erschienen in der FIfF-Kommunikation,
herausgegeben von FIfF e.V. - ISSN 0938-3476
www.fiff.de

FIfF-Konferenz 2016

Un.Sichtbare Datenpraktiken?

Big Data in Wirtschaft, Wissenschaft & Politik

Zusammenfassung des Vortrags von Judith Simon

Die Proliferation von Big-Data-Praktiken ist ein relativ neues und komplexes Thema. Insbesondere die Intransparenz der verwendeten – häufig zudem proprietären – Systeme bereitet dabei Schwierigkeiten. Dabei muss man zwischen funktionaler Intransparenz („kein Zugriff") und epistemischer Intransparenz („Schnelligkeit", „Komplexität der Prozesse", usw.) unterscheiden, um sinnvolle Desiderate für die Governance von Big Data ableiten zu können. Ein wesentlicher Teil solcher Governance-Bemühungen ist neben den rechtlichen Regelungen natürlich auch die transparenzförderliche Gestaltung der IT-Systeme (Stichwort: value-sensitive Design).

Warum und in welcher Weise müssen wir uns mit Big Data beschäftigen?

Die meisten werden das folgende Beispiel kennen: 2012 ging durch die Medien die Geschichte eines US-amerikanischen Vaters, der sich bei der Supermarktkette Target beschwerte, weil er Werbematerial zu schwangerschaftsbezogenen Produkten zugesendet bekommen hatte. Schließlich stellte sich jedoch heraus, dass seine 16-jährige Tochter tatsächlich schwanger war und es ihm bis dahin verschwiegen hatte. Target wusste also mehr als der eigene Vater. Bei aller anfänglichen Empörung oder Überraschung über diesen Fall müssen wir jedoch kritisch fragen: Liegt hier überhaupt ein Problem vor? Wenn ja, wo, und wie kann man es konzeptuell fassen?

Der naheliegendste Ansatz sind die Punkte *Verletzung der Privatsphäre* und *illegitime Datensammlung* – durch die rechtlich legitime Praxis der *Informed-consent*-Verfahren (Einwilligung nach vorheriger Aufklärung) stimmen wir allerdings der Nutzung unserer Daten zu, sodass im Grunde rechtlich hier gar kein Problem vorliegt, denn auch die Tochter hat dem Sammeln und Auswerten ihrer Daten zugestimmt. Die Verletzung der Privatsphäre ist hier jedoch nicht auf Grund des Datensammelns passiert, sondern erst durch die Inferenzen, die auf Basis der Daten, d. h. der Verarbeitung der Daten und der Prognosen, gemacht wurden. Big-Data-Praktiken müssen also zunächst als epistemische, d. h. wissenschaftliche Praktiken betrachtet werden, die

Judith Simon

als solche wiederum ethische und politische Auswirkungen haben, denn Big Data durchdringt mit seinen Verfahren unsere Lebenswelt ganz grundlegend und ist damit natürlich keine rein erkenntnistheoretische Frage. Der Fokus muss dabei auf den konkreten Praktiken des Datensammelns selbst liegen, erst im zweiten Schritt auf ihrer Interpretation, denn ich kann Daten erst dann interpretieren, wenn ich weiß, wo, wie und in welcher Qualität sie erhoben worden sind. Hierzu sind detaillierte und vor allem interdisziplinäre Analysen der Datenpraktiken in Wirtschaft, Politik und Wissenschaft nötig. Die Blickweisen der Ethik müssen wiederum mit der Erkenntnistheorie und den Entscheidungsprozessen der Politik verknüpft werden, sodass daraus rechtliche oder auch ökonomische Konsequenzen abgeleitet werden können.

