationsgesellschaft.[2] Wenn vom „Wachstums- und Beschäftigungstreiber Internet“ die Rede ist, steht zugleich das Thema „Big Data – Big Chances“ auf der Tagesordnung, ohne dass die Risiken oder Probleme auch nur thematisiert werden.[3] Tatsächlich erleben wir derzeit eine gewaltige Werbekampagne zum Verkauf neuer informationstechnischer Angebote, bei denen Speicherplatz, Rechengeschwindigkeit, Lokalität, Sprachen, Kulturen, Kontexte und konkrete Zwecke und Fragestellungen nur noch untergeordnete Rollen spielen sollen.

Es ist verblüffend, wie ungestraft wertefrei hierbei geworben und argumentiert wird. Dass sich unsere westlichen Informationsgesellschaften von denen Indiens, Chinas oder des Irans durch demokratische Prozesse, die Gewährleistung von Freiheitsrechten, das Prinzip der Solidarität und durch rechtsstaatliche Verfahren qualitativ unterscheiden und dass dies das Wesen unserer westlichen Informationsgesellschaften ausmacht, ist zumeist nicht erkennbar. Der Algorithmus wird zum angeblich unbestechlichen Maßstab für die gesellschaftlichen Phänomene gemacht; Fragen nach Interessen, Macht und Kontrolle, Rechte, Pflichten und *Werte werden ausgeklammert*. Die bisherige Diskussion über Big Data ist ein Beispiel für die von Lawrence Lessig im Jahr 2000 gemachte Feststellung „Code is Law“ sowie daraus folgend, dass der elektronische Code unsere Werte und Traditionen des Humanismus und der Aufklärung zu zerstören droht.[4]

Eine der zentralen Errungenschaften moderner Werte und Traditionen ist – von Deutschland und Europa ausgehend – der Datenschutz bzw. genauer das *Grundrecht auf informationelle Selbstbestimmung*.[5] Dieses Grundrecht ist eine Grundbedingung einer informationstechnisch hochentwickelten freiheitlichen Demokratie. Es ist damit unverzichtbarer Bestandteil einer digitalen Menschenrechtscharta, eines Codex Digitalis Universalis. Diese Erkenntnis hat sich noch nicht in den USA und in vielen Ländern Asiens und Afrikas durchgesetzt.[6]

Der folgende *Grundlagentext* macht sich zur Aufgabe, die Mythen, die Chancen und die Risiken, die sich um Big Data ranken, zu analysieren, diese ins Verhältnis zum Grundrecht auf informationelle Selbstbestimmung zu stellen sowie Bedingungen zu definieren und Perspektiven aufzuzeigen, wie die technischen Möglichkeiten des Big Data mit unserem modernen Wertekanon in Einklang gebracht werden können. Hierbei wird insbesondere auf das deutsche Datenschutzrecht Bezug genommen, das eine Grundlage auch im europäischen Datenschutzrecht findet.

3 Technische Möglichkeiten der Datenanalyse

[2] Die Industrie schaltet ganzseitig begeisterte Werbeanzeigen zum Thema, z. B. IBM in Der Spiegel 7/2013, 19.
[3] So z. B. eine Veranstaltung des Wirtschaftsrats Deutschland am 20.03.2013 in Berlin.
[4] Lessig, Code is Law, http://harvardmagazine.com/2000/01/code-is-law-html.
[5] BVerfG, NJW 1984, 419 ff.
[6] Weichert, Codex Digitalis Universalis, Datenschutz und Überwachung in ausgewählten Staaten, in: Schmidt/Weichert, Datenschutz, 2012, 344 ff., 418 ff.

„Big Data“ steht für den erkenntnisorientierten Umgang mit den *digitalen Daten*, welche die Menschheit produziert. Nach Schätzungen sind dies bisher 2,8 Zettabytes (1 Zettabyte (ZB) sind eine Billion Gigabyte; 1 ZB in Byte umgerechnet ergibt eine Zahl, die aus einer 1 gefolgt von 21 Nullen besteht).[7] Allein die Daten im Internet werden auf ein Volumen von 1,9 Zettabyte geschätzt. Es wird von einer jährlichen Steigerungsrate hinsichtlich verfügbarer Daten von 50 % ausgegangen.[8]

Personenbezogene Daten entstehen in mehr oder weniger strukturierter Form vor allem im Web 2.0, also in sozialen Netzwerken wie Facebook, Videoplattformen wie Youtube, Foto-Plattformen wie Flickr. Jede Minute teilen die 1 Milliarde Facebook-Nutzenden ca. 685.000 Inhalte. Jede Minute werden auf Youtube ca. 72 Stunden Videomaterial hochgeladen. Durch das „Internet der Dinge“, also die Datenerhebung aus dem Auto, den mobilen Devices, dem Stromnetz, den Haushaltsgeräten … steigt die Zahl der Datenquellen exponentiell. Nach Angaben von Cisco waren schon 2011 10,3 Mrd. Geräte mit dem Internet verbunden.[9]

Das digitale Material ist in immer größerem Umfang potenziell global verfügbar. Das reine Datenvolumen ermöglicht noch kein Big Data. Hinzukommen müssen die Variabilität bzw. die *Konvergenz* der unterschiedlichen Datentypen und Datenquellen, was deren Verbindung und Analyse ermöglicht. Weiterhin ist die Rechengeschwindigkeit von Bedeutung, die innerhalb kurzer Zeit hochkomplexe Auswertungen und sogar Echtzeitergebnisse verspricht.[10] Es ist nachvollziehbar, dass sich Unternehmen, Behörden oder sonstige Stellen mit Zugriff auf einen nennenswerten Datenbestand Gedanken machen, wie sie diesen neuen Rohstoff nutzen bzw. ausbeuten können.

Das, was heute „Big Data“ genannt wird, ist nichts völlig Neues, sondern hat sich aus bestehenden Instrumenten weiterentwickelt. Grundlagen sind die staatliche Statistik sowie das privatwirtschaftliche soziodemografische und ökonomische Auswerten von eigenen Datenbeständen, die in sog. *Data Warehouses* strukturiert und auswertbar gemacht wurden. Die Auswertung nach Gesetzmäßigkeiten, Kausalitäten und Korrelationen erfolgt mit immer mächtiger bzw. komplexer werdenden Analyse-Werkzeugen (Data Analytics).[11] Beim „Data Mining“ werden u. U. Fragen beantwortet, die zuvor überhaupt nicht gestellt wurden.

---

[7] IDC, The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East, Dezember 2012, http://germany.emc.com/leadership/digital-universe/iview/.

[8] Lohr, Mining an Information Explosion, New York Times, selected für SZ, 20.02.2012, 1; Fessler, Die Datenflut steigt – wie können wir sie nutzen? Datareport 1/2013, 11.

[9] Reichel, Gut vorbereitet, Vitako aktuell 1/2013, 6 f.

[10] Fesser (Fn. 8) S. 12.

[11] Hansen/Meißner (ULD) Clauß/Steinbrecher/Pfitzmann (TU Dresden), Verkettung digitaler Identitäten, Studie im Auftrag des Bundesministeriums für Bildung und Forschung, 2007.

Die *anfallenden Daten* können sich stark unterscheiden: Daten zu Inhalten und Nutzungen des Internet (Suchanfragen, Surfverhalten, Kommunikation, Online-Banking, Online-Einkauf, E-Government) sowie aus der sonstigen Telekommunikation, Angaben zu Kunden, Warenbeständen und Lieferwegen, zu Finanztransaktionen, aus der Werbung, Bilder und Metadaten aus Bildern und Videoüberwachung, Angaben aus der Statistik, der Markt- und Meinungsforschung, wissenschaftlichen Registern, Medieninformationen, Inhalte aus Datenbanken medizinischer Einrichtungen, Ergebnisse aus Mustererkennungsverfahren, Angaben aus der Nutzung von Funkchips (RFID) etwa im Bereich der Logistik, Karteninformationen, Mikrosensorergebnisse über Standorte, Maschinen, Bewegungen, Vibrationen, Temperaturen, Feuchtigkeit, Chemie ... Bei der aktuellen Diskussion noch wenig berücksichtigt ist die Entwicklung im Bereich der Biotechnologie durch die Sequenzierung der Genome von Lebewesen einschließlich des Menschen. Etwa 10.000 menschliche Genome gelten heute als sequenziert, was einer Datenmenge von 5 Petabyte entspricht. Die Kosten für die Sequenzierung eines menschlichen Genoms bewegen sich heute nur noch im vierstelligen Dollarbereich und nehmen weiter ab.[12]

