#06 DIGITALKOMPAKT LfM

KLEINE DATEN, GROSSE WIRKUNG

Big Data einfach auf den Punkt gebracht.

lfm:
Landesanstalt für Medien
Nordrhein-Westfalen (LfM)

#06 DIGITALKOMPAKT LfM

KLEINE DATEN, GROSSE WIRKUNG – BIG DATA

INTRO
3 # Wir alle speisen den Ozean der Daten!

HINTERGRUND
4 # Rudern im Ozean der kleinen und großen Daten

ÜBERBLICK
10 # Wie man Petabytes bändigt

PERSPEKTIVE
16 # Big Data beginnt mit Small Data

HINTERGRUND
24 # Das Einmaleins der Daten

PERSPEKTIVE
26 # Big Data ist Big Business

32 # Gefahren und Nachteile für den Nutzer, oder: Die Ethik der Daten

FAZIT & AUSBLICK
38 # Leben in der Big-Data-Welt

39 Der Autor dieser Ausgabe
40 Glossar
41 Impressum

„WIR ALLE SPEISEN DEN OZEAN DER DATEN!“

Daten sind der Treibstoff und das Schmiermittel der vernetzten Wirtschaft. Ohne sie geht fast gar nichts mehr. Wir alle speisen den Ozean der Daten mit jedem Klick, jedem Einkauf, jedem Griff zur Bonuskarte oder der Eingabe auf dem Navigationsgerät. Experten haben für diese unerhörte Flut von Informationen, die Menschen wie Maschinen erzeugen, einen imposanten Begriff geprägt: „Big Data." Das klingt nach Größe, Vehemenz, Tragweite, nach einer Zahl mit zu vielen Nullen, als dass sie ein Mensch noch verstehen könnte. Und „Big Data" ist zugleich vage genug, um jede Menge Erklärungsversuche herauszufordern.

Viele der Daten, die die moderne Internetwirtschaft am Laufen halten, sind keineswegs nur binäre Informationen wie die Messdaten eines Schalters am Fließband oder eines Strichcodes auf einem Päckchen. Es sind Meinungen, Urteile, Klatsch und Tratsch. Marketingchefs wollen wissen, was rund um die Uhr im Netz über ihre Marke gepostet wird. Kunden erwarten, dass ihre Hilferufe und Beschwerden prompt wahrgenommen und beantwortet werden. Personalchefs benutzen Software, um soziale Medien nach Kandidaten zu durchkämmen, und sie verlassen sich ebenso immer mehr auf Software, um Bewerber auszusieben. Leser erwarten, dass die App ihrer Zeitung oder ihres Senders weiß, welche Themen sie schätzen – und Medienunternehmen wollen ihrerseits Inhalte und die rundherum platzierte Werbung möglichst auf jeden einzelnen Kunden zuschneiden.

Die Datenfülle hat eine mindestens ebenso große Schattenseite, denn sie erlaubt völlig neue Formen der Benachteiligung und Ausgrenzung, die einen einzelnen Verbraucher oder Bürger ins Visier nehmen – von der Vorenthaltung von Informationen bis schlimmstenfalls zur genetischen oder sozialen Diskriminierung. Einmal angelegte Datensätze haben zudem ein beinahe ewiges Leben und können Jahre oder Jahrzehnte später wieder auftauchen, um etwa eine Karriere oder Beziehung zu ruinieren.

Diese Ausgabe von Digitalkompakt der Landesanstalt für Medien Nordrhein-Westfalen (LfM) wird die verschiedenen Facetten, die Chancen und Risiken von „Big Data" zu beleuchten versuchen. Wie fügen sich viele kleine Datenpunkte zum großen Ganzen? Wie lassen sich aus Unmengen an Daten Informationen und vor allem Bedeutung destillieren? Wie wird „Big Data" Gesellschaft und Volkswirtschaft verändern? Und welche Herausforderungen und Gefahren bringt der Wandel zur datengetriebenen Gesellschaft für das Leben jedes Nutzers mit sich?

RUDERN IM OZEAN DER KLEINEN UND GROSSEN DATEN

Datenverarbeitung ist nichts Neues, schon antike Gesellschaften entwickelten Systeme, um etwa ihren Viehbestand oder die Ernte zählen und besser verwalten zu können. Aber das Internet hat alle unsere Vorstellungen dessen gründlich erschüttert, was sich messen lässt – und wie oft und von wem.