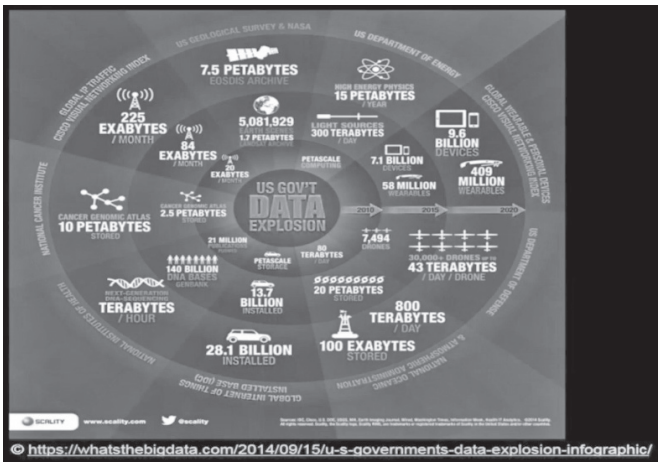
Was ist Big Data?

Nach der klassischen Definition der Industrie gehören zu Big Data (nach IBM) die vier Vs: Volume (*scale of data*), Velocity (*analysis of streaming data*), Variety (*different forms of data*) und Veracity (*uncertainty of data*). Interessanter allerdings ist eine Definition aus den Sozialwissenschaften (Boyd & Crawford 2012) als „*cultural, technological, and scholarly phenomenon that rests on the interplay of Technology, Analysis and Mythology*", wobei hier insbesondere interessant ist, dass als entscheidender Faktor in die Begriffsbestimmung auch der Glauben an die Allmacht der Daten mit aufgenommen wurde, der Glaube daran, große Datenmengen ermöglichen „*a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy.*" Dieser Punkt spielt insbesondere für die Rhetorik in politischen Debatten um die angeblich unbestechlichen und neutralen Big-Data-Analysen eine große Rolle.

Über welche Datensätze reden wir aber eigentlich konkret, wenn wir von Big Data sprechen? Unterschieden werden können einerseits die Daten einer Person innerhalb ihres Alltags, etwa danach, wo sie angefallen sind: Die Daten, die aus den sozialen Medien über die Nutzer.innen gewonnen werden, und zwar hierbei die expliziten wie Kommentare, Fotos, Likes, ebenso wie auch die impliziten Daten, also die Spuren, die wir hinterlassen, wenn wir handeln (Ort, Zeit, Verweildauer, etc.). Andere Datensätze erfassen wiederum jegliche Formen von Transaktionsdaten, die orts- und zeitbezogenen (Meta-)Daten, Sensordaten oder Daten aus dem Internet der Dinge. Unterscheiden können wir die Daten andererseits auch etwa innerhalb der Wissenschaft, z. B. in Daten der Astronomie, der Physik, der Medizin, etc., und weiter differenziert schließlich innerhalb der Disziplinen – z. B. die medizinischen Daten in Daten aus Versuchsreihen, Daten aus elektronischen Patientenakten, in administrative Klinikdaten oder Daten des Personal Health Monitoring (aus Sensoren, Apps, Wearables, etc.). Im Kontext von Big Data ist noch eine dritte Gruppe an Daten relevant: Bürgerdaten, sog. Open Government Data, wobei diese jedoch nicht notwendigerweise offen sind. Hierzu zählen Daten aus dem Geburtsregister, Finanzdaten, Zensusdaten, usw.

Das Interessante an Big Data aber ist nicht, dass es all diese Daten gibt, sondern dass sie mit Hilfe der Datenanalyse vielfältig in Bezug zueinander gesetzt werden können. Erkenntnistheoretisch lässt sich dabei festhalten, dass bestimmte Unterscheidungen, die einst relevant waren, im Big-Data-Kontext zunehmend obsolet werden, wie etwa die Unterscheidung zwischen sensiblen persönlichen Daten und unverfänglichen sonstigen Daten. Ein Beispiel, an dem sich dies zeigt: Der eingangs genannte Fall der Zusendung von Werbung für Schwangerschaftsprodukte. Konsumdaten wie der Kauf einer parfümfreien Body Lotion oder von bestimmten Vitaminpräparaten sind zunächst nicht per se medizinische Daten. Sie werden jedoch genutzt für Prognosen über den medizinischen Zustand des oder der Käufer.in, in diesem Falle zur Prognose einer Schwangerschaft. Das Problem ist also, allgemeiner gesprochen: Je nach Nutzungskontext können scheinbar unverfängliche Daten sensibel werden. Dennoch sind diese (Proxy-)Daten weniger geschützt als herkömmliche Gesundheitsdaten, Finanzdaten, etc.