Personenbeziehbare Daten können außerhalb des Menschen entstehen und bei Dritten, vor allem in Unternehmen oder Behörden, gespeichert und ausgewertet werden. Denkbar und nicht unerhebliche Praxis ist auch, dass *Menschen selbst* sich digital erfassen und diese Daten auswerten und damit sich und ihre Umwelt zu verstehen versuchen. Beim „Quantified Self“ wie auch bei den vielfältigen individuell genutzten Applikationen etwa auf Smartphones verlassen sich die Menschen nicht nur auf sich selbst und auf die von ihnen beherrschten Hilfsmittel, sondern nehmen die „Hilfe“ von Dienstleistern in Anspruch, die sich häufig nebenbei die Befugnis zur anderweitigen Auswertung dieser Daten geben lassen.[13]

Digitale Daten sind nicht die einzige potenzielle Quelle für Big Data. Weitere Wachstumsfelder erschließen sich durch die Digitalisierung bisher nur analog vorhandener Informationen. Geschieht die Digitalisierung nicht im Rahmen einer Diensteerbringung, z. B. im Rahmen der Nutzung eines Internetdienstes oder eines Kfz, so kann diese ausschließlich zwecks Anlieferung für Big Data erfolgen. Informationstechnisch werden hierfür *Mustererkennungsverfahren* eingesetzt, mit denen analoge Sachverhalte in eine digitale „Ordnung“ gebracht werden, z. B. Menschenbewegungen, Stimmen, Handschrift, Gesichter, Mimik und Gestik.[14]

Schon beim Data Warehousing besteht eine wichtige Aufgabe darin, die mehr oder weniger unstrukturierten Daten in eine für Auswertungen *sinnvolle Struktur* zu bringen.

---

[12] Rauchhaupt, Dicke Daten, www.faz.net, 06.10.2012.

[13] „Big Data“ verändert unser Leben radikal, www.welt.de, 22.01.2013.

[14] Monroy, Bundesregierung lässt Bevölkerungsscanner zur Erkennung „bedrohlicher Handlung“ an Mimik und Gestik beforschen, netzpolitik.org, 26.02.2013; Behrens, Wie Überwachung „intelligent“ werden soll, www.sueddeutsche.de 09.03.2013.

Diese Aufgabe stellt sich beim Verknüpfen von Datenbeständen unterschiedlicher Herkunft beim Big Data verstärkt.

Hinsichtlich der personenbezogenen Datenauswertung sind derzeit zweifellos *soziale Netzwerke* für Big Data am attraktivsten, zumal hier die Verfügbarkeit für die Dienste- und interessierten Applikationsanbieter quantitativ wie qualitativ groß ist. Die Angebote in diesem Bereich sprießen geradezu aus dem Boden. An vorderster Front stehen dabei die großen Dienste- und Applikationsanbieter selbst, allen voran Facebook und Google.

Von den Unternehmen, denen umfangreich Internetdaten zur Verfügung stehen, haben viele ihren *Stammsitz in den USA*, wo sie hinsichtlich der Datenauswertung bisher nur wenigen nennenswerten rechtlichen Restriktionen unterworfen sind.[15]

Es tummeln sich immer mehr *Spezialanbieter* auf dem Markt. So wertet z. B. die US-Rüstungsfirma Raytheon mit ihrem Programm RIOT (Rapid Information Overlay Technology) öffentlich zugängliche Daten bis hin zu Metadaten von publizierten Fotos zu Personen in sozialen Netzwerken aus, um Ortsangaben, Kontaktnetzwerke, Interessen und vieles mehr darzustellen, zu analysieren und zu prognostizieren.[16] Angesichts des Big-Data-Hypes rechnen Unternehmen, die Lösungen für die Integration und Analyse von Daten anbieten, wie z. B. die deutsche Software AG, mit nachhaltigem Wachstum und Profit.[17]

Insbesondere hinsichtlich der *inhaltlichen Auswertung* von Internet-Kommunikation und -Inhalten aus Social Media, etwa Netzwerke, Blogs, oder Foren, etablieren sich immer mehr Lösungen und Angebote. Es geht dabei darum, aus dem riesigen Strom von Daten für den eigenen Kontext Relevantes auszusieben, um hierauf z. B. durch eine politische oder eine Werbe-Strategie antworten zu können. Damit können sowohl einzelne Unmutsäußerungen als auch Massenkritiken bis hin zu Shitstorms detektiert und untersucht werden, möglicherweise in Echtzeit bzw. in einem sehr frühen Stadium. Beispielsweise spielt bei Inhaltsanalyse von Online-Äußerungen die Anzahl der Beiträge zu bestimmten Themen eine Rolle, für die Schwellenwerte definiert werden. Auswertbar ist auch, von wem die Beiträge stammen und wie (häufig) diese von anderen aufgegriffen, kommentiert und weiterverteilt werden. Wortführer und Multiplikatoren können so erkannt und der Dialog mit ihnen kann gesucht werden. Eine Grundlage der Analyse sind als relevant festgelegte Begriffe, wobei diese in einer sog. Kontextextraktion auf die interessierende Fragestellung hin selektiert werden. Über eine Sentiment-Analyse sollen nicht nur Personen, Themen und Fragestellungen erkannt werden, sondern auch Gefühle oder gar Ironie: „Die Software arbeitet also weitgehend wie ein intelligenter menschlicher Zuhörer, der abstrahieren, zwischen den Zeilen lesen

[15] Hofmann, Die Versuchungen von Big Data, Böll.Thema 2/2012 – Digitale Ökonomie – S. 33, Weichert, Privatheit und Datenschutz im Konflikt zwischen den USA und Europa, RDV 2012, 113 ff.
[16] US-Firma erprobt Facebook-Überwachung, www.spiegel.de, 11.02.2013.
[17] Tauber, „IT-Standort Deutschland hat viel Innovationskraft“, www.welt.de, 10.02.2013.

und Nuancen erkennen kann“.[18]

4 Der Baum der Erkenntnis

Die heute mit Big Data verbundenen Versprechungen sind gewaltig. Tatsächlich sind diese aber so alt wie die moderne Informationstechnik allgemein. Schon in den 70er Jahren des letzten Jahrhunderts glaubte der damalige Präsident des Bundeskriminalamtes *Horst Herold*, mit Hilfe der kriminalgeografischen und zugleich individualisierten informationstechnischen Auswertungen von Kriminalitäts- und sozialökonomischen Daten vor dem Straftäter am Tatort sein zu können, um das geplante Vergehen oder Verbrechen zu verhindern.[19]

Die Diskussion um die *Volkszählungen* 1983 und 1987 waren geprägt von der Erwartung, mit den Volks-, Berufs- und Arbeitsstättenzählungen eine „aktuelle, umfassende und zuverlässige Datenbasis für gesellschafts- und wirtschaftspolitische Entscheidungen des Bundes, der Länder und Gemeinden“ zu liefern.[20]

Das *Versprechen von Big Data* ist, durch Analyse von Daten Probleme zu erkennen und lösen zu können, bevor sie sich gesellschaftlich ausgebreitet haben. Gesundheitsmanager wollen durch Auswertung von Google-Suchanfragen vorhersagen, welchen Weg eine Grippewelle einschlägt. Durch Bewegungsdaten sollen Verkehrsstaus prognostiziert und Ausweichstrecken ausgemacht werden. Anhand von strukturierten oder unstrukturierten Massendaten sollen verlässliche Prognosen über Klimaveränderungen, Konjunkturentwicklungen, demografischen Wandels oder Vorhersagen über Preisentwicklungen erstellt werden. Selbst im Bereich der Sozialverwaltung, von der finanziellen Grundsicherung über die Kinder- und Jugendhilfe bis hin zur Arbeitsvermittlung, wird gehofft, durch „intelligente“ Datenanalyse, die einem dauernden „automatisierten Selbstlerneffekt“ ausgesetzt wird, das Mittel zur Problemlösung gefunden zu haben.[21]

Folgende *Fähigkeiten* werden Big-Data-Analysen zugeschrieben:

- „Query“ und Reporting,
- Data-Mining,
- Datenvisualisierung,
- Vorhersagemodelle (Prognosen),
- Optimierung (v. a. von Prozessen).,

---

[18] So Nold, Am Puls der Zeit, Vitako aktuell 1/2013, 17.
[19] Weichert, Informationelle Selbstbestimmung und strafrechtliche Ermittlung, 1990, S. 228 m. w. N.
[20] Appel, Volkserfassung – zweiter Versuch, in Hummel/Pollähne/Ruhne/Sögtrop, „Kein Staat mit diesem Staat?“, 1986, Frage, S. 267 ff.; Appel in Appel/Hummel, Vorsicht Volkszählung, 1987, S. 21 ff.; Rottmann/Strohm, Was Sie gegen Mikrozensus und Volkszählung tun können, 1986, S. 134 ff.
[21] Sauter, Weiterdenken erlaubt, in Vitako aktuell 1/2013, 9.