Wer heute online einkauft, kann sich darauf verlassen, dass ihn der e-Shop wiedererkennt. Der Server am anderen Ende der Verbindung weiß, wer wann welches Paar Schuhe angeschaut, in den Warenkorb gelegt, aber dann doch nicht bestellt hat. Wer eben noch auf einer Reise-Webseite nach Flügen in die Türkei gesucht hat, darf erwarten, bei den unmittelbar folgenden Stationen im Web Anzeigen für Pauschalurlaube in Antalya oder besonders preiswerte Flüge serviert zu bekommen. Rechenzentren haben hinter den Kulissen in Sekundenbruchteilen den individuellen Nutzer erkannt, sein Surfverhalten analysiert und dem meistbietenden Anzeigenkunden Werbeflächen verkauft, die wie von Zauberhand auf der Startseite der Tageszeitung des Nutzers auftauchen.

ALLES KOMMUNIZIERT MIT ALLEM

Auch durch die Offline-Welt fließen sichtbare wie unsichtbare Datenströme, von denen die meisten nichts wissen. Ein Mitglied checkt sich im Fitness-Studio mit einer Chipkarte ein, und das Laufband oder der Crosstrainer kennt sein Stresslevel. Die Geräte wissen unter Umständen,

dass er seit Freitag keinen Sport mehr getrieben hat. Die Rabattmarke für ein neues Waschmittel, die die Kassiererin einscannt, findet ihren Weg zum Rechenzentrum des Grossisten, der damit seine morgige Lieferung anpassen kann. Der Container, der gerade im Hafen auf einen Güterzug umgeladen wird, hat sich mit einem Funksensor schon mehrfach an- und wieder abgemeldet, sodass das Unternehmensplanungs-System hunderte Kilometer entfernt bereits die Bauteile, die in ihm verstaut sind, einer Schicht zuweisen kann. Datenströme treiben so inzwischen fast alle Lebensbereiche an und erlauben eine bislang ungeahnte Verfolgung einzelner Güter und jedes einzelnen Verbrauchers, obwohl diese Vorgänge den meisten Menschen ebenso verborgen bleiben wie die genaue Funktionsweise des Mobilfunknetzes.

JEDER MENSCH TRÄGT ZUM DATENSTROM BEI

Parallel dazu erzeugen wir aktiv und wissentlich Daten für andere Menschen. Jeder Eintrag in einem sozialen Netzwerk, jede Kurznachricht in einem Mikroblogging-Dienst wie Twitter, jedes Foto von der Wanderung, das wir samt den in der Bilddatei enthaltenen Standortdaten hochladen, fließt umgehend in den endlosen Datenozean.

Die Liste der Datensätze, die heute gesammelt werden, ist lang:

- Finanzielle Transaktionen
- Einkäufe, online wie offline
- Web-Protokolle aus einem Browser oder einer mobilen App
- Verbindungsdaten von SMS und Telefonaten
- Standortdaten von vernetzten Geräten, vom Smartphone bis zur Digitalkamera
- Verkehrsdaten aus einem Navigationsgerät, einem Fahrzeug oder in die Straße eingebetteten Sensoren und Mautstationen
- Sensordaten aus ganzen Fertigungsstraßen oder Warenlagern, von Containern und einzelnem Stückgut
- Biometrische und Vitaldaten vom Fitness-Studio bis zum Krankenhaus
- Einträge in sozialen Medien
- Video- und Tondateien

RUDERN IM OZEAN DER KLEINEN UND GROSSEN DATEN

Das addiert sich auf. Während der Speicherplatz auf einem Handy oder einem Laptop in Gigabyte bemessen wird, rechnen Unternehmen, die Daten sammeln, verwalten und auswerten, längst in Terabyte, Petabyte, Exabyte und Zettabyte. Letzteres ist eine Zahl mit 21 Nullen:

1.000.000.000.000.000.000.000

Wenn es stimmt, dass Daten die neue Währung der Informationsgesellschaft sind, leben wir in Zeiten der Hyperinflation. Experten haben hochgerechnet, dass die Menschheit vom Beginn der Zeitrechnung bis zum Jahr 2003 rund fünf Milliarden Gigabyte an Daten erzeugt hat. Der Siegeszug des Internets, immer leistungsfähigerer Rechner und tragbarer Geräte sowie immer billigerer Speichermedien hat dazu geführt, dass wir mehr Daten denn je erzeugen und auch aufbewahren: Im Jahr 2011 sammelte sich dieselbe Datenmenge – 4,7 Exabyte – bereits alle 48 Stunden an. Wenn sich der Trend so fortsetzt, und alles sieht danach aus, wird es 2013 nur noch zehn Minuten dauern, bis diese Datenmenge anfällt.

Das IT-Marktforschungsunternehmen IDC hat einen eindrücklichen Vergleich angestellt: Mit all den Daten, die alleine im Jahr 2009 geschaffen und auf andere Datenträger kopiert wurden, ließen sich genügend DVDs füllen, um sie einmal zum Mond und zurück zu stapeln. Bis 2020, schätzen die Experten, wird der Stapel 44 Mal so hoch sein!

DIE BEZWINGUNG DES DATENMEERS

Einsen und Nullen sind aber nur das Rohmaterial der digitalen Wirtschaft. Big Data dreht sich indes nicht um die schiere Menge an Daten, in denen wir alle zu ertrinken drohen, sondern um gezielt herausgefilterte und auf individuelle Bedürfnisse angepasste Rinnsale. Big Data ist insofern weniger eine Zustandsbeschreibung für Experten der Exponentialrechnung als vielmehr eine Vision für das datengestützte Leben im 21. Jahrhundert. Big Data und die darunter subsumierten Technologien sollen das Chaos kanalisieren, Sinn stiften, Fragen beantworten und letztlich Verbrauchern, Unternehmen und Behörden bei der Entscheidungsfindung helfen.

Big Data, richtig umgesetzt, kann alle möglichen Aspekte unseres Lebens effektiver und effizienter machen – von Konsum und Kommerz über Unterhaltung bis zu Forschung, Wissenschaft und Bildung. Die Betonung liegt auf „kann“, denn wie alle großen Trends hat auch dieser seine Schattenseiten.


BYTE
= 8 BIT
KILOBYTE
= 1.000 BYTE
MEGABYTE
= 1.000.000 BYTE
GIGABYTE
= 1.000.000.000 BYTE
TERABYTE
= 1.000.000.000.000 BYTE
PETABYTE
= 1.000.000.000.000.000 BYTE
EXABYTE
= 1.000.000.000.000.000.000 BYTE
ZETTABYTE
= 1.000.000.000.000.000.000.000 BYTE

RUDERN IM OZEAN DER KLEINEN UND GROSSEN DATEN

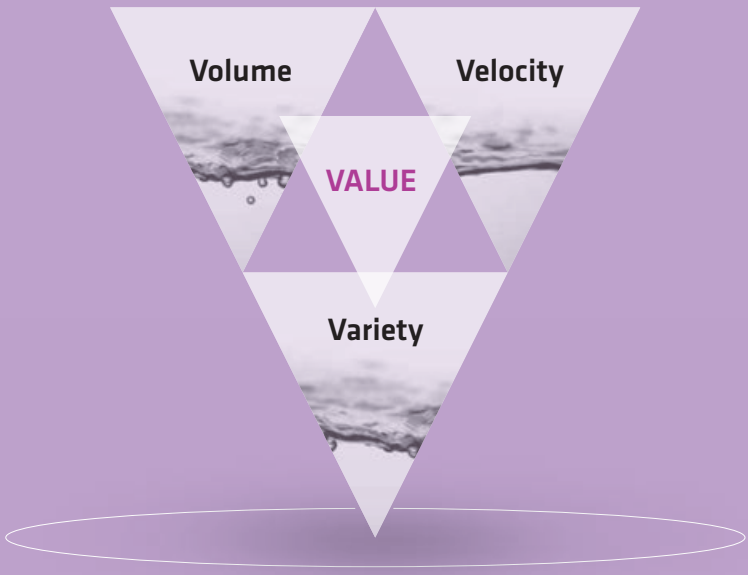
WERDEN WIR MASCHINENLESBAR?