Eine zweite Differenzierung, die im Big-Data-Zusammenhang obsolet wird, ist die zwischen personenbezogenen und anonymen Daten. Durch Aggregation und Datenverarbeitung können anonyme Daten zunehmend leicht de-anonymisiert werden. Hier ein banales Beispiel: Ich wurde im Zusammenhang mit einem Forschungsprojekt in Österreich um eine Bewertung gebeten und musste hierfür einige Daten über mich angeben, wie meine wissenschaftliche Disziplin und mein Alter. Vier dieser Datenpunkte reichen jedoch aus, um mich eindeutig zu identifizieren. Zu einer Diskriminierung, die möglicherweise aus dieser Zuordnung folgt, reicht jedoch schon die Identifikation als Teil einer Gruppe aus, etwa die Zuordnung zu einer bestimmte Wohngegend, die bei Kreditinstituten schlechtere Konditionen erhält (vgl. Vortrag Corinna Bath: *Sozial gerechte Algorithmen?*).



Datenexplosion

Wer ist Teil von Big Data?

Wer ist nun aber involviert in diese Big-Data-Prozesse? Manovich (2011) unterscheidet hier in „*those who create data (both consciously and by leaving digital foot prints), those who have the means to collect it, and those who have expertise to analyze it*". Ersteres sind wir alle, die weiteren beiden Gruppen jedoch nicht, und dies wiederum führt zu Unterschieden bezüglich der – mitunter unsichtbaren und nicht uns allen verfügbaren – technischen Auswertungsfähigkeiten, was wieder eine Machtasymmetrie erzeugt zwischen denen, die die Daten bereitstellen und denen, die über sie verfügen.

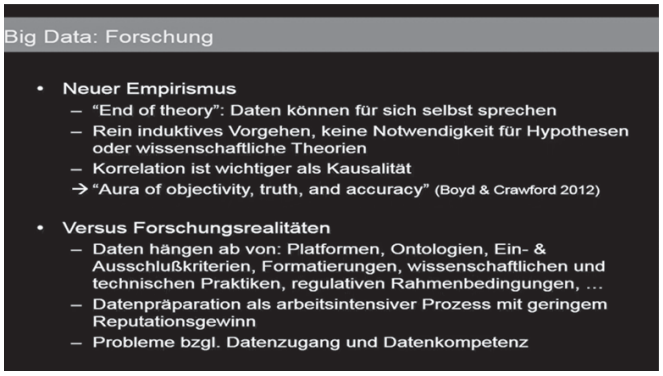
Wer sammelt also die (personenbezogenen) Daten? Schematisch können wir die Bereiche Wissenschaft, Staaten und Unternehmen unterscheiden sowie – eingeschränkt – auch die Nutzer.innen selbst, die bis zu einem gewissen Grad und in geringerem Ausmaß ebenfalls in der Lage sind, Daten zu sammeln und auszuwerten, etwa via Self-Tracking. Die Funktionen von Big Data für diese vier Gruppen sowie die Auswirkungen sind jedoch sehr verschieden. Daher ist es besonders wichtig, den jeweiligen Rahmen sehr genau zu formulieren, wenn von Big Data die Rede ist.