- Simulation,
- Integration verschiedener Datenformate (z. B. Umwandlung von Sprache in Text),
- Geodaten- und raumbezogene Analysen,
- Videoanalyse,
- Sprachanalyse.[22]

Um den Baum der Erkenntnis ernten zu können, werden die vorhandenen Datenmassen wissenschaftlich durchdrungen. Big Data ist also das Basismaterial für die wissenschaftlich unabhängige wie auch die kommerzielle bzw. zweckgerichtete *Forschung*. Big Data soll „Spielraum für Experimente" liefern, bei der Performance-Daten in Echtzeit kontrollierte Simulationen ermöglichen, um Bedürfnisse und Variabilität zu identifizieren.[23] Durch Auswertung von Realdaten sollen Erkenntnisse gewonnen werden, die der Menschheit zugutekommen. Forschende der Harvard-University leiteten aus der Analyse der Handydaten von 15 Mio. Kenianern die Ausbreitungswege von Tropenkrankheiten ab und meinen damit neue Instrumente gegen deren Verbreitung entwickeln zu können.[24]

Hinter dem Versprechen auf Erkenntnis stecken weitergehende Versprechen: Mit Einsätzen in der *Wirtschaft* sollen sich neue Geschäftsmodelle entwickeln, die Arbeitsplätze, wirtschaftliches Wachstum, Profit und Steuereinnahmen generieren. Eine zentrale Anwendung von Big Data soll darin liegen, den „Zugang zu Kunden" zu verbessern, indem feinkörnig Bevölkerungs- und Kundensegmente erstellt und Waren und Dienstleistungen hierauf zugeschnitten werden.[25] Es besteht die Erwartung, dass datenbasierte Entscheidungsfindung eine Erhöhung der Produktivität bewirkt.

Auch für die *öffentliche Verwaltung* stehen die Versprechen der Rationalisierung und der Effektivierung der eigenen Aufgabenerfüllung im Raum, das Einsparen von Ressourcen und größere Bürgernähe durch Datenbereitstellung (Open Data), Transparenz und verstärkte digitale Kommunikation. Über Benutzer-Feedback und Crowd-Sourcing sollen die demokratische Meinungsbildung und die Entscheidungsfindung optimiert werden. Sicherheitsbehörden setzen Big-Data-Instrumente zu Zwecken der Strafverfolgung und der Gefahrenabwehr, aber auch zur Gefahrenprognose und nachrichtendienstlichen Analyse ein.[26]

---

[22] Fesser (Fn. 8), S. 14.
[23] Urbanski, Projekt mit Breitenwirkung, Vitako aktuell 1/2013, 10.
[24] Meister, Das Unbehagen im Datenhaufen, www.taz.de, 21.11.2012.
[25] Urbanski (Fn. 23), S. 10.
[26] Kleine Anfrage der Fraktion Die Linke und Antwort der Bundesregierung, Automatisierte Strafverfolgung, Data Mining und sogenannte erweiterte Nutzung von Daten in polizeilichen Informationssystemen, BT-Drs. 17/11582 v. 22.11.2012; zur polizeilichen Recherche in sozialen Netzwerken Weichert, Facebook, der Datenschutz und die öffentliche Sicherheit, in: Möllers/van Ooyen, Jahrbuch Öffentliche Sicherheit 2012/2013, S. 380 ff., https://www.datenschutzzentrum.de/facebook/JBOES-2012-2013-Sonderdruck-Weichert.pdf.

Big Data soll den Menschen *Informationsangebote* liefern über alles, was interessiert: Infrastrukturangebot für Konsum, Verkehr, Arbeit, Freizeit, Kommunikation, aber auch Daten über Wirtschaft und Politik, kombiniert nach individuellem Interesse und Belieben. Die US-Zeitung Washington Post präsentierte den Prototyp einer Software mit dem Namen „Truth Teller", die in Echtzeit automatisch von Politikerreden oder Talkshow-Diskussionen Transkripte erstellt, deren sachlichen Gehalt dann mit einem Algorithmus mit Datenbanken und Archiven abgleicht und zu den Äußerungen schließlich die Anzeigen „wahr" oder „falsch" auswirft.[27]

Welche Bedeutung Big Data für die *demokratische Willensbildung* haben kann, zeigt der Umstand, dass dem Chief Technology Officer im Wahlkampfteam von Barack Obama im Jahr 2012 der Verdienst zugesprochen wird, dem US-Präsidenten die zweite Amtszeit gesichert zu haben, indem dieser mit seinem Big-Data-Werkzeug „the optimizer" „Milliarden von Daten potenzieller Wähler (identifiziert hat), bevor sie via Facebook, E-Mail oder Hausbesuch kontaktiert wurden". Zuvor wurden Antworten auf folgende Fragen gegeben: „Wann schaut wer welche TV-Sendung, damit Obama-Spots ihr Publikum auch wirklich erreichen? Wann ist wer zu Hause, damit die Wahlkämpfer vor Ort nicht vergeblich vor der Haustür stehen? Bei welchem Thema muss man jeden Einzelnen abholen?"[28]

Angesichts derartiger Perspektiven gibt es einige Menschen, die – wieder einmal – meinen, perspektivisch könne auf die *Humanwissenschaften* verzichtet werden, zumal Politik, Soziologie, Psychologie, Ökonomie ja auch nichts anderes seien als die Verarbeitung von Informationen.[29] Dabei werden Korrelation mit Kausalität und komplexe, von Menschen programmierte Technik mit Intelligenz verwechselt.

5 Eingriffe in das Grundrecht auf Datenschutz

Erste *öffentliche Auseinandersetzungen* zeigen, dass die Akzeptanz der Öffentlichkeit für Big-Data-Anwendungen nicht zu erreichen ist, wenn sich der Eindruck ergibt, hiermit erfolge eine Entmündigung oder gar eine Diskriminierung von Einzelpersonen oder Gruppen. So musste O2 seine Planungen zurückziehen, die Standortdaten der Mobilgeräte seiner Kundschaft für Marketingzwecke zu monetarisieren.[30] Nachdem bekannt wurde, dass das Hasso-Plattner-Institut im Auftrag der Schufa Daten aus sozialen Netzwerken auswerten möchte, um Rückschlüsse auf die Bonität zu ziehen,

[27] Hofmann, Echtzeit-Wahrheitscheck für die politische Debatte, www.sueddeutsche.de, 04.02.2013.
[28] President´s Nerd, T-Systems, Best Practice 1/2013, 07.
[29] Graff, Wenn Daten sprechen, www.sueddeutsche.de, 02.01.2013, zitiert Anderson von „Wired".
[30] O2 stoppt Datenprojekt, SZ, 02.11.2012, 19; Bernau, Der gläserne Bürger hat seinen Schrecken verloren; www.sueddeutsche.de, 02.11.2012; Bernau/Paukner/Zydra, Marktforschung mal anders, SZ, 31.10./01.11.2012, 19.

war der drohende öffentliche Ansehensverlust so groß, dass das Projekt wieder aufgegeben wurde.[31] Und der niederländische Navigationsgeräteanbieter TomTom hatte ein gewaltiges Imageproblem, als bekannt wurde, dass das Unternehmen anonymisierte Verkehrsbewegungsdaten an die Polizei verkaufte, damit diese ihre Geschwindigkeits- oder sonstigen Verkehrskontrollen effektivieren konnte.[32]

Die *Grundrechtssensibilität der Menschen* in Bezug auf die Datenverarbeitung nimmt in Europa zu. Es war vor wenigen Jahren in Großbritannien noch kein politisches Problem, selbst personenbezogene Prognosen in der Polizei oder in der Sozialverwaltung vorzunehmen.[33] Als im Herbst 2012 bekannt wurde, dass zum Zweck einer Verbesserung des Gesundheitswesens angeblich anonymisiert die Gesundheitsdaten der gesamten britischen Bevölkerung ausgewertet werden sollten, wurde hiergegen jedoch heftiger Widerstand laut.[34]

In den *USA* sind vergleichbare Angebote noch weitgehend akzeptiert, etwa das Personenscoring und -ranking auf der Basis von Daten aus sozialen Netzwerken durch den Anbieter Klout.[35] Während sich in den USA die Visionen der Big-Data-Lobby auch rechtlich noch weitgehend ungehindert ausleben können, steht dem in Europa der Datenschutz und das Grundrecht auf informationelle Selbstbestimmung entgegen. Dieses beinhaltet, dass die Menschen grundsätzlich selbst bestimmen können, „wer was wann bei welcher Gelegenheit über sie weiß“.[36] Dies steht der Kollektivierung personenbezogener Daten durch Big Data schon im Grundsatz entgegen.