Da sind einmal übertriebene Versprechungen und überzogene Erwartungen zu nennen, wenn sich Unternehmen oder Behörden vom Sammeln und Auswerten großer Datensätze die Lösung aller Probleme erwarten. Für den Nutzer wirft Big Data zahlreiche, meist ungelöste rechtliche wie ethische Fragen auf, was den Umgang mit diesen Daten angeht. Wenn Datensätze darüber entscheiden, wer was wann zu welchem Preis einkaufen kann oder wer bestimmte Informationen oder Dienstleistungen zu sehen oder vorenthalten bekommt, dann entstehen neue Formen der gleichsam automatischen Diskriminierung. Im schlimmsten Fall entsteht der maschinenlesbare Mensch, dem bei jedem Handgriff von der Wiege bis zur Bahre ein Algorithmus über die Schulter blickt und Buch führt.

Die allgemein akzeptierte Definition von Big Data klingt unschuldig genug: All jene Daten, die sich mit herkömmlichen Software-Werkzeugen und Technologien nicht mehr bearbeiten lassen. Andere Fachleute haben das etwas salopper gefasst: Big Data sind alle Datensätze, die sich nicht mehr in eine Excel-Tabelle pressen lassen. Das mag simpel klingen, doch mit diesem Werkzeug arbeitet ein Großteil der modernen Wirtschaft – von den Mainframe-Rechnern multinationaler Konzerne und den weltweit verteilten Rechenzentren von Internetriesen wie Amazon, Google oder Microsoft einmal abgesehen. Zugleich versinnbildlicht das Dilemma der Tabellenkalkulation die Probleme, die ungebremste Datenströme schaffen.

WIE MAN PETABYTES BÄNDIGT

Big Data lässt sich anhand von drei Aspekten beschreiben, die im Fachjargon als die „drei Vs“ bezeichnet werden: Datenmenge oder Volume, Geschwindigkeit oder Velocity und Vielfalt oder Variety. Wer das Phänomen „Big Data“ erfassen möchte, sollte alle drei betrachten.



Das erste Kriterium – das Volumen – ist noch am einfachsten nachvollziehbar. Werfen Sie einen Blick auf Ihren Rechner zu Hause und sehen Sie einmal nach, wie viele digitale Fotos sich angesammelt haben. Gleiches gilt für Dokumente im Textverarbeitungsprogramm und auf einem Webmail-Konto gehortete Korrespondenz. Manche dieser Datensätze haben Sie bearbeitet, in Ordnern abgelegt oder verschlagwortet, viele sind einfach nur abgelegt und vergessen worden.

Ein Unternehmen wie ein Verlag oder Sender, der tagtäglich neue Inhalte produziert, steht demselben Problem gegenüber: einem Archiv aus Tausenden von Textbeiträgen, interaktiven Karten, Leserbriefen und Kommentaren, Tweets, die einen Artikel erwähnen, Notizen und Rohmaterial aus der laufenden Produktion. Mindestens ebenso groß ist das Daten-Volumen bei einem Hersteller, der Komponenten entlang seiner gesamten Lieferkette verwaltet, die Fertigungsstraßen in mehreren Fabriken überwacht, die Logistik mit an Fahrzeugen installierten Sensoren oder GPS-Sendern verwaltet und gleichzeitig alle internen wie externen Prozesse in seine Steuerungs- und Analysesoftware einspeist. Das kann jede Kauforder sein, jede gestellte und bezahlte Rechnung, sowie alle Kommunikation, die seine Zulieferer, Mitarbeiter und Kunden auf elektronischem Wege abwickeln.

UNTERNEHMEN SITZEN AUF DATENBERGEN

Das Beratungshaus McKinsey schätzte in einer wegweisenden Studie aus dem Jahr 2011, dass das durchschnittliche US-Unternehmen mit 1.000 Beschäftigten auf mindestens 200 Terabyte an Daten sitzt, in vielen Fällen sogar auf einem Petabyte oder mehr. Für europäische Unternehmen veranschlagen die Forscher das Datenvolumen auf 70 Prozent der amerikanischen Konkurrenz. Selbst kleine und mittelständische Betriebe kommen inzwischen auf so große Datenmengen, dass sie vor erheblichen Problemen bei Erfassung, Speicherung und Verarbeitung stehen.