Big Data in der Industrie

Für Unternehmen geht es primär darum, Konsument.innen zu tracken und zu profilieren, um auf der Basis der Datenspuren, die sie hinterlassen, Rückschlüsse etwa für zielgenaue Werbung ziehen zu können. Hierzu zählen nicht nur die großen bekannten Datensammler Facebook und Google, sondern auch viele andere Unternehmen, die schon sehr lange Kompetenzen in der Datenanalyse haben, die uns mit dieser Praxis aber nicht unbedingt geläufig sind, etwa IBM, das sich inzwischen längst als Datenunternehmen versteht. Darüber hinaus gibt es viele Hintergrundakteure, die für Unternehmen arbeiten, deren Namen uns Nutzer.innen oft nicht bekannt sind, die jedoch trotzdem Daten aus diversen Kontexten über uns besitzen – ein geläufigeres Beispiel hier wäre Oracle.

Big Data in der Wissenschaft

Datensammlung und -analyse ist ein Grundbestandteil jeglicher wissenschaftlicher Forschung. Inzwischen wird sowohl aus philosophischer als auch aus wissenschaftlicher Perspektive allerdings diskutiert, ob es durch die neue Form der Datenanalyse zu einem Paradigmenwechsel kommt. Nach den Thesen des sogenannten *Neuen Empirismus* sprechen die Daten inzwischen für sich selbst, sodass es in der Wissenschaftspraxis keine Notwendigkeit mehr für Hypothesen gibt und Experimente oder Kausalitätszusammenhänge obsolet werden (*End of Theory*). Dem gegenüber stehen die Forschungsrealitäten, die zeigen, dass Daten keinesfalls aus sich selbst heraus neutral entstehen, sondern immer abhängig sind von verschiedenen Faktoren wie den Rahmenbedingungen, die für das Sammeln der Daten und für ihre Auswertung gesetzt werden. Beim Neuen Empirismus bleibt zudem u. a. das Problem aller Forschenden unberücksichtigt, stets nur einen limitierten oder kostenpflichtigen Zugang auf insbesondere kommerzielle Daten zu haben. Wissenschaftlich sind diese Thesen daher höchst umstritten, allerdings ist die Rhetorik dahinter für die Politik wiederum enorm verführerisch.



Big Data in der Politik

In der Politik wird auf Big Data große Hoffnungen (z. B. in der Verkehrs- oder Gesundheitspolitik) gesetzt, denn das Sammeln und Verarbeiten von Daten wie etwa das Führen von Geburts- und Sterberegistern (anfangs noch allein von der Kirche übernommen) war immer ein Grundbestandteil staatlicher Kontrolle. Seit dem 18. Jahrhundert bezeichnen wir diese systematische Sammlung demographischer und ökonomischer Daten speziell durch den Staat mit dem Terminus Statistik. Staatliche Verwaltung, die Konstruktion eines Staates und das Regieren als solches sind eng mit dem Erheben von Daten verbunden (Desrosières 1998), sodass sich die Geschichte von Nationalstaaten zugleich als Geschichte der Statistik und damit als eine Geschichte des Sammelns und Verarbeitens von Daten lesen lässt. Wissen und Macht gehören hier eng zusammen.