Art. 8 Abs. 2 der *Europäischen Grundrechtecharta* legt fest, dass personenbezogene Daten nach „Treu und Glauben“ verarbeitet werden müssen „für festgelegte Zwecke und mit Einwilligung der betroffenen Person oder auf einer sonstigen gesetzlich geregelten legitimen Grundlage“. Es ist gesetzlich gefordert, dass im Interesse des digitalen Grundrechtsschutzes „so wenig personenbezogene Daten wie möglich“ verarbeitet und diese deshalb anonymisiert oder pseudonymisiert werden (§ 3a BDSG). Der Grundsatz der Datenvermeidung und der Datensparsamkeit ist angesichts der digitalen Datenmassen nicht anachronistisch[37], sondern sowohl für die Betroffenen individuell als auch für manche Systemverantwortlichen eine Antwort auf den Trend zu Big Data.

---

[31] Hasso-Plattner-Institut kündigt Forschungsauftrag der Schufa, DANA 3/2012, 118 f.
[32] TomTom verkauft Bewegungsprofile an Polizei, DANA 2/2011, 86.
[33] Korff, Guaranteeing Liberty or Big Brother, 2007, https://www.datenschutzzentrum.de/sommerakademie/2007/sak2007-korff-surveillance-in-the-united-kingdom.pdf.
[34] Meister, Big Data vs. Privacy: Großbritannien zentralisiert 52 Millionen Krankenakten für Forschung, Anonymisierung fraglich, netzpolitik.org, 30.08.2012.
[35] Klout: Personenscoring in sozialen Netzwerken, DANA 3/2012, 130; Weichert, www.datenschutz.de, 11.08.2012.
[36] BVerfG NJW 1984, 422.
[37] So Ulbricht, Schwieriges Terrain, Vitako aktuell 1/2013, 15.

Hinsichtlich der *Sensibilität* der Datenverarbeitung kann aus persönlichkeitsrechtlicher Sicht danach differenziert werden, ob und inwieweit ein Personenbezug besteht: Wann hat im Rahmen eines Big-Data-Verfahrens ein Personenbezug keine Relevanz (mehr)? Ist tatsächlich ausgeschlossen, dass ein solcher Personenbezug hergestellt wird? Sind schon die Quelldaten anonymisiert, so besteht ein geringeres Risiko, wobei jedoch beachtet werden muss, dass mit Menge und Qualität der Daten die Reidentifizierungsrisiken steigen. Dient ein Verfahren der Erlangung personenbeziehbarer oder gar personenbezogener Ergebnisse, so ist die Sensibilität regelmäßig sehr hoch.

Die Frage der datenschutzrechtlichen Zulässigkeit von Big-Data-Analysen durch private, d. h. nicht-öffentliche Stellen orientiert sich an den §§ 27 ff. BDSG. Bestehen keine Legitimationsmöglichkeiten durch Vertrag oder Einwilligung, so setzt jede personenbezogene Datennutzung das *Überwiegen der Auswertungsinteressen* gegenüber den schutzwürdigen Betroffeneninteressen voraus (§ 28 Abs. 1 S. 1 Nr. 2 BDSG). Für Forschungsauswertungen gelten zumeist privilegierende Spezialregelungen.

Hinsichtlich der Auswertung durch öffentliche Stellen besteht zudem eine Privilegierung im Bereich der *Statistik*.[38] Wegen der Aufgabenbezogenheit jeglicher Datenverarbeitung und einem insofern strengeren Zweckbindungsgrundsatz sind Big-Data-Auswertungen nach allgemeinem Datenschutzrecht durch öffentliche Stellen grundsätzlich ausgeschlossen. Spezialregelungen, die gewisse Auswertungen erlauben, können insbesondere im Sicherheitsrecht bestehen.[39]

6 Stadien der Datenverarbeitung

Hinter dem kurzen Begriff Big Data verbirgt sich zumeist eine lange Reihe von Verarbeitungsschritten, die aus rechtlicher Sicht gesondert betrachtet und legitimiert werden müssen. Am Anfang steht das *Erheben der Daten* – aus allgemein zugänglichen Quellen, bei den betroffenen Menschen, bei Dienstleistern für diese, bei Unternehmen und aus der Verwaltung. Die sammelnde Stelle muss als verantwortliche Stelle für die weiteren Verarbeitungsschritte eine Befugnis zum Speichern und Weiterverarbeiten haben. Bestehen insofern gesetzliche Privilegien, z. B. für Forschende oder für Statistikbehörden, so sind diese regelmäßig verbunden mit Restriktionen hinsichtlich der Nutzung und Weitergabe der Auswertungsergebnisse.[40]

Keinen Beschränkungen hinsichtlich der Datenerhebung unterliegen grundsätzlich allgemein zugängliche Daten. Wurden diese zulässigerweise veröffentlicht, so eröffnet dies zweckungebunden weitere Nutzungen. Daten, die mit Einwilligung der Betroffenen,

[38] BVerfG, NJW 1984, 423 ff.
[39] Dazu Kleine Anfrage der Fraktion Die Linke und Antwort der Bundesregierung, BT-Drs. 17/11582 vom 22.11.2012.
[40] BVerfG NJW 1984, 423.

zur Durchführung von Verträgen oder zur gesetzlichen oder geschäftlichen Aufgabenerfüllung erlangt wurden, unterliegen dagegen jeweils einer mehr oder weniger strengen *Zweckbindung*, die fortgilt, solange ein Personenbezug besteht. Generell gilt, dass Daten nicht zusammengeführt werden dürfen, deren ursprüngliche Zwecke miteinander nicht vereinbar sind und sich gegenseitig ausschließen.[41]

Die Möglichkeiten bei der Auswertung sind das, was die Versprechungen des Big Data in den Himmel sprießen lassen. Aus persönlichkeitsrechtlicher Sicht sind Art, Umfang und denkbare Verwendungen der erhobenen Daten relevant, sowie die *Gefahr von deren Missbrauch*. Unzumutbare intime Angaben, Selbstbezichtigungen sowie Gefahren der sozialen Abstempelung sind auszuschließen.[42] Nicht nur der Gefahr der Diskriminierung, sondern auch der Manipulation ist vorzubeugen.

Im Rahmen der Auswertung personenbezogener Daten wird in Bezug auf Fragestellung und Art der Verarbeitung unterschieden: Beim *Tracking* wird das Verhalten einer Person auf der Zeitachse hinsichtlich einer bestimmten Eigenschaft, insbesondere des Aufenthaltsortes, verfolgt. Eine Untersuchung hat ergeben, dass mit Bewegungsdaten von nur zwei Wochen zu einer Person zu 93 % exakt künftige Aufenthalte vorhergesagt werden können.[43] Internet-Tracking erfolgt hinsichtlich unseres Surf-, Nutzungs- und Kommunikationsverhaltens und bestätigt durch seine Werthaltigkeit auf dem Werbemarkt, wie leicht Menschen im Hinblick auf ihr Konsumverhalten berechen- und vorhersehbar sind.

Beim *Scoring* erfolgt eine zahlenmäßige Bewertung einer Eigenschaft einer Person auf der Grundlage vieler, möglichst für die Eigenschaft relevanter Merkmale. Nachdem Missbräuche im Bereich von Finanzdienstleistern bekannt wurden[44], wurde das Scoring vom Bundesgesetzgeber rechtlich in § 28a BDSG eingegrenzt. Die berechneten Scores können zur Grundlage eines vergleichenden Ratings, also der Festlegung einer Rangliste, verwendet werden.

Beim *Personalizing* erfolgt auf Grund von persönlichen Merkmalen keine zahlenmäßige Bewertung, sondern eine Antwort auf relevante Fragestellungen in Bezug auf einzelne Personen, etwa das Interesse als Konsument oder die Effektivität einer bestimmten pharmazeutischen Behandlung oder Therapie (personalisierte Medizin).[45]

Verschiedene Formen personenbezogener Big-Data-Auswertungen werden unter dem Begriff *Profiling* zusammengefasst. Das BVerfG hat schon früh ein Verbot des Erstellens totaler Persönlichkeitsbilder ausgesprochen und dies definiert als die zwangsweise

[41] BVerfG, NJW 1984, 427.
[42] BVerfG, NJW 1984, 422 f.
[43] Song et al., Limits of Predictability in Human Mobility, Science 327, 1018, 2010.
[44] Kamp/Weichert, Scoringsysteme zur Beurteilung der Kreditwürdigkeit, 2005.
[45] Rasmussen, Das vermessene Ich, Böll.Thema 3/2012, 30 f.