Das hängt wiederum mit der Geschwindigkeit, dem zweiten V, zusammen. Während man früher Daten in Schüben erhielt und verarbeitete – etwa wenn ein Mitarbeiter Zeit hatte, die Tabelle zu aktualisieren oder die Buchhaltung fällige Zahlungen bearbeitete – strömen Daten heute dank vernetzter Sensoren, Smartphones, Tablets und elektronischer Kommunikation rund um die Uhr auf uns ein.

Wer bei Big Data mitspielen will, muss die generierten oder von außen einlaufenden Daten immer schneller, im Idealfall in Echtzeit einspeisen und verarbeiten. Das passiert in der Regel über fest eingerichtete Schnittstellen, bei denen ein System mit dem anderen „spricht“ und neue Daten automatisch abfragt und einpflegt – von allen Tweets über die Marke oder andere Stichwörter bis zu Online-Bestellungen und den Logdateien, die den Verkehr von und zu einer Webseite protokollieren.

WIE MAN PETABYTES BÄNDIGT

DATEN MÜSSEN VERSTANDEN WERDEN

Womit wir beim dritten V wären, der Vielfalt – einer der größten Herausforderungen von Big Data. In der alten Welt waren Daten strukturierte Einträge meist numerischer Art, beispielsweise ein Produkt mit einer fest zugewiesenen Nummer in einer bestimmten Stückzahl an einem fest definierten Standort oder eine Überweisung von einem Konto an ein anderes. Solche Werte lassen sich relativ einfach in einer Datenbank anlegen, pflegen und wiederfinden.

Heute stehen Verbraucher wie Unternehmen einer wachsenden Anzahl von Datenquellen und -formaten gegenüber, die wenig bis gar nicht strukturiert sind und irgendwo im Web kursieren. Tweets oder Einträge auf einem sozialen Netzwerk wie Facebook sind frei von der Leber weg geschriebene Texte mit Doppeldeutigkeiten und Ironie. Software versucht inzwischen weit mehr, als diese Einträge nur zu quantifizieren. Das Ziel lautet, alle unstrukturierten Datensätze maschinenlesbar zu machen, sie also auf Inhalt und Stimmung zu analysieren. Programme sollen die Erwähnung von Marken oder sogar Gefühle und Emotionen erkennbar machen.

Die so extrahierten Daten werden in Größen übersetzt, mit denen Menschen und Maschinen buchstäblich rechnen können: Sind Nutzer mit einem Produkt zufrieden oder nicht? Machen sie sich über schlechten Kundendienst lustig? Sprechen Patienten vor allem von Nebenwirkungen, wenn sie über ein neues Medikament posten? Kommt der unzufriedene Blogger bereits in der Kundendatei vor und kann er mit einem gezielten Sonderangebot umgestimmt werden?

AUCH BILD- UND TONDATEN WERDEN AUSGEWERTET

Ton- und Bilddateien sind nicht nur vom Datenaufkommen umfangreicher, sondern auch weitaus schwieriger zu übersetzen, denn hier müssen Programme Sprache verstehen und transkribieren, urheberrechtlich geschützte Musik identifizieren, sowie Gesichter, Objekte oder Logos „erkennen." Das setzt erhebliche technische und semantische Fähigkeiten voraus, etwa die Unterscheidung zu treffen, ob mit „Paris Hilton" das platinblonde Sternchen oder ein Hotel in der französischen Hauptstadt gemeint ist. Und es stellt einen erheblichen, wenn nicht sogar illegalen, Eingriff in die Privatsphäre dar, wenn etwa soziale Netzwerke oder Regierungsbehörden ohne das Wissen der Nutzer Bewegungsdaten aus Fotoalben gewinnen oder Gesichter in Schnappschüssen analysieren, um den Freundeskreis des Nutzers zu ermitteln.

All diese Datenquellen müssen nicht nur in ein maschinenlesbares Format, sondern auch miteinander in Verbindung gebracht werden. Nur so ergibt sich aus einem steten Strom von Kurznachrichten, kombiniert mit Standortdaten von Handys und Lieferwagen sowie Transaktionen im Einzelhandel, ein lebendiges Geflecht, das sich jede Sekunde ändert und dennoch wertvolle