Big Data soll als sogenannte *evidenzbasierte Politik* zunehmend Anwendung finden. Ist dieses Vorgehen aber überhaupt epistemisch, politisch und ethisch gerechtfertigt? Die bereits angesprochenen ethischen Fragen um Privatsphäre, Datenschutz, Schutz vor Diskriminierung, Herstellung von Gleichheit, Autonomie, usw., sind beim Einsatz von Big Data in der Politik zwingend zu berücksichtigen. All diese Fragen sind bislang ungelöst. Die Rechtfertigung von Big Data in der Politik ist epistemisch jedoch bereits aufgrund des beschränkten Zugangs zu Daten (bzw. Algorithmen oder Software) und der mangelnden Kompetenz in kritischer Datenanalyse schwierig. Diejenigen, die auf der Basis von Datenanalyse politische Entscheidungen treffen, müssen in die Lage versetzt werden, sie zu verstehen, d. h. es muss zunächst einmal die Wissensgrundlage geschaffen sein, um einschätzen zu können: Wie ist die Qualität der Daten, der Berechnungen, der Algorithmen, der Systeme? Liefern sie tatsächlich zuverlässige Prognosen? Sind die Daten tatsächlich die argumentativ beste Grundlage, auf der ein politischer Entscheidungsprozess beruhen sollte? Auch die Rhetorik von Big Data als neutraler theorie- und verzerrungsfreier Wissenschaft erschwert die kritische Analyse, weil sie ermöglicht, dass Politiker.innen sich auf diese scheinbar unbestechlichen Zahlen berufen und damit die eigenen unliebsamen Entscheidungen begründen, sie so aber zugleich als nicht selbstgetroffen von sich weisen können.

Big Data: (Un)sichtbare Systeme?

Um die Schwierigkeit der epistemischen Rechtfertigung von Big Data zu verstehen, ist ein Blick auf die Intransparenz von Big-Data-Praktiken hilfreich. Unterscheiden lassen sich zwei Arten: Die funktionale und die epistemische Intransparenz. Die funktionale umfasst den eingeschränkten Zugang zu Systemen, Daten, Algorithmen, etc., insbesondere von proprietären Systemen, die nicht Open Source sind und als solche ein großes Problem darstellen, denn wie lässt sich funktionale Intransparenz auflösen, wenn Geschäftsgeheimnisse als Wettbewerbsvorteil gelten und daher ein Unternehmen nicht gezwungen werden kann, Daten und Algorithmen offenzulegen? Möglich wäre dies nur durch neue Geschäftsmodelle, die Big Data anders regulieren. Hierzu

gibt es derzeit drei Argumentationslinien: das Verteidigen des Status Quo in Berufung auf die Selbstregulierung des Marktes, die Forderung nach verbesserter Regulation durch effektivere Regulierungsmöglichkeiten großer Internetfirmen und drittens den Ansatz, dass sich die ökonomischen Grundlagen von Datenmärkten verändern müssen. Dieser Ansatz folgt u. a. aus der Erkenntnis, dass oft nicht mehr die Qualität der Algorithmen, sondern die Menge der Daten über die Qualität der Auswertung entscheidet und damit zu einem entscheidenden Wettbewerbsvorteil wird, der nicht durch bessere Algorithmen oder andere Maßnahmen ausgeglichen werden kann. Bereits allein aus ökonomischer Perspektive ist das höchst problematisch, denn so entsteht ein großes Ungleichgewicht zwischen den Unternehmen eines Marktes, sodass sich die Frage stellt, ob nicht längst eine Monopolbildung den freien Markt verunmöglicht hat. Auflösen lässt sich dies nur durch etwa alternative Infrastrukturen oder Konzepte, bei denen die Daten nicht mehr den Unternehmen gehören. Hier gibt es im Wesentlichen zwei Richtungen: Gehören Daten allein dem Nutzer oder sind sie als öffentliches Gut zu verstehen? Alexey Morosov etwa ist einer der Vertreter, die die Auffassung von Daten als Gemeingut vertreten, nach der Daten demnach gar nicht gehandelt werden können (was allerdings für den Schutz und die Sicherheit der Daten neue Herausforderungen mit sich bringt). Jaron Lanier nach sollten die Daten dem Nutzer gehören, der dann an Gewinnen auch direkt partizipieren kann.