Registrierung und Katalogisierung des Menschen in seiner ganzen Persönlichkeit.[46] Beim Erstellen von Persönlichkeitsprofilen kann zwischen einer zeitlichen (Langzeitprofile) und einer Sektor übergreifenden Dimension (Querschnittsprofile) unterschieden werden. Das Verbot des Erstellens totaler Persönlichkeitsbilder soll verhindern, dass die Vergangenheit prägend für die Zukunft eines Menschen wird und dass das individuelle Ausfüllen einer gesellschaftlichen Rolle sich für den Menschen auf andere Rollen auswirkt.[47] Beim Profiling werden zumindest digitale Persönlichkeitsteilbilder erstellt.

Bei der Auswertung von Big Data ist nicht nur die Frage der Quantität und der Qualität der Datenanalysen relevant. Die Art der Daten kann für eine Bewertung entscheidend sein. Restriktionen gelten generell für *sensible Daten*. Dabei kann es sich um Berufsgeheimnisse handeln (vgl. § 203 StGB), um besondere Arten personenbezogener Daten (Herkunft, Ethnie, politische Meinung, religiöse oder philosophische Überzeugung, Gewerkschaftszugehörigkeit, Gesundheit, Sexualleben; § 3 Abs. 9 BDSG) und/oder um besondere Diskriminierungsrisiken (Herkunft, Nationalität, Ethnie, Religion, Weltanschauung, Behinderung, Alter, Geschlecht, sexuelle Identität (Allgemeines Gleichbehandlungsgesetz – AGG)).

Generell gilt, dass die rechtliche Bewertung einer personenbeziehbaren Datenverwendung vom jeweiligen verfolgten *Zweck* abhängt. Hinsichtlich bestimmter Zwecke hat § 2 Abs. 1 AGG besondere Restriktionen vorgesehen: Bildung, berufliche Betätigung, politische und gewerkschaftliche Betätigung, Sozialschutz, Gesundheitsschutz, Zugang zur öffentlichen Daseinsvorsorge.

7 Prinzipien und Schutzziele des Datenschutzes

Die Überprüfung der *datenschutzrechtlichen Zulässigkeit* von Big-Data-Anwendungen setzt voraus, dass diese konkret dokumentiert sind und auf dieser Basis bewertet werden können. Für diese Bewertung können aus nationalen und europäischen Regelungen einige übergreifende Anforderungen abgeleitet werden.

7.1 Grundsätze

Generell gilt bei jeder personenbezogenen Datenverarbeitung das *Verbot mit Erlaubnisvorbehalt*. Dies bedeutet, dass entweder eine Erlaubnis durch den Betroffenen (durch eine Einwilligung, vgl. §§ 4a BDSG, 13 Abs. 2 TMG) oder durch eine gesetzliche Regelung gegeben sein muss (§ 4 Abs. 1 BDSG, § 12 Abs. 1 TMG, Art. 7 EU-DSRL).[48] Als gesetzliche Legitimation kommt das allgemeine Datenschutzrecht (bei nicht-öffentlichen Stellen §§ 27 ff. BDSG) in Betracht. Denkbar ist auch die Anwendung bereichsspezifischen Datenschutz- und -verarbeitungsrechts.

[46] BVerfG, NJW 1969, 1707.
[47] Weichert in Däubler/Klebe/Wedde/Weichert, BDSG, 3. Aufl. 2010, Einl. Rz. 46.
[48] Weichert, DuD 2013, 246 ff..

Eine *Einwilligung* setzt voraus, dass der jeweils Betroffene über verarbeitende Stelle, Art der Verarbeitung, Zweck und Datenumfang informiert ist und hierzu ausdrücklich und freiwillig seine Zustimmung erteilt hat (§ 4a BDSG, Art. 2 lit. h EU-DSRL). Eine Opt-out-Möglichkeit genügt nicht. Das Einholen massenhafter Einwilligungen kann schon aus logistischen Gründen ein Problem darstellen. Das Einholen von Einwilligungen mit dem Kauf eines Gerätes oder dem Abschluss eines Vertrags ist zwar grundsätzlich nicht ausgeschlossen. Zu beachten ist aber, dass tatsächlich Wahlmöglichkeit und Freiwilligkeit bewahrt bleiben, was bei einer Koppelung eines Kauf- oder Dienstleistungsvertrags mit der Einwilligung regelmäßig nicht der Fall ist. Praktisch unlösbare Probleme bzgl. der Einholung von Einwilligungen entstehen, wenn eine Big-Data-Anwendung für eine unüberschaubare Zahl von Stellen mit nicht eindeutig bestimmten Daten für noch nicht bestimmbare Zwecke evtl. erst nachträglich zur Verfügung gestellt werden soll.

Das datenschutzrechtliche Prinzip der *Zweckbindung* zielt darauf ab, dass Daten nur für den Zweck verwendet werden dürfen, für den sie erhoben wurden. Dies schließt Big Data, bei dem keine konkreten Zwecke verfolgt werden, tendenziell aus. Je offener ein Zweck festgelegt wird, desto mehr Auswertungsmöglichkeiten können sich eröffnen, desto risikobehafteter ist aber zugleich auch die Verarbeitung für die Betroffenen. Die Zweckbindung kann im Rahmen der Gesetze nur aus Gründen eines überwiegenden Interesses der Allgemeinheit oder bestimmter Dritter begrenzt aufgehoben werden.[49]

Mit der Zweckbindung eng verbunden ist der *Grundsatz der Erforderlichkeit* bzw. der Datensparsamkeit. Eine Verarbeitung ist nur zulässig, soweit sie für die Erreichung des konkreten Zweckes nötig ist. Wird kein Zweck von vornherein festgelegt oder wird dieser nur allgemein beschrieben, so ist eine Erforderlichkeitsprüfung unmöglich oder erheblich erschwert. Datensparsamkeit steht der häufig vorzufindenden Praxis entgegen, zunächst einmal so viel Daten wie möglich zu erheben, auch wenn der Bedarf daran noch nicht festgestellt ist.

Zentraler Bestandteil des Schutzes des Rechts auf informationelle Selbstbestimmung ist die Gewährleistung der *Betroffenenrechte*. Dabei handelt es sich insbesondere um das Recht auf Auskunft sowie das Recht auf Datenkorrektur (Berichtigung, Löschung, Sperrung, Gegendarstellung). Diese Rechte bestehen, solange die Daten eine Personenbeziehbarkeit aufweisen.

Die *Datenschutzkontrolle* ist operativ für den Grundrechtsschutz von großer Bedeutung. Diese erfolgt innerhalb einer Organisation durch die jeweilige Leitung sowie in Delegation durch den betrieblichen bzw. behördlichen Datenschutzbeauftragten. Auf einer zweiten Ebene ist eine externe unabhängige staatliche Kontrolle durch die Datenschutzaufsichtsbehörden verfassungs- und europarechtlich gefordert und tatsächlich vorgesehen. Je komplexer eine Datenverarbeitung ist, desto höher ist der

[49] BVerfG, NJW 1984, 419, 422.

Kontrollbedarf und desto schwieriger die Kontrolle. Dies gilt in noch stärkerem Maße, wenn verschiedene verantwortliche Stellen und damit oft auch verschiedene Kontrollinstanzen beteiligt sind, so wie dies bei Big-Data-Anwendungen oft der Fall ist.

Mit einer Gestaltung der Systeme, die schon von Anfang an Datenschutzanforderungen integriert, mit technisch-organisatorischen Maßnahmen kann der Schutz des Rechts auf informationelle Selbstbestimmung unterstützt werden. Dies betrifft sowohl die *Datensicherheit* als auch weitere rechtliche Anforderungen. Die Maßnahmen verfolgen hierbei unterschiedliche Schutzziele.

7.2 Schutzziele

Aus dem materiell-rechtlichen Erfordernis des Schutzes des Rechts auf informationelle Selbstbestimmung für die Betroffenen lassen sich sechs Schutzziele mit technisch-organisatorischer Relevanz ableiten.[50] Ein *Schutzbedarf* besteht bei Personenbeziehbarkeit der Daten in jedem Fall.[51] Er ist umso höher, je sensibler und je aussagekräftiger die Daten und je relevanter die verfolgten Zwecke für das Persönlichkeitsrecht der Betroffenen sind.[52]

Das Ziel der *Verfügbarkeit* zielt darauf ab, dass die eingesetzten Verfahren und die eingeführten Daten zeitgerecht zur Verfügung stehen und ordnungsgemäß anwendet werden. Das Ziel wird relevant, wenn Big-Data-Anwendungen von akuter Entscheidungsrelevanz sind.

