Viktor Mayer-Schönberger | Kenneth Cukier

Big Data

Die Revolution, die unser Leben verändern wird

Übersetzung aus dem Englischen
von Dagmar Mallett

REDLINE | VERLAG

Kapitel 1

Heute

Im Jahr 2009 wurde ein neues Grippevirus entdeckt. Diese neue, als H1N1 bezeichnete Variante kombinierte Elemente des Vogelgrippe- und Schweinegrippevirus und breitete sich rasch aus. Schon nach wenigen Wochen warnten die Gesundheitsbehörden weltweit vor einer möglichen Pandemie. Einige Stimmen befürchteten eine der Spanischen Grippe von 1918 vergleichbare Seuche; damals hatten sich eine halbe Milliarde Menschen angesteckt, von denen einige Dutzend Millionen gestorben waren. Schlimmer noch war, dass vorerst kein Impfstoff gegen das neue Virus zur Verfügung stand. Die Gesundheitsbehörden konnten nur darauf setzen, die Ausbreitung der Seuche möglichst zu verlangsamen. Dazu musste man allerdings zunächst einmal das Ausmaß der Ausbreitung erfahren.

Die Centers for Disease Control and Prevention (CDC), die amerikanische Seuchenbekämpfungsbehörde, führte eine Meldepflicht der Ärzte für neue Grippefälle ein. Allerdings war das so gewonnene Bild des Fortschreitens der Epidemie immer um eine oder zwei Wochen veraltet, da die meisten Menschen nicht sofort zum Arzt gehen, wenn sie sich krank fühlen. Auch die Übermittlung der Meldungen an die Zentralstellen dauerte ihre Zeit, und die CDC fasste die Zahlen nur einmal wöchentlich zusammen. Bei einer sich rasch ausbreitenden Epidemie sind zwei Wochen Zeitverzug eine Ewigkeit. Diese Verspätung machte die Gesundheitsbehörden im entscheidenden Zeitraum praktisch blind.

Zufällig nur wenige Wochen davor hatten Software-Entwickler des Internetriesen Google in der Fachzeitschrift *Nature* einen interessanten Aufsatz veröffentlicht.[1] Bei Gesundheitsbehörden und Computerexperten sorgte er durchaus für Aufsehen, blieb aber in der Öffentlichkeit ziemlich unbemerkt. Die Autoren erklärten darin, wie die Suchmaschine Google die Ausbreitung der jährlichen Grippeepidemie in den USA »voraussagen« könne, und zwar nicht nur landesweit, sondern auch regional und sogar für die einzelnen Bundesstaaten. Das Unternehmen wertete dazu die Suchanfragen seiner Kunden im Internet aus. Weil die Suchmaschine täglich über drei Milliarden solcher Anfragen erhält und sie alle speichert, stand genug Datenmaterial zur Verfügung.

Google verglich die 50 Millionen am häufigsten von US-Bürgern eingegebenen Suchbegriffe mit den Daten der CDC zur Ausbreitung der jährlichen Grippeepidemien von 2003 bis 2008, um so eine Korrelation zwischen Suchanfragen und befallenen Gebieten zu ermitteln. Das hatten auch andere schon versucht, aber niemand verfügte über so viele Daten, so viel Rechnerleistung und so großes statistisches Können wie Google.

Die Google-Mitarbeiter vermuteten dabei zwar durchaus, dass es sich bei diesen Suchanfragen um grippespezifische Begriffe handeln könne – etwa »Medikamente gegen Husten und Fieber« –, doch war weder der Inhalt der Begriffe tatsächlich von Bedeutung, noch beruhte das entwickelte System darauf. Das System suchte stattdessen nur nach Korrelationen zwischen der Häufigkeit bestimmter Suchbegriffe und der Ausbreitung der Grippewelle über Zeit und Raum. Insgesamt wurde die enorme Zahl von 450 Millionen unterschiedlicher mathematischer Modelle auf ihre Tauglichkeit geprüft, wobei jeweils die Voraussagen mit den tatsächlichen Grippedaten der CDC von 2007 und 2008 verglichen wurden. Und so fanden die Google-Entwickler tatsächlich das richtige Modell, das bei 45 Suchbegriffen eine starke Korrelation zwischen der darauf basierenden Grippevorhersage und den amtlichen landesweiten Zahlen zur Verbreitung der Epidemie aufwies. Google konnte damit die Ausbreitung der Grippe genauso gut wie die CDC feststellen, aber nicht mit ein oder zwei Wochen Verspätung, sondern praktisch unmittelbar.[2]

Während der H1N1-Krise des Jahres 2009 erwies sich das Google-System daher als nützlicherer und schnellerer Indikator als die Regierungsstatistiken mit ihren unvermeidlichen Verzögerungen. Die Gesundheitsbehörden gewannen einen wertvollen Informationsvorsprung im Kampf gegen die Seuche.

Neu ist dabei, dass Google keine Gewebeproben einsammelt oder Berichte von Hausärzten auswertet. Stattdessen beruht die Methode auf »Big Data« – der Fähigkeit, Informationen so zu nutzen, dass neue Erkenntnisse, Güter oder Dienstleistungen von bedeutendem Wert gewonnen werden. Mit dieser Methode verfügt die Menschheit über ein neues Instrument, um im Falle einer Pandemie die Ausbreitung vorauszusagen und damit zu verhindern.

Das Gesundheitswesen ist nur ein Bereich, in dem Big Data große Auswirkungen hat. Ganze Geschäftsfelder werden durch Big Data ebenfalls grundlegend verändert. Ein gutes Beispiel dafür ist der Kauf von Flugtickets.

Oren Etzioni flog 2003 zur Hochzeit seines jüngeren Bruders von Seattle nach Los Angeles. Bereits Monate vor dem Termin kaufte er das Ticket im Internet, in der Annahme, dass es umso günstiger sei, je früher man den Flug buche. Während des Fluges wurde er dann neugierig und fragte seinen Nachbarn, wie viel dieser für sein Ticket bezahlt und wann er es gekauft habe. Wie sich herausstellte, hatte der trotz späterer Buchung seinen Flugschein viel günstiger als Etzioni bekommen. Der verärgerte Etzioni befragte noch andere Fluggäste, und tatsächlich hatten die meisten weniger bezahlt als er.

Bei den meisten Menschen hätte sich das unangenehme Gefühl, verloren zu haben, wieder gelegt, bevor es an der Zeit war, die Tische hochzuklappen und die Sitzlehnen senkrecht zu stellen. Nicht so bei Etzioni, einem der führenden Informatiker der USA. Für ihn ist die Welt eine Reihe von Big-Data-Problemen – also lösbaren Herausforderungen. Und er löst derartige Probleme, seit er 1986 in Harvard seinen Abschluss als erster Student im Hauptfach Informatik gemacht hat.

Von seinem Lehrstuhl an der Universität von Washington aus hatte er bereits eine ganze Reihe von Big-Data-Unternehmen gegründet, bevor der Begriff »Big Data« überhaupt aufkam. Er half bei der Gründung von MetaCrawler, einer der ersten Internet-Suchmaschinen, die rasch wuchs und 1994 von InfoSpace aufgekauft wurde. Er war Mitbegründer von Netbot, der ersten großen Versandhandel-Vergleichsseite, die er an Excite weiterverkaufte. Sein Start-up-Unternehmen ClearForest, das den Sinngehalt von Textdokumenten erschließt, wurde von Reuters angekauft.

Nach der Landung suchte Etzioni nach einer Methode, mit der sich feststellen ließe, ob der Internet-Buchungspreis für einen Flug jeweils wirklich günstig ist. Ein Flugticket ist eine Handelsware, und jeder Sitz eines gegebenen Fluges hat den gleichen Wert. Dennoch gibt es große Preisunterschiede zwischen ihnen, die von einer Vielzahl von Faktoren bestimmt werden, die nur den Fluglinien selbst bekannt sind.

Etzioni kam zu dem Schluss, dass er die Gründe für die Preisunterschiede gar nicht verstehen musste, sondern dass es genügte, wenn er voraussagen konnte, ob der aktuelle Ticketpreis voraussichtlich in den nächsten Tagen und Wochen steigen oder sinken werde. Das ist zwar nicht einfach, aber möglich. Dazu müssen die Ticketpreise für eine bestimmte Route analysiert werden, und zwar für jeden Tag vor dem Flug.

Wenn der durchschnittliche Preis im Laufe der Zeit eher sank, war es sinnvoll, das Ticket erst kurz vor dem Flug zu kaufen; stieg er dagegen eher an, sollte das System die sofortige Buchung empfehlen. Mit anderen Worten, erforderlich war eine ausgereifte Version der spontanen Umfrage, die Etzioni im Flugzeug durchgeführt hatte. Das stellte zwar ein komplexes Informatikproblem dar, aber auch dieses wollte er lösen. Also machte er sich an die Arbeit.

Mithilfe von 12.000 Preisangaben für Tickets, die Etzioni im Verlauf von 41 Tagen durch Beobachtung einer Reise-Webseite gewann, schuf er ein Vorhersagemodell, das den simulierten Passagieren eine Menge Geld sparte. Das Modell verstand nicht *warum*, sondern nur *was*. Das

heißt, es kannte nicht die Variablen, die für die Preisgestaltung eines Flugtickets herangezogen werden, zum Beispiel die Zahl der unverkauften Sitzplätze, die Saison oder das Wochenend-Sonderangebot. Es gründete seine Voraussage auf das, was es wusste: Wahrscheinlichkeiten, die es aus den Daten anderer Flüge gewonnen hatte. »Kaufen oder nicht kaufen, das ist hier die Frage«, sagte sich Etzioni und nannte sein Forschungsprojekt »Hamlet«.

Das kleine Projekt entwickelte sich zu einem risikokapitalgestützten Start-up-Unternehmen namens Farecast. Die Software ermöglichte es den Kunden durch die Vorhersage, ob der Preis für ein Flugticket steigen oder fallen würde, eine begründete Entscheidung zu treffen, wann sie am besten den »Buchen«-Button anklicken sollten. Sie erhielten Zugang zu Informationen, die ihnen vorher verschlossen gewesen waren. Auch die Entscheidungskompetenz von Farecast konnte jeder Kunde selbst beurteilen, da es transparent handelte, denn Farecast gab an, wie viel Vertrauen es selbst in seine konkrete Vorhersage hatte.

Um zu funktionieren, brauchte dieses Programm eine große Datenmenge. Etzioni verschaffte sich daher Zugang zu einer Flugreservierungs-Datenbank der Luftfahrtgesellschaften. Mit diesen Informationen konnte das System Vorhersagen auf der Basis sämtlicher Buchungen für die meisten Routen auf amerikanischen Linienflügen über das ganze Jahr hinweg machen. Farecast verarbeitete so fast 200 Milliarden Flugpreisdaten für seine Voraussagen. Seine Anwender konnten dadurch viel Geld sparen.[3]

Mit seinem sandfarbenen Haar, dem breiten Grinsen und dem unschuldig-gesunden Äußeren wirkt Etzioni kaum wie ein Mensch, der die Fluglinien um Millionen Dollar potenzieller Einkünfte bringt. Dabei hatte er sogar noch mehr vor. Im Jahre 2008 plante er, dieselbe Methode auch auf andere Bereiche wie Hotelbuchungen, Konzertkarten und Gebrauchtwagen anzuwenden: Sie sollte bei allen Verkaufsvorgängen mit geringen Produkt-, aber hohen Preisunterschieden und einer großen Datenmenge funktionieren. Bevor aber Etzioni diese Pläne weiterverfolgen konnte, klopfte Microsoft an seine Tür, kaufte Farecast für

110 Millionen Dollar auf und integrierte das Programm in die Suchmaschine Bing.[4] Im Jahr 2012 sagte das System in 75 Prozent der Fälle richtig voraus und sparte Ticketkäufern durchschnittlich 50 Dollar.

Farecast ist ein ausgezeichnetes Beispiel für eine Big-Data-Firma und ein Ausblick in die Zukunft. Noch vor fünf oder zehn Jahren hätte Etzioni ein solches Unternehmen nicht gründen können. »Es wäre unmöglich gewesen«, sagt er, denn die erforderliche Rechenleistung und Speicherkapazität wären zu teuer gewesen. Aber obwohl die technologischen Fortschritte ein entscheidender Faktor waren, hatte sich noch etwas anderes, nicht so Auffälliges verändert: die Einstellung zum Gebrauch von Daten.

Plötzlich wurden Daten interessant und waren nicht länger statisch oder langweilig und überflüssig, sobald ihr unmittelbarer Zweck erfüllt war, zum Beispiel nach der Landung des Flugzeugs oder der Beantwortung der Suchanfrage. Sie waren zum Rohmaterial für Transaktionen geworden und zu einem wichtigen wirtschaftlichen Input, mit dem neue ökonomische Werte geschaffen wurden. Mit der richtigen Einstellung kann man Daten so geschickt wiederverwenden, dass sie zu einer Quelle für Innovationen und neuartige Dienstleistungen werden. Diese Daten eröffnen den Zugang zu verblüffenden Einsichten, wenn man bereit ist, sich auf sie einzulassen, und über die richtigen Werkzeuge verfügt, um sie zu analysieren.

Daten sprechen lassen

Die Auswirkungen der Informationsgesellschaft sind nicht zu übersehen, mittlerweile trägt fast jeder ein Mobiltelefon in der Tasche, einen Laptop im Rucksack und im Büro stehen Desktop-Computer. Die Information selbst ist allerdings weniger augenfällig. Und trotzdem: Ein halbes Jahrhundert nach dem Einzug des Computers in die Gesellschaft haben sich so viele Daten angesammelt, dass sich nun etwas Neues und Besonderes anbahnt. Die Welt ist so voll von Information wie nie zuvor und auch die Informationsmenge nimmt immer schneller zu. Die-

se quantitativen Veränderungen haben zu einer qualitativen Veränderung geführt. In den Naturwissenschaften, etwa der Astronomie oder der Genetik, wo diese Datenexplosion um die Jahrtausendwende zuerst sichtbar wurde, entstand der Begriff »Big Data«. Dieses Konzept breitet sich jetzt auf alle menschlichen Tätigkeitsfelder aus.

Es gibt für Big Data keine exakte Definition.[5] Ursprünglich verstand man darunter eine Informationsmenge, die zu groß für den Arbeitsspeicher des verarbeitenden Computers geworden war und von den Entwicklern neue Technologien verlangte. Das war der Anlass für technologische Innovationen wie Google MapReduce und sein Open-Source-Pendant Hadoop von Yahoo. Mit diesen neuen Werkzeugen konnten sehr viel größere Datenmengen als zuvor verarbeitet werden, und zwar – das war ein entscheidender Fortschritt – nicht nur dann, wenn sie bereits sauber in klassischen Datenbanken strukturiert und in einem einheitlichen Format zusammengefasst waren. Neue Werkzeuge, die sich noch weiter von den bislang notwendigen strengen Hierarchien und Strukturen verabschieden, sind gerade im Entstehen. Gleichzeitig kristallisierten sich die Internetanbieter als führende Anwender dieser neuen Werkzeuge heraus, weil sie über die größten Datenmengen verfügten und ein brennendes finanzielles Interesse hatten, aus den gesammelten Daten möglichst großen Nutzen zu ziehen. So überrundeten sie traditionelle Firmen, die mitunter schon mehrere Jahrzehnte Erfahrung auf diesem Gebiet besaßen.

Man kann sich die Sache so vorstellen, wie wir es in diesem Buch tun wollen: Big Data ist das, was man in großem, aber nicht in kleinem Maßstab tun kann, um neue Erkenntnisse zu gewinnen oder neue Werte zu schaffen, sodass sich Märkte, Organisationen, die Beziehungen zwischen Bürger und Staat und vieles mehr verändern.

Aber das ist nur der Anfang. Die Ära von Big Data wird sich auch auf unsere Lebensweise und unsere Weltsicht auswirken. Vor allem muss die Gesellschaft sich gewohnter Vorstellungen von Kausalität entledigen und stattdessen vermehrt auf Korrelationen verlassen: Man wird oft nicht mehr wissen *warum,* sondern nur noch *was.* Das ist das En-

de jahrhundertelang eingeführter Prozesse und verändert tiefgreifend die Art, wie wir Entscheidungen treffen und die Wirklichkeit verstehen.

Big Data steht am Anfang einer grundlegenden Umwälzung. Wie so viele neue Technologien, wird Big Data ganz sicher auch der Übertreibungsmaschinerie von Silicon Valley zum Opfer fallen: Erst kommen die begeisterten Zeitschriftentitel und Branchentagungen, dann ebbt der Trend wieder ab, und der Glanz vieler Start-up-Gründungen verblasst. Aber sowohl das Hochjubeln wie das anschließende Verdammen verkennen, dass hier etwas Bedeutendes passiert. Genau wie die Erfindung des Fernrohrs das Verständnis des Kosmos und die Erfindung des Mikroskops die Entdeckung der Mikroben ermöglichten, werden uns die neuen Datensammlungs- und Datenanalyse-Werkzeuge in großem Stil dabei helfen, die Welt auf eine Weise neu zu verstehen, die wir erst erahnen können. In diesem Buch sind wir weniger die Evangelisten von Big Data als vielmehr seine Herolde. Die wirkliche Revolution, das sei noch einmal gesagt, findet nicht in der Technik statt, sondern in den Daten selbst und in der Art ihrer Analyse.

Um zu verstehen, wie weit diese Informationsrevolution bereits fortgeschritten ist, lassen Sie uns einen Blick auf einige Trends aus allen gesellschaftlichen Bereichen werfen. Unser digitales Universum expandiert ständig. Nehmen wir nur die Astronomie. Als im Jahr 2000 das Projekt zur Himmelskartografie Sloan Digital Sky Survey startete, sammelte das damit befasste Teleskop in Neu-Mexiko in den ersten Wochen bereits mehr Daten, als in der gesamten bisherigen Geschichte der Astronomie gesammelt worden waren. Im Jahr 2010 verfügte das Archiv des Survey bereits über 140 Terabyte an Informationen. Ein geplanter Nachfolger, das Large Synoptic Survey Telescope in Chile, soll 2016 in Betrieb gehen und alle fünf Tage dieselbe ungeheure Datenmenge sammeln.

Solche astronomischen Zahlen findet man aber nicht nur draußen im Weltall. Als 2003 das menschliche Genom entschlüsselt wurde, stand dahinter ein Jahrzehnt intensiver Arbeit, um die drei Milliarden Basenpaare vollständig zu sequenzieren.[6] Heute, ein weiteres Jahrzehnt spä-

ter, kann ein einziges Labor dieselbe Menge DNA an einem Tag sequenzieren. Auf den amerikanischen Finanzmärkten wechseln jeden Tag etwa sieben Milliarden Aktien den Besitzer, etwa zwei Drittel davon aufgrund von Computeralgorithmen, die auf mathematischen Modellen beruhen und Berge von Daten verarbeiten, um Gewinne vorherzusagen und Verluste zu minimieren.[7]

Internetunternehmen spüren die Datenflut besonders stark. Allein Google sammelt pro Tag 24 Petabyte an Daten, ungefähr tausendmal so viel wie alle gedruckten Werke in der US-Kongressbibliothek zusammen.[8] Facebook, ein Unternehmen, das es vor einem Jahrzehnt noch gar nicht gab, erhält pro Stunde über zehn Millionen neuer Fotos. Facebook-Nutzer geben pro Tag etwa drei Milliarden Kommentare oder »Gefällt mir«-Klicks ab; die digitale Spur, die sie so hinterlassen, kann der Konzern auswerten, um die Vorlieben der einzelnen Kunden zu erfassen.[9] Die 800 Millionen monatlichen Nutzer des Google-Videodienstes YouTube laden pro Sekunde eine Stunde Videos hoch.[10] Die Anzahl der Twitter-Kurznachrichten wächst jährlich um 200 Prozent und lag 2012 bei über 400 Millionen Tweets pro Tag.[11]

Ob in der Naturwissenschaft oder dem Gesundheitswesen, auf dem Börsenparkett oder im Internet – alle diese so verschiedenen Bereiche erzählen dieselbe Geschichte: Weltweit steigt die Datenmenge so rasend schnell, dass sie nicht nur die Verarbeitungskapazitäten der Rechner, sondern auch unser Vorstellungsvermögen übersteigt.

Oft wurde versucht, eine konkrete Zahl für die uns umgebende Informationsmenge zu ermitteln und festzustellen, wie schnell sie wirklich wächst. Diese Projekte waren unterschiedlich erfolgreich, da sie sich verschiedener Messverfahren bedienten. Eine der umfassendsten Studien stammt von Martin Hilbert, der an der Annenberg School for Communication and Journalism der University of Southern California lehrt. Hilbert versucht wirklich alles zu erfassen, was erzeugt, gespeichert und übermittelt wird – nicht nur Bücher, Gemälde, E-Mails, Fotografien, Musikstücke und Videos (analoge und digitale), sondern auch Videospiele, Telefonanrufe und sogar Auto-Navigationssysteme

und traditionelle Briefe. Ebenfalls inbegriffen sind Radio- und Fernsehsendungen, die je nach Zuhörer- und Zuschauerzahl eingeordnet werden.

Hilbert kommt zu dem Ergebnis, dass 2007 über 300 Exabyte gespeicherter Daten existierten.[12] Übertragen auf verständlichere Maßstäbe bedeutet das: Ein Kinofilm in digitaler Form kann zu einer Datei von etwa einem Gigabyte komprimiert werden. Ein Exabyte ist eine Milliarde Gigabyte. Kurz gesagt: eine ganz schöne Menge. Interessanterweise waren 2007 nur mehr etwa 7 Prozent der Daten in analoger Form gespeichert (Druckerzeugnisse, Papierbilder und so weiter); der Rest digital. Es ist noch gar nicht lange her, dass dieses Verhältnis ganz anders aussah. Zwar spricht man schon seit den 1960er Jahren von der »Infomationsrevolution« und dem »digitalen Zeitalter«, aber beides beginnt gerade erst Realität zu werden. Noch im Jahr 2000 war nur ein Viertel der weltweit gespeicherten Informationen digital – die anderen drei Viertel befanden sich auf Papier, Film, Vinyl-LPs, Audiokassetten et cetera.

Die Gesamtmenge digitaler Information war damals nicht sehr groß – was diejenigen, die schon seit Jahrzehnten im Web surfen und ihre Bücher online kaufen, nachdenklich machen sollte. (Noch 1986 wurden etwa 40 Prozent der nichtspezialisierten Computerkapazität der Welt von Taschenrechnern repräsentiert, die damals mehr Rechenleistung als alle PCs zusammen hatten.) Weil der digitale Datenberg aber so rasch wächst – laut Hilbert verdoppelt er sich jeweils in weniger als drei Jahren –, kehrte sich dieses Verhältnis bald um. Für 2013 wird die Gesamtmenge gespeicherter Informationen auf 1.200 Exabyte geschätzt, und weniger als 2 Prozent davon sind nicht digital.[13]

Man kann sich nicht wirklich vorstellen, was diese Datenmenge eigentlich bedeutet. Würde man alle diese Daten ausdrucken und zu Büchern binden, bedeckten sie die gesamte Landfläche der USA in 52 Schichten. Würde man die Datenmenge auf CD-ROM brennen, könnte man mit diesen CDs fünf Stapel bis zum Mond errichten. Im dritten Jahrhundert v. Chr. versuchte der ägyptische König Ptolemaios II., von jedem da-

mals vorhandenen Buch ein Exemplar für seine große Bibliothek in Alexandria zu bekommen, die damals das Wissen der gesamten Welt in sich vereinte. Die digitale Sintflut von heute bedeutet, dass auf jeden einzelnen lebenden Menschen 320 solcher Bibliotheken im Umfang derjenigen des antiken Alexandria kämen.

Die Entwicklung beschleunigt sich weiter. Die gespeicherte Informationsmenge wächst viermal rascher als die Weltwirtschaft, die Rechenleistung von Computern sogar neunmal schneller. Kein Wunder, dass man überall Klagen über die Informationsflut hört. Die Veränderungen treiben uns vor sich her.

Versuchen wir also ein wenig Distanz zu gewinnen, indem wir die gegenwärtige mit einer früheren Informationsrevolution vergleichen, die aus der Erfindung der Druckerpresse durch Gutenberg 1439 folgte. Die Historikerin Elizabeth Eisenstein ermittelte, dass in den 50 Jahren zwischen 1453 und 1503 etwa acht Millionen Bücher gedruckt wurden.[14] Das sind mehr, als die Schreiber in ganz Europa seit der Gründung von Byzanz etwa 1.200 Jahre zuvor handschriftlich vervielfältigt hatten. Mit anderen Worten: Damals dauerte es 50 Jahre, bis die Informationsmenge sich in Europa verdoppelt hatte, im Gegensatz dazu braucht es dafür heute nur noch drei Jahre.

Was bedeutet diese Zunahme? Peter Norvig, Experte für künstliche Intelligenz bei Google, zieht zum Vergleich gerne Bilder heran.[15] Stellen Sie sich als Erstes die bekannte Abbildung eines Pferdes aus den altsteinzeitlichen Höhlenmalereien von Lascaux in Frankreich vor, die etwa 17.000 Jahre alt sind. Dann denken Sie an die Fotografie eines Pferdes – oder besser noch, an Pablo Picassos Farbtupfer, die den Höhlenmalereien gar nicht so unähnlich sind. Als Picasso die Zeichnungen von Lascaux sah, soll er bemerkt haben: »Wir haben seitdem nichts Neues erfunden.«[16]

Einerseits hat Picasso recht, andererseits irrt er aber. Denn es dauert sehr lange, ein Pferd zu zeichnen; fotografiert ist es sehr viel schneller. Dies stellt zwar eine Veränderung dar, aber noch keine grundlegende,

da das Ergebnis immer noch dasselbe ist: die Abbildung eines Pferdes. Stellen Sie sich dagegen, so Norvig, 24 Bilder eines Pferdes pro Sekunde vor. Hier führt die quantitative zu einer qualitativen Veränderung, denn ein Film ist etwas grundlegend anderes als ein fotografisches Standbild. So ist es auch mit Big Data: Indem wir die Menge verändern, verändern wir das Wesen der Aufzeichnung.

Eine Analogie findet sich in der Nanotechnologie, wo alles immer kleiner anstatt immer größer wird. Nanotechnologie beruht auf dem Prinzip, dass sich die physikalischen Eigenschaften auf der molekularen Ebene verändern. Wenn man diese neuen Eigenschaften ausnutzt, kann man Materialien mit ganz neuen Fähigkeiten entwickeln. Im Nanobereich sind zum Beispiel sehr flexible Metalle und sogar dehnbare Keramik möglich. Genauso können wir durch eine Veränderung der Größenordnung der Daten, mit denen wir arbeiten, Ergebnisse erzielen, die im kleineren Maßstab unmöglich wären.

Manchmal sind die Einschränkungen unserer Existenz, die wir als allgemeingültig voraussetzen, lediglich Funktionen der Größenordnungen, in denen wir uns bewegen. Hier ein dritter Vergleich, wieder ein naturwissenschaftlicher. Das wichtigste Naturgesetz in unserem Alltag ist das der Gravitation: Die Schwerkraft bestimmt alles, was wir tun. Für manch winziges Insekt dagegen spielt sie kaum eine Rolle. Für den Wasserläufer zum Beispiel ist die Oberflächenspannung des Wassers eine weit stärkere und seine Lebensweise bestimmende Kraft.

Wie in der Physik kommt es auch bei der Information auf die Größenordnung an. Deshalb kann Google die Verbreitung einer Grippewelle genauso gut vorhersagen wie die offizielle Statistik, die auf tatsächlich stattgefundenen Arztbesuchen beruht. Das ist möglich, weil Hunderte Milliarden Suchanfragen durchkämmt werden – und die Antwort kommt fast unmittelbar und weit schneller als aus amtlichen Quellen. Ebenso kann Etzionis Farecast die Preisentwicklung eines Flugtickets voraussagen und gibt so dem Verbraucher beträchtliche ökonomische Macht in die Hand. Das funktioniert aber in beiden Fällen nur, weil Hunderte Milliarden Datensätze verarbeitet werden.

Diese beiden Beispiele zeigen die wissenschaftliche und die gesellschaftliche Bedeutung von Big Data sowie das Ausmaß der damit möglichen ökonomischen Wertschöpfung. Dies sind die beiden Felder, auf denen Big Data die Welt verändern wird – vom Geschäftsleben über das Finanzwesen, das Gesundheitswesen, die Verwaltung, das Bildungssystem und die Wirtschaft bis hin zu den Geisteswissenschaften und jedem anderen gesellschaftlichen Bereich.

Obwohl wir uns erst am Anfang des Big-Data-Zeitalters befinden, sind wir doch bereits täglich darauf angewiesen. Spam-Filter zum Beispiel passen sich automatisch an, wenn die Merkmale unerwünschter E-Mails sich verändern: Man könnte die Software gar nicht auf die Erkennung jeder einzelnen Variante verräterischer Begriffe, wie etwa »via6ra«, programmieren. Partnerschaftsvermittlungs-Webseiten stellen Paare auf der Basis der Übereinstimmung ihrer angegebenen Eigenschaften mit denen früherer erfolgreicher Vermittlungen zusammen. Die »Autocorrect«-Funktion in Smartphones verfolgt unsere Eingaben und fügt ihrem internen Wörterbuch entsprechend neue Einträge hinzu. Aber diese Anwendungen sind erst der Anfang. Von Autos, die selbsttätig ausweichen oder bremsen, bis hin zum Watson-Computer von IBM, der in der Quizsendung *Jeopardy!* menschliche Kandidaten schlägt, wird Big Data viele Aspekte unserer Lebenswelt verändern.

Im Grunde geht es dabei um Vorhersagen. Das gilt zwar als Bereich der künstlichen Intelligenz innerhalb der Informatik, genauer gesagt eines Spezialgebietes namens maschinelles Lernen, aber diese Einordnung ist eigentlich irreführend. Big Data versucht nicht, einem Rechner »beizubringen« wie ein Mensch zu »denken«. Vielmehr geht es um die mathematische Verarbeitung riesiger Datenmengen zur Gewinnung von Wahrscheinlichkeiten: der Wahrscheinlichkeit, dass eine E-Mail Spam ist, dass die eingetippte Zeichenfolge »dei« in Wirklichkeit »die« lauten soll, dass Weg und Geschwindigkeit eines Passanten, der die Straße überquert, keine Gefahr bedeuten und das selbstgesteuerte Auto nur leicht bremsen muss. Wesentlich ist, dass alle diese Systeme ihre Vorhersagen auf der Grundlage ungeheuer großer Datenmengen machen.

Außerdem sind sie so konzipiert, dass sie sich immer weiter selbst verbessern, indem sie darauf achten, welche Eingaben und Datenmuster zu den besten Vorhersagen führen.

In Zukunft – und zwar schneller, als wir glauben – werden viele Elemente unserer Lebenswelt, die bis jetzt der menschlichen Beurteilung unterliegen, durch Computersysteme ergänzt oder ersetzt werden, nicht nur Autofahren oder Partnervermittlung, sondern auch viel komplexere Abläufe. Amazon kann schon heute das passende Buch empfehlen und Google die am ehesten relevante Webseite, Facebook kennt unsere Vorlieben und LinkedIn kann sagen, wen wir kennen. Dieselben Technologien werden bald auf die Diagnose von Krankheiten, die Empfehlung von Therapien und womöglich sogar auf die Identifikation von »Kriminellen« angewandt werden, bevor sie ein Verbrechen begehen. Genau wie das Internet die Welt radikal verändert hat, indem es dem Computer das Kommunizieren ermöglichte, wird auch Big Data das Leben fundamental verändern, indem es eine völlig neue quantitative Dimension hervorbringt.

Mehr, unscharf, gut genug

Big Data wird zu einer Quelle für neuen wirtschaftlichen Wert und neue Innovationen werden. Aber es geht um noch viel mehr. Der Aufstieg von Big Data steht für drei Umwälzungen in unserer Art der Informationsanalyse, die sich auf Selbstverständnis und Organisation der Gesellschaft auswirken.

Die erste Umwälzung wird in Kapitel 2 beschrieben. In dieser neuen Welt können wir sehr viel mehr Daten analysieren. In manchen Fällen können wir sogar alle für ein bestimmtes Phänomen relevanten Daten heranziehen. Seit dem 19. Jahrhundert ist die Gesellschaft auf Stichproben (sog. Samples) angewiesen, wenn es um sehr große Zahlen geht. Dieser Zwang zu repräsentativen Stichproben ist ein Relikt aus einer Zeit, als Information etwas Außergewöhnliches war, ein Ergebnis der Grenzen der Informationsverarbeitung im analogen Zeital-

ter. Vor der Einführung leistungsfähiger Digitaltechnologien erkannten wir die Samples nicht als Behinderung, sondern nahmen sie als unvermeidlich hin. Aber indem wir alle Daten sammeln, sind wir nun in der Lage, Einzelheiten zu sehen, die vorher nicht erkennbar waren, solange wir nur kleine Datenmengen verarbeiten konnten. Big Data gibt uns einen besonders deutlichen Einblick in Details: Unterkategorien und Marktnischen etwa, die der Stichprobenmethode unzugänglich bleiben.

Die Verarbeitung einer sehr viel größeren Datenmenge lässt uns auch unser Bedürfnis nach Exaktheit überwinden. Das ist die zweite Umwälzung; mit ihr befassen wir uns in Kapitel 3. Ihr liegt ein Kompromiss zugrunde: Weil wir mehr Daten sammeln, können wir auch eine gewisse Unschärfe in der Datensammlung akzeptieren. Als unsere Fähigkeit zu messen begrenzt war, zählten wir nur das Wichtigste. Es war sinnvoll, die Daten möglichst exakt zu ermitteln. Man kann keine Kuhherde verkaufen, solange man nicht sagen kann, ob sie 100 oder nur 80 Tiere umfasst. Bis vor Kurzem gründeten alle unsere digitalen Werkzeuge auf dieser Exaktheit. Wir wollten, dass Datenbankabfragen exakte Auskunft liefern, die unsere Anfrage genau beantwortet.

Diese Denkweise war durch die Welt von »Small Data« (also der Vorbedingung nur weniger verfügbarer Daten) bedingt: Weil das Messen so aufwendig war, wollten wir die wenigen Daten so präzise wie möglich messen. Das ist nachvollziehbar: In einem Geschäft werden die Tageseinnahmen bis auf den Cent genau abgerechnet, beim Bruttoinlandsprodukt eines ganzes Staates ist das hingegen nicht mehr sinnvoll (und auch nicht mehr möglich). Mit der Größe wachsen auch die Ungenauigkeiten.

Exaktheit erfordert sorgfältig erhobene Daten. Sie funktioniert bei kleinen Datenmengen und ist für bestimmte Fälle natürlich auch in Zukunft unabdingbar: Entweder man hat genug Geld auf dem Konto für eine bestimmte Überweisung – oder nicht. Aber in einer Big-Data-Welt können wir diesen Fokus auf Genauigkeit im Gegenzug einer umfassenderen Sammlung von Daten zum Teil hinter uns lassen.

Big Data ist oft unscharf, von sehr verschiedener Qualität und über zahllose Server weltweit verteilt. Big Data gibt uns oft nur eine allgemeine Richtung vor, anstatt uns einen Sachverhalt bis auf den Zentimeter, den Cent oder das Atom genau zu erklären. Wir geben die Exaktheit nicht gänzlich auf, sondern nur unsere Versessenheit darauf. Was wir an Genauigkeit auf der Mikroebene verlieren, gewinnen wir an Erkenntnis auf der Makroebene.

Diese beiden Umwälzungen führen zu einer dritten, der wir uns im vierten Kapitel zuwenden: Eine Abwendung von der jahrtausendealten Suche nach kausalen Zusammenhängen. Als Menschen sind wir darauf ausgelegt, bei allem nach seiner Ursache zu fragen, auch wenn das oft schwierig ist und uns vielleicht auf eine falsche Fährte führt. In der Big-Data-Welt dagegen müssen wir uns nicht auf Kausalitäten festlegen, sondern können viel öfter nach Mustern und Korrelationen in den vorliegenden Daten Ausschau halten, die uns neuartige und wertvolle Erkenntnisse gewähren. Die Korrelationen sagen uns nicht *warum* etwas geschieht, aber sie machen uns darauf aufmerksam, *dass* etwas geschieht.

Und in vielen Fällen genügt das bereits. Wenn Millionen elektronischer Patientenakten zeigen, dass Krebskranke, die eine bestimmte Kombination von Aspirin und Orangensaft einnehmen, eine Remission der Krankheit erfahren, dann ist die genaue Ursache dieses Phänomens vielleicht nicht so wichtig wie die Tatsache, dass die Patienten überleben. Ebenso wenig müssen wir die Methode hinter dem Wahnsinn der Preisgestaltung von Flugtickets verstehen – es genügt, wenn wir einfach den besten Zeitpunkt zum Kauf ermitteln können. Bei Big Data geht es um das *Was*, nicht um das *Warum*. Nicht immer müssen wir die Ursache eines Sachverhalts kennen, sondern können mitunter auch die Daten für sich selbst sprechen lassen.

Vor Big Data war unsere Analyse oft auf das Prüfen einer bestimmten Hypothese beschränkt, die schon vor der Sammlung der Daten klar sein musste. Wenn wir hingegen die Daten sprechen lassen, ergeben sich Zusammenhänge, an die niemand zuvor gedacht hat. Einige Hedge-Fonds

verfolgen zum Beispiel den Kurznachrichtendienst Twitter, um die Kursentwicklung am Aktienmarkt vorhersagen zu können. Amazon und Netflix gründen ihre Produktempfehlungen auf eine Analyse der unzähligen Interaktionen ihrer Kunden auf ihren Webseiten. Twitter, LinkedIn und Facebook zeichnen den »Social Graph« der Beziehungen ihrer Kunden auf, um deren Vorlieben zu erfahren.

Natürlich analysiert der Mensch schon seit Jahrtausenden Daten. Die Schrift entstand im alten Mesopotamien, weil die Verwaltung sich ein effizientes Instrument zur Aufzeichnung und Verarbeitung von Information wünschte. Schon in der Bibel lesen wir von amtlichen Volkszählungen, mit denen die Regierungen riesige Datenmengen über ihre Bürger erhoben, und die Versicherungsstatistiken sammeln seit 200 Jahren ebenfalls große Datenschätze, um die Risiken, mit denen sie sich befassen, besser verstehen oder wenigstens vermeiden zu können.

Im analogen Zeitalter war das Sammeln und Analysieren von Daten allerdings sehr kosten- und zeitintensiv. Neue Fragestellungen bedeuteten damals oft auch die Notwendigkeit einer erneuten Datensammlung und -analyse.

Der große Schritt zu einem besseren Datenmanagement war die Einführung der digitalen Verarbeitung: Indem man analoge Informationen für Computersysteme lesbar macht, kann man sie gleichzeitig leichter und billiger speichern und verarbeiten. Dieser Fortschritt bedeutete eine dramatische Steigerung der Effizienz. Datenanalysen, die früher Jahre gedauert haben, waren jetzt eine Sache von höchstens noch Tagen. Sonst aber veränderte sich wenig. Die Menschen, die mit den Daten umgingen, waren allzu oft vom analogen Paradigma geprägt, dass Daten einem speziellen Zweck dienten und darüber hinaus wertlos waren. Unser eigenes Handeln hielt dieses Vorurteil aufrecht. So wichtig die Digitalisierung auch für die Umwälzung durch Big Data war, wurde diese doch nicht von der bloßen Einführung der Computer bewirkt.

Es gibt keinen wirklich treffenden Begriff, um die gegenwärtig stattfindende Umwälzung zu beschreiben, aber ein pragmatischer Versuch ei-

ner Näherung ist das Wort *Datafizierung*, ein Konzept, das wir im fünften Kapitel vorstellen. Es bezeichnet die Umwandlung von allem nur Vorstellbaren – auch von Dingen, die wir nie als Informationen betrachtet hätten, etwa den Standort eines Menschen, die Vibrationen eines Motors oder die statische Belastung einer Brücke – in Datenform, um sie damit quantifizieren zu können. Dadurch können wir diese Informationen auf ganz neue Arten verwenden, zum Beispiel für Analysen und Vorhersagen. So kann man etwa an der Wärmeabgabe oder den Vibrationsmustern eines Motors erkennen, dass er bald versagen wird. Auf diese Weise gewinnen wir Zugang zum impliziten, latenten Wert einer Information.

Schon ist eine Schatzsuche ausgebrochen nach den Erkenntnissen, die aus Daten gewonnen werden können, indem man dank einer Abkehr von der Suche nach Ursachen ihren verborgenen Wert freilegt. Nahezu jede Datensammlung, jedes Datenstück hat intrinsische, verborgene, noch unentdeckte Nutzen und damit auch ökonomischen Wert, und das Rennen, alle diese Datenschätze zu heben, ist in vollem Gange.

Big Data bringt eine radikale Veränderung des Geschäftslebens, der Märkte und der Gesellschaft mit sich, wie wir im sechsten und siebten Kapitel schildern. Im 20. Jahrhundert traten zu physischen Werten wie Grund und Boden oder Anlagevermögen immaterielle Güter hinzu, wie Markennamen oder geistiges Eigentum. Diese Entwicklung setzt sich jetzt mit Daten fort, die zu einem bedeutenden Wirtschaftsfaktor werden, indem sie einen unabhängigen ökonomischen Wert erlangen und die Grundlage neuer Geschäftsmodelle bilden. Sie sind das Öl im Getriebe der Informationswirtschaft. Bis jetzt werden Daten selten als Aktivposten in einer Unternehmensbilanz verzeichnet, aber das dürfte sich in Zukunft ändern.

Obwohl es Verfahren zur Verarbeitung großer Datenmengen schon länger gibt, waren sie bisher nur Geheimdiensten, Forschungseinrichtungen und wirklich großen Konzernen vorbehalten. Schließlich haben Walmart und Capital One den Einsatz von Big Data im Einzelhandel und im Bankgeschäft eingeführt und dadurch ihre Branche jeweils ent-

scheidend neu geprägt. Inzwischen sind viele dieser Verfahren demokratisiert (die Daten allerdings nicht).

Der größte Schock wird vermutlich die Rückwirkung auf den Einzelnen sein. Fachkenntnisse auf einem spezifischen Gebiet werden weniger wichtig, wenn Wahrscheinlichkeit und Korrelation entscheidend sind. In dem Film *Die Kunst zu gewinnen – Moneyball* konnte man sehen, wie Baseball-Talentscouts von Statistikern in den Hintergrund gedrängt werden, weil das Bauchgefühl für die Fähigkeiten eines Nachwuchsspielers gegenüber einer ausgefeilten Analysetechnik nicht mehr zählt. Es wird zwar immer Fachleute für einzelne Gebiete geben, aber sie werden sich mit den Urteilen der Big-Data-Analytiker auseinandersetzen müssen. Das wiederum wird eine Anpassung traditioneller Vorstellungen von Management, Entscheidungsfindung, Personalsuche und Ausbildung nach sich ziehen.

Die meisten unserer Institutionen beruhen auf der Voraussetzung, dass menschliche Entscheidungen auf der Grundlage weniger, exakter und kausaler Informationen getroffen werden. Aber wenn die Datengrundlage extrem umfangreich ist, blitzschnell verarbeitet werden kann und trotz Unschärfe neue Einsichten bietet, dann werden wir unseren Umgang mit Daten grundlegend überdenken müssen. Außerdem werden aufgrund der Menge an Daten Entscheidungen in Zukunft oft nicht mehr von Menschen, sondern von Maschinen getroffen werden. Der daraus resultierenden dunklen Seite von Big Data wenden wir uns im achten Kapitel zu.

Unsere Gesellschaft hat lange Erfahrung mit der Einschätzung und Regulierung menschlichen Verhaltens. Wie aber reguliert man das Verhalten eines Algorithmus? Schon zu Beginn der Informatik erkannten Fachleute, dass diese neue Technologie zur Aushöhlung der Privatsphäre des Einzelnen führen kann. Inzwischen gibt es gesellschaftliche Regeln zum Schutz personenbezogener Daten. Im Zeitalter von Big Data ist dieser Datenschutz allerdings oftmals genauso nutzlos wie im Zweiten Weltkrieg die Maginot-Linie gegen die feindlichen schnellen Panzerverbände. Denn heute stellen viele Menschen freiwillig persönliche

Informationen ins Netz – das ist ein wichtiges Merkmal der Netzkultur, und kein Fehler, den man verhindern müsste.

Die Gefahr verlagert sich heute vom Angriff auf die Privatsphäre des Einzelnen hin zur ungewollten Beurteilung des Einzelnen aufgrund von Wahrscheinlichkeiten: Algorithmen sagen vorher, dass man wahrscheinlich einen Herzinfarkt erleiden wird (was die Krankenversicherung verteuert), dass man seine Kreditraten nicht zahlen können wird (was zur Ablehnung des Hypothekenantrags führt) oder gar, dass man ein Verbrechen begehen wird (woraufhin man vorbeugend verhaftet wird). Das wirft die ethische Frage nach der Rolle des freien Willens gegenüber der Diktatur der Daten auf. Zählt die Entscheidungsfreiheit des Individuums mehr als die Voraussagen von Big Data, auch wenn die Statistik etwas anderes sagt? Genau wie die Erfindung der Druckerpresse letztlich zur Formulierung des Grundrechts auf freie Meinungsäußerung führte – das es vorher nicht gegeben hatte, weil so wenig Möglichkeit bestand, seine Meinung zu verbreiten –, wird es im Zeitalter von Big Data erforderlich sein, neue Regeln zum Schutz der individuellen Freiheit zu finden.

In vielerlei Hinsicht wird sich unsere Datenverarbeitung und -kontrolle verändern müssen. Wir nähern uns einer Zeit ständiger Vorhersagen aufgrund von Datenbeständen, in der wir unsere Entscheidungen womöglich nicht mehr begründen können. Was bedeutet es, wenn ein Arzt eine medizinische Behandlung nicht mehr rechtfertigen kann, ohne den Patienten aufzufordern, einer »Black Box« zu vertrauen, wie es der Fall sein wird, wenn die Diagnose von Big Data bestimmt wird? Wird die Polizei nicht mehr bei »Gefahr im Verzug«, sondern womöglich schon bei »wahrscheinlicher Gefahr« zugreifen? Wenn ja, was sind die Folgen für die Freiheit und die Würde des Menschen?

Für dieses neue Zeitalter brauchen wir neue Prinzipien; wir stellen sie in Kapitel 9 vor. Obwohl die Prinzipien sich auf Werte gründen, die für die Welt der wenigen Daten entwickelt und kodifiziert worden sind, sind sie nicht einfach eine neue Version für veränderte Umstände, sondern selbst etwas prinzipiell Neues.

Die Gesellschaft wird auf vielfältige Weise Nutzen daraus ziehen, wenn Big Data auf die Lösung drängender weltweiter Probleme wie der Erderwärmung, auf die Seuchenbekämpfung, die öffentliche Verwaltung und die Förderung guter Regierungsführung und wirtschaftlichen Fortschritts angewandt wird. Sie sieht sich allerdings auch der Herausforderung gegenüber, sich besser auf die Veränderungen vorzubereiten, die diese Technologie für unsere Institutionen wie für uns selbst darstellt.

Big Data ist ein wichtiger Schritt auf der Suche der Menschheit nach Quantifizierung und Verständnis der Welt. Eine enorme Masse von Information, die nie zuvor gemessen, gespeichert, analysiert und verbreitet werden konnte, wird jetzt datafiziert. Der Einsatz großer Datenmengen anstatt nur kleiner Samples und die Bevorzugung eines Mehr an Daten anstatt möglichst großer Exaktheit kleiner Datenmengen öffnet neuen Erkenntnissen die Tür. Sie führen dazu, dass die Gesellschaft ihre traditionelle Neigung zur ausschließlichen Suche nach Ursachen aufgeben und in vieler Hinsicht die Vorteile der Korrelation nutzen wird.

Denn die Vorstellung, wir könnten einfach Ursachen erkennen, ist eine überschätzte Illusion, die von Big Data infrage gestellt wird. Wieder einmal befinden wir uns an einem historischen Scheidepunkt, an dem wir feststellen »Gott ist tot«. Das, was wir als gesichert angenommen haben, verändert sich auf einmal. Dieses Mal allerdings wird es ironischerweise durch bessere Belege ersetzt. Welche Rolle bleibt da noch der Intuition, dem Glauben, der Ungewissheit, dem Handeln wider besseres Wissen und dem Lernen durch Erfahrung? Wenn sich die Welt von der Suche nach Kausalitäten hin zum Erkennen von Korrelationen bewegt, wie können wir dies pragmatisch tun, ohne die Grundlagen der Gesellschaft, der Menschlichkeit und des Fortschritts zu untergraben, die auf der Vernunft beruhen? Dieses Buch möchte erklären, wo wir heute stehen, nachverfolgen, wie wir dorthin gelangt sind, und einen dringend benötigten Leitfaden für die Vorteile und Gefahren bieten, auf die wir zugehen.