Die zweite Art der Intransparenz nun, die epistemische, entsteht durch die Grenzen jedes Einzelnen, komplexe Systeme und Verfahren verstehen zu können. Sie ist damit relativ zur Kompetenz der Nutzer.innen: Je mehr ich über Datenanalyse weiß, desto mehr kann ich verstehen. Das Informed-consent-Verfahren ist in dieser Hinsicht ein für die meisten Nutzer.innen intransparentes Vorgehen, denn auch wenn offengelegt wird, was mit den eigenen Daten geschieht, d. h., wenn ich über die *Terms of Agreement* informiert werde, bevor ich sie unterschreibe, ist dies nur insofern transparent, als dass das Vorgehen nur bis zu einem gewissen Grad tatsächlich offengelegt wird. In den seltensten Fällen haben Nutzer als Nicht-Experten ausreichende Handlungskompetenz, um das Gelesene wirklich zu verstehen. Lösungsansätze hierzu sind Konzepte wie *Privacy* oder *Transparency by Design*, das heißt im Sinne des Vortrags von Leon Hempel werden die Dinge *transparent to use*: Wir schützen die Nutzer als Default. Für Expert.innen stellen sich in Bezug auf die epistemische Intransparenz zudem weitere Fragen: Wo sind Erkenntnisgrenzen mit welchen Auswirkungen, wenn ich verstehen will, wie komplexe Systeme im Kontext von z. B. Maschinenlernen oder neuronalen Netzen funktionieren? MacKenzie z. B. hat in *Mechanizing Proof* bereits 2001 danach gefragt, ob ein Beweis durch Computerberechnungen formal ein Beweis ist, wenn ihn keiner der Experten verstehen kann, denn im Mathematischen beruht ein Beweis darauf, dass er epistemisch nachvollziehbar ist. Auch im Vortrag von Corinna Bath ging es um die Frage, wo sich auf Expertenseite die Grenzen des Verständnisses verschieben lassen und inwiefern sich Verzerrungen und etwa daraus abgeleitete Diskriminierung durch Software nachweisen und visualisieren lassen.

Big Data: Was nun?

Das Problem der *Un.Sichtbarkeit* als einerseits funktionales, andererseits epistemisches System verlangt demzufolge nach sehr unterschiedlichen Lösungen – einerseits nach grundlegenden ökonomischen Veränderungen, andererseits nach dem Auflösen oder Verschieben der Grenzen der Verstehbarkeit von Big-Data-Praktiken für Nicht-Expert.innen wie für Expert.innen. Hierzu müssen wir uns von einem reinen Big Data for Governance hinwenden zur Notwendigkeit der Governance for Big Data. Es stellen sich Fragen wie: Wie erhöht man die Vertrauenswürdigkeit von Datenpraktiken? Durch welche technischen und legalen Verfahren lässt sich epistemische, politische und ethische Überprüfbarkeit unterstützen? Big Data Governance muss als eine Kombination gedacht werden aus rechtlichen Lösungen (Regeln der Hard Law), Selbstverpflichtungen (Soft Law), Governance by Design (die Bemühungen von Privacy by Design oder Transparency by Design) und auch einem Umdenken in der Bildung. Keine dieser Säulen wird allein ausreichen, um die bislang etablierte Praxis und derzeit größte Krux der In.Transparenz abzulösen: Die Bürde des Informiertseins auf die Schultern von Nutzer und Nutzerin zu legen.

Referenzen

Boyd, D./Crawford, K. (2012): Critical questions for big data – provocations for a cultural, technological, and scholarly phenomenon. In Information, Communication & Society 15 (5) 662–679

Manovich, L. (2011): Trending: the promises and the challenges of big social data. In M. K. Gold (ed.): Debates in the digital humanities. The University of Minnesota Press

Desrosières, A. (1998): The politics of large numbers – a history of statistical reasoning. Cambridge: Harvard University Press

Judith Simon

Judith Simon ist Associate Professorin für Wissenschaftstheorie und Technikphilosophie an der IT University of Copenhagen. Sie ist zudem Editorin der Journals *Philosophy & Technology* und *Big Data & Society* sowie Vorstandsmitglied in der *International Association of Computing and Philosophy* (IACAP) sowie der *International Society for Ethics and Information Technology* (INSEIT).