Das Ziel der *Integrität* ist angesichts des Ausmaßes der bei Big Data verwendeten Daten in der Praxis nur annähernd zu erreichen. Integrität bedeutet, dass die Daten unversehrt, vollständig, zurechenbar und aktuell sind. Die schiere Masse der verarbeiteten Daten verursacht beim Big Data eine hohe „Datenunsicherheit“. Bei Verarbeitungen mit dem Ziel einer personenbezogenen Auswertung kommt der Integrität eine hohe Bedeutung zu; dies gilt umso mehr, je relevanter die abgeleiteten Schlüsse sind. Auch die Herauslösung aus dem ursprünglichen Kontext kann bereits die Aussage der Daten und damit die Integrität der Ergebnisse massiv beeinträchtigen. Vielen Datenquellen ist mangelnde Integrität immanent, etwa den Inhaltsdaten aus Social Media.[53] Entsprechendes gilt, wenn, wie oft beim Big Data, die Daten sich nicht in alleiniger Verfügungsgewalt einer verantwortlich zu machenden Stelle befinden.[54] Die Komplexität einer zumeist wenig transparenten Auswertung dient zwar regelmäßig der Optimierung der Ergebnisse, sie hat aber zugleich zwangsläufig zur Folge, dass die Integrität aus Betroffenensicht weiter beeinträchtigt wird.

---

[50] Vgl. § 5 Abs. 1 LDSG SH; zum normativen Gehalt der Schutzziele Bock/Meissner, Datenschutz-Schutzziele im Recht, DuD 2012, 425 ff.
[51] BVerfG, NJW 1984, 422.
[52] Rost, Standardisierte Datenschutzmodellierung, DuD 2012, 436.
[53] Fesser (Fn. 8), S. 12.
[54] Probst, Generische Schutzmaßnahmen für Datenschutz-Schutzziele, DuD 2012, 441.

Vertraulichkeit bedeutet, dass nur Berechtigte auf das Verfahren und die Daten zugreifen können. Das Ziel der Vertraulichkeit steht einer Praxis, möglichst viele Daten zusammenzuführen und auszuwerten, nicht zwangsläufig entgegen. Die Standardmaßnahme zur Sicherung der Vertraulichkeit ist die Verschlüsselung. Diese beeinträchtigt jedoch bei Big Data regelmäßig die angestrebte Verknüpf- und Analysemöglichkeit. Weitere Instrumente sind solche der technisch-organisatorischen Abschottung.[55] Zur Wahrung der Vertraulichkeit unter Anstreben größtmöglicher Verfügbarkeit wurde z. B. im Bereich der gesetzlichen Krankenversicherung ein technisch-organisatorisch abgeschottetes Instrument zur Datentransparenz vorgesehen (§§ 303a ff. SGB V).

Das Ziel der *Nicht-Verkettbarkeit* dient den Datenschutzgrundsätzen der Zweckbindung und der Erforderlichkeit/Datensparsamkeit (s. o. 7.1) und steht dem Grundanliegen von Big Data, eine Verkettung so umfassend wie möglich zu realisieren, diametral entgegen. Um eine einfachere Verkettbarkeit zu ermöglichen, können (vor der Analyse) Standardisierung, Aufbereitung und Formatierung eine Rolle spielen. Unterliegen sämtliche in die Auswertung einfließenden Daten sowie die Auswertung selbst einem gemeinsamen Zweck (bei einer Stelle), so kann eine unbegrenzte Verkettbarkeit rechtlich ermöglicht werden. Werden jedoch mit den Datensätzen (von unterschiedlichen Stellen), die bei Big Data einfließen, unterschiedliche Zwecke verfolgt, so eröffnet dies hohe Anforderungen an die Zwecktrennung durch logische Separierung, durch zweckdienliche Metadaten oder durch Anonymisierung und Aggregierung vor der Verkettung.

Transparenz bedeutet, dass die Verarbeitung der Daten mit zumutbarem Aufwand nachvollzogen, überprüft und bewertet werden kann. Die Transparenz dient allen Beteiligten, der verantwortliche Stelle, den Betroffenen, der Datenschutzkontrolle sowie der Öffentlichkeit im Rahmen einer demokratischen Kontrolle. Eine lückenlose Dokumentation des Verfahrens sowie eine umfassende Protokollierung der konkreten Verarbeitungsprozesse sind notwendig, um Transparenz zu gewährleisten (siehe auch unten 10).

Intervenierbarkeit bedeutet, dass das Verfahren so gestaltet werden muss, dass den Betroffenen die Ausübung der ihnen zustehenden Rechte wirksam möglich ist. Die wesentlichen Betroffenenrechte sind die der Auskunft (§ 34 BDSG), der Berichtigung, Sperrung und Löschung (§ 35 Abs. 1-4, 6 BDSG), des Widerspruchs (§ 35 Abs. 5 BDSG) und der Gegendarstellung. Die Datenlöschung zielt u. a. darauf ab, dass den Betroffenen der Anspruch zukommt, mit bestimmten Daten im Leben nicht mehr konfrontiert zu werden (right to be forgotten). Dieses Recht lässt sich auch durch eine wirksame Anonymisierung realisieren. Die Intervenierbarkeit setzt nicht das Vorhandensein von Klarnamen voraus, sondern gilt auch bei personenbezogen zuordenbaren Pseudonymen.

[55] Probst (Fn. 54), S. 441 f.

8 Datenschutzrechtliche Einzelfragen

8.1 Personenbezogenes Datum

Ein weit verbreitetes Legitimationsmuster für Big-Data-Analysen ist das Negieren eines Personenbezuges. So wird immer wieder die Ansicht vertreten, bestimmte Geo- oder *Sachdaten* hätten keinen Personenbezug und könnten deshalb uneingeschränkt genutzt werden. Diese Annahme trifft für Orte oder Dinge, die einer konkreten Person mehr oder weniger eindeutig zugeordnet werden können, nicht zu. Daten zur eigenen Wohnung mit Kühlschrank, Strom-, Wasser- und Wärmeverbrauch oder digitalen Media-Geräten sind i. d. R. ebenso personenbezogen wie Daten zu einem Handy oder einem anderen Mobilgerät, einem Kraftfahrzeug, einer Kreditkarte oder einem mit RFID-Technik ausgestatteten Kleidungsstück. Personenbezogen sind regelmäßig auch Identifikatoren von Geräten mit einer personalen Zuordnung, also etwa IP-Adressen, Cookie-IDs, Browser-„Fingerabdrücke“ oder Kartennummern.

An einem Personenbezug ändert sich nichts, wenn an Stelle des Klarnamens ein *Pseudonym* verwendet wird, wenn es also der verarbeitenden Stelle möglich ist, ohne unverhältnismäßigem Aufwand das Identifikationsmerkmal aus dem Pseudonym und evtl. weiteren Angaben im Datensatz herzuleiten (§ 3 Abs. 5 Nr. 6a BDSG).

Ein Personenbezug besteht erst dann nicht mehr, wenn die Einzelangaben über persönliche oder sachliche Verhältnisse überhaupt nicht mehr oder zumindest nur mit einem unverhältnismäßigen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können (§ 3 Abs. 6 BDSG). Dieser *Ausschluss einer personalen Zuordnung* ist umso weniger möglich, je mehr qualifizierte Daten von möglicherweise unterschiedlichen Quellen bei einer Big-Data-Anwendung zusammengeführt werden können. Im Internet veröffentlichte Daten sind eine Fundgrube zur Reidentifizierung von vermeintlich anonymisierten Datensätzen.

8.2 Öffentlich zugängliche Quellen

Es trifft nicht zu, dass Daten, die allgemein zugänglich sind oder veröffentlicht werden dürften, für Big-Data-Anwendungen keinen Einschränkungen unterliegen. Vielmehr ist im Grundsatz hinsichtlich jeder einzelnen Auswertung der Daten zur prüfen, ob nicht „das schutzwürdige Interesse des Betroffenen ... gegenüber dem berechtigten Interesse der verantwortlichen Stelle offensichtlich überwiegt“ (§§ 28 Abs. 1 S. 1 Nr. 3, 29 Abs. 1 S. 1 Nr. 3 BDSG). Die allgemeine Zugänglichkeit von Daten hat sich mit dem *Internet* und den dortigen Web2.0-Diensten massiv erhöht, etwa in sozialen Netzwerken, in Blogs und Informationsportalen und insbesondere wegen der Erschließbarkeit über Suchmaschinen.

Öffentlich zugänglich sind nur Daten, die sowohl von der Intention als auch von der eingesetzten Technik her nicht auf den Zugriff durch einen *eingeschränkten Nutzerkreis*

beschränkt sind. Internetdaten sind öffentlich zugänglich, wenn diese zulässigerweise für jedermann einsehbar sind. Bedarf es einer zusätzlichen Authentisierung oder Autorisierung, etwa in Bezug auf „Freunde“ in einem sozialen Netzwerk, so sind die Daten nicht mehr öffentlich zugänglich. Dies gilt z. B. für die Daten, die im Rahmen der Graph-Suche von Facebook erschlossen werden.[56]

Es ist nun nicht möglich, in jedem Fall eine Einzelfallprüfung durch Interessenabwägung zu verlangen. Vielmehr genügt es bei der Weiterverarbeitung von öffentlich zugänglichen Massendaten, eine *Pauschalierung* möglicher schutzwürdiger Interessen vorzunehmen. Dieser Weg wird jedoch versperrt, wenn die betroffene Person ihren Widerspruch gegen die Auswertung erklärt hat (§ 35 Abs. 5 BDSG, Art. 14 EU-DSRL). Im Rahmen einer Pauschalierung stehen insbesondere dann schutzwürdige Interessen entgegen, wenn besondere Arten personenbezogener Daten (§ 3 Abs. 9 BDSG; Art. 8 EU-DSRL) mit erfasst werden, wenn Zwecke mit einer Anwendung verfolgt werden, die möglicherweise konkrete negative Folgen für die Betroffenen auslösen können, oder wenn bei der Anwendung eine Erstellung eines Persönlichkeitsprofils erfolgt.[57] Je unbestimmter die herangezogenen Datenquellen und die verfolgten Zwecke sind, umso offensichtlicher sind entgegenstehende schutzwürdige Interessen.

9 Auswertungsmethode

Eine große Blackbox beim Big Data und allen seinen praktischen und ideologischen Vorläufern liegt in der verwendeten *Analysemethode*. Zwar verpflichtet z. B. § 28b Nr. 1 BDSG beim Scoring zu nachweisbaren „wissenschaftlich anerkannten mathematisch-statistischen Verfahren“, doch gibt es hierzu bisher nur wenige Erfahrungen bei Datenschutzbehörden. Unternehmen betrachten ihre Analysealgorithmen als Betriebs- und Geschäftsgeheimnis. Dies hat zur Folge, dass eine unabhängige Überprüfung der Analyseverfahren und damit auch der Analyseergebnisse faktisch oft (bisher) nicht möglich ist. Dies wiederum ist der fruchtbare Boden für statistische Tricksereien und vorurteilsbelastete und manipulierte Ergebnisse.[58]

Viele Auswertungen zielen auf *statistische, anonymisierte und aggregierte Ergebnisse*, die nicht direkt einzelnen Personen zugeordnet werden können. Dies ist aus persönlichkeitsrechtlicher Sicht weniger problematisch als das „Personalizing“. Dessen ungeachtet können auch aggregierte Ergebnisse individuelle Beeinträchtigungen zur Folge haben. Ein persönlichkeitsrechtlich wenig sensibles, wenngleich anschauliches Beispiel ist die Weitergabe aggregierter Verkehrs- und Geschwindigkeitsdaten des Navigationsdienstes TomTom an die niederländische Polizei, die auf dieser Grundlage ihre Kontrollpraxis ausrichtete. Diese Weitergabe wirkt sich direkt auf die Personen aus, die auf bestimmten Auswertungsstrecken mit ihrem Auto unterwegs sind und deren

[56] Rubenzer, Profilputz bevor die Zuckerberg-Suche kommt, www.chip.de, 06.02.2013; https://graph.facebook.com/.
[57] Weichert (Fn. 47), Einl. Rz. 45 f.
[58] Lohr (Fn. 8) zitiert Goldin von der George Mason University, Virginia.

frühere Bewegungsdaten anonymisiert, ausgewertet und aggregiert wurden.[59]

Höher ist die persönlichkeitsrechtliche Relevanz, wenn die Aggregierung sich an der *Zugehörigkeit zu identitätsstiftenden Gruppen* etwa politischer, religiöser, sozialer oder ethnischer Art orientiert. Insofern kann selbst bei aggregierten Ergebnissen ein aus dem Gleichheitsgrundsatz oder dem Sozialstaatsprinzip abzuleitendes Diskriminierungsverbot verletzt sein.[60]

Eine rechtliche Konsequenz hat der Gesetzgeber aus fehlender Transparenz und dem Risiko von Fehlprogrammierung dadurch gezogen, dass er relevante negative *automatisierte Einzelentscheidungen* auf einer großen Datenbasis davon abhängig machte, dass dem Betroffenen „durch geeignete Maßnahmen“ die Wahrung seiner berechtigten Interessen ermöglicht wird (§ 6a BDSG, Art. 15 EU-DSRL). Hierzu gehört auch die Informationen, dass überhaupt eine automatisierte Entscheidung erfolgt, welche die wesentlichen Entscheidungsgründe sind und wie das Verfahren logisch aufgebaut ist.

10 Transparenz

Transparenz zu Big-Data-Verfahren hat vielfältige Dimensionen (s. o. 7.2). Transparenz hat neben der Funktion für die Betroffenen eine *gesamtgesellschaftlich-demokratische Funktion*. Sie muss gegenüber der Öffentlichkeit, wenigstens aber gegenüber einer unabhängigen Forschungsgemeinschaft oder einer Kontrollbehörde hergestellt werden. Je relevanter ein Verfahren für Politik, Wirtschaft und soziales Leben ist, desto weniger kann die Offenlegung mit dem Ziel des Schutzes von Betriebs- und Geschäftsgeheimnissen verhindert werden. Transparenz entsteht nicht alleine durch den Markt, sondern erfordert Marktregulierung. Marktteilnehmer versuchen durch Vorenthaltung ihres Know-hows regelmäßig, ihre Marktstellung zu verbessern.

Besondere öffentliche Transparenz ist im Hinblick auf die *Ergebnisse der Auswertung* gefordert. Angesichts der Mächtigkeit von Big-Data-Instrumenten besteht die Gefahr, dass durch nicht transparente und unkontrollierte Nutzung der Ergebnisse, denen der Nimbus der wissenschaftlichen, unangreifbaren Objektivität anhaftet, Informationsungleichgewichte verstärkt werden.[61] Es ist daher im Sinne einer demokratischen Bürgergesellschaft, Big Data mit Open Data bzw. Open Government zu verbinden und eine Pflicht zur Veröffentlichung bestimmter großer Datenbestände und der damit durchgeführten Verfahren vorzusehen.[62]

---

[59] TomTom verkauft anonymisierte Bewegungsprofile an Polizei, DANA 2/2011, 86.
[60] Weichert (Fn. 47), Einl. Rz. 51.
[61] Boyd/Crawford, Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon, Information, Communication & Society 15:5, 2012, S. 662 ff.
[62] Schallaböck, Die Antwort auf Big Data lautet Open Data, Böll.Thema 2/2012 – Digitale Ökonomie – S. 35.

Individuelle *Transparenz für den Betroffenen* ist zugleich Grundvoraussetzung für dessen informationelle Selbstbestimmung. Im Hinblick darauf, dass bei Big Data oft eine unübersehbare Zahl von Betroffenen existiert, muss die öffentliche Transparenz die individuelle Benachrichtigung und Auskunft im Regelfall ersetzen. Dies darf aber nicht als Legitimation dafür herangezogen werden, in Bedarfsfall die verfassungsrechtlich geforderte Einzelauskunft oder -benachrichtigung zu unterlassen.[63]

Transparenz kann eine direkte Rückwirkung auf die *Qualität der verwendeten Daten* haben. Der Chef von PayPal meinte, dass Nutzende bereit seien, für ein besseres Erlebnis mehr von sich preiszugeben, „solange sie die Kontrolle über ihre Daten behalten".[64] Nur wenn die individuell einfließenden Daten den Betroffenen bekannt sind, können sie diese auf ihre Aktualität und Richtigkeit hin überprüfen. Dies ist eine wichtige Bedingung für die Richtigkeit der Analyseergebnisses.

11 Technische Lösungen

Nach Darstellung der persönlichkeitsrechtlichen Voraussetzungen kann der Eindruck entstehen, dass es rechtlich praktisch ausgeschlossen ist, personenbeziehbare und damit datenschutzrelevante Big-Data-Anwendungen zu realisieren. Big-Data-Anwendungen zielen regelmäßig nicht auf die Erlangung personenbeziehbarer, sondern allgemeiner Erkenntnisse. Deshalb sind Konzepte gefordert, bei denen eine hinreichende Anonymisierung der personenbezogenen Daten erfolgt. Mit Hilfe eines durchdachten Zusammenwirkens von verschiedenen *Gestaltungs- und Schutzmechanismen* können[65] rechtlich zulässige und zugleich gesellschaftlich akzeptable Einsätze erreicht werden.[66]

Die aktuelle Entwicklung von Big-Data-Anwendungen erfolgt bisher weitgehend ohne Berücksichtigung von datenschutzrechtlichen Anforderungen. Wirtschaft und Verwaltung werden immer mehr feststellen, dass Datenschutz ein zentrales Entwicklungshindernis für Big Data sein kann. Ziel muss es deshalb sein, durch bestimmte Gestaltungsmaßnahmen „*Privacy Preserving Data Mining*" zu verwirklichen, bei dem der Datenschutz hinreichend gewahrt wird.[67]

---

[63] Ein anschauliches nicht akzeptables Beispiel ist etwa die Auskunftserteilung bzw. -verweigerung an Betroffene durch Facebook, siehe Europe-v-Facebook, http://www.europe-v-facebook.org/DE/Anzeigen/anzeigen.html.

[64] Analyse: „Big Data" nimmt Fahrt auf, www.ln-online.de, 21.01.2013.

[65] Etwas zu euphorisch Ulbricht, Schwieriges Terrain, Vitako aktuell 1/2013, 15: „im Regelfall".

[66] Hierzu ausführlich Hansen/Meissner u. a., Verkettung digitaler Identitäten, 2007 (Fn. 11).

[67] Urbanski (Fn. 21), S. 11; grundlegend beispielsweise Agrawal/Srikant, Privacy-Preserving Data Mining, ACM SIGMOD Int'l Conf. on Management of Data, 2000.

11.1 Anonymisierung

Werden mit Big Data statistische, nicht personenbezogene Erkenntnisse angestrebt, so ist das adäquate Mittel zur Einführung von Daten die frühzeitige Anonymisierung und die Verhinderung der Reidentifizierung. *Anonymisierung* ist das Verändern von personenbezogenen Daten so, dass die Einzelangaben nicht mehr oder nur mit einem unverhältnismäßigen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten Person zugeordnet werden können (§ 3 Abs. 6 BDSG). Liegt das für die Deanonymisierung nötige Zusatzwissen bei einer anderen als der die Personendaten verarbeitenden Stelle vor, so kann nicht mehr von Anonymität gesprochen werden. Dabei spielt es keine Rolle, ob dieses Zusatzwissen nur unter Rechtsverletzung beschafft werden kann.[68]

Für eine wirksame Anonymisierung genügt es nicht, *Identifikatoren* (beispielsweise Namen, IP-Adressen oder Ordnungsnummern) zu verändern. Vielmehr müssen diese Daten entweder ersatzlos gelöscht oder so verändert werden, dass über sie keine Zuordnung mehr möglich ist (etwa durch Zufallszahlen).

Eine große Herausforderung im Hinblick auf die Anonymität sind beim Big Data die *Merkmalsdaten*. Je umfangreicher und detaillierter diese sind, umso größer ist die Wahrscheinlichkeit, dass durch einen Abgleich dieser Merkmalsdaten eine Reidentifizierung erfolgt.

Eine valide Form der Anonymisierung kann in der *Aggregierung* von Daten liegen, d. h. in der Zusammenführung einer größeren Zahl von Einzeldatensätzen und deren weitere Verarbeitung als einheitlicher Gruppendatensatz. Dabei muss darauf geachtet werden, dass die Aggregierung mit derart vielen Datensätzen erfolgt, dass hinsichtlich einzelner Merkmale keine Einzelnennungen erfolgen bzw. erfolgen können.

Am Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS) in Sankt Augustin wird erforscht, wie durch die Analyse von *Mobilfunk- und Internetinformationen* notdienstrelevante Daten extrahiert werden können, ohne gegen Datenschutzbestimmungen zu verstoßen. Die hierfür entwickelte Lösung besteht darin, dass die Datensätze in Teile zerlegt werden und diese Teile durchgewürfelt werden.[69] Eine vergleichbare Zielsetzung verfolgt die Telekom im Hinblick auf eine Nutzung von Mobilfunkbewegungsdaten für lokalisierte Werbung. Hier soll eine Anonymisierung durch kurzfristige Vergabe einer Hash-ID für die Lokalisierungsdaten erreicht werden.[70]

11.2 Verarbeitungsmanagement

Aus technischer Sicht höchst voraussetzungsvoll sind Big-Data-Anwendungen, bei denen nicht von Anbeginn mit anonymisierten Daten gearbeitet werden kann. Derartige

[68] Weichert, Das Geschäft mit den Verordnungsdaten, DuD 2013, 130.
[69] Fischermann/Hamann, Wer hebt das Datengold? www.zeit.de, 06.01.2013.
[70] Motionlogic – Dienste, Technik & Datenschutz, Februar 2013.

Verfahren bedürfen regelmäßig eines wirksamen *Mixes von technisch-organisatorischen Absicherungen*, die möglicherweise durch weitere rechtliche Vorkehrungen und Verfahren abgesichert werden. Diese Verfahren müssen auf einem validen Datenschutzmanagement aufsetzen. Dabei muss es darum gehen, die Ziele Zweckbindung und Datensparsamkeit miteinander zu verbinden. Hierzu kann Metadaten von Datensätzen eine wichtige Funktion zukommen, die den Kontext der Daten beschreiben und sogar maschineninterpretierbare Vorgaben zur Datennutzung beinhalten können (Sticky Policies[71]). Auch können gekapselte Anonymisierungswerkzeuge, wie sie in zertifizierten Verfahren des Online Behavioural Advertising angewendet werden, eine wichtige Rolle spielen.[72] Weiterer Forschungsbedarf besteht zudem bei Ansätzen wie „Homomorphic Encryption“[73], wodurch Berechnungen auf der Basis von verschlüsselten Daten ermöglicht werden, sowie „Differential Privacy“[74], bei der Datenbankabfragen und -antworten einen Filter passieren, der Unschärfen hinzufügt.

Datenschutzverträgliche Big-Data-Technologien sind bisher kaum *erforscht und entwickelt*. Soll Big Data in unserer freiheitlich-demokratischen Informationsgesellschaft eine Zukunft erhalten, so sind diese Technologien unabdingbar. In einem zweiten Schritt bedarf es der Festlegung gewisser Standards im Hinblick auf den Datenschutz.

12 Schlussfolgerung

Es wäre die falsche Antwort auf die Möglichkeiten des Big Data, angesichts der bestehenden datenschutzrechtlichen Hindernisse nach dem *Gesetzgeber* zu rufen.[75] Die Wirkungszusammenhänge sowie die möglichen technischen, organisatorischen und rechtlichen Lösungen beim Big Data müssen nicht mit neuen Gesetzen geregelt werden. Vielmehr geht es zunächst darum, die bestehenden Gesetze anzuwenden. Diese Regelungen verlangen nach einem Ausgleich der berechtigten Verarbeitungs- und der schutzwürdigen Betroffeneninteressen. Gegen diese Vorgaben ist auch verfassungsrechtlich kein Kraut gewachsen.

Daher ist es dringend nötig, die Spreu vom Weizen zu trennen, die Propaganda zur IT-Förderung von den wirklich sinnvollen Anwendungsmöglichkeiten. Dann müsssen technische und organisatorische Lösungen erforscht und entwickelt werden, die die Datenschutzanforderungen für Big Data umsetzen, beispielsweise für eine Nutzung der

---

[71] Pearson/Casassa Mont, Sticky Policies: An Approach for Managing Privacy across Multiple Parties, IEEE Computer Magazine, Vol. 44, Nr. 9, 2011, S. 60 ff.
[72] EuroPriSe, https://www.european-privacy-seal.eu/results/fact-sheets/EuroPriSe%20Behavioural%20Targeting%20FS-Follow%20Up.pdf.
[73] Lauter/Naehrig/Vaikuntanathan, Can Homomorphic Encryption be Practical? 3rd ACM Workshop on Cloud Computing Security (CCSW’11), 2011.
[74] Dwork, Differential Privacy, 33rd International Colloquium on Automata, Languages and Programming, 2006.
[75] So aber Ulbricht, Schwieriges Terrain, Vitako aktuell 1/2013, 15.

Daten ohne Personenbezug. Letztlich muss eine *öffentliche Debatte* darüber stattfinden, inwieweit unsere offene demokratische Gesellschaft die Generierung digitalisierter Planungs- und Entscheidungsgrundlagen zulassen möchte, deren Ursprünge nicht mehr im gesunden Menschenverstand und in demokratischen transparenten Prozessen, sondern in Algorithmen zu finden sind.

Kiel, 19.03.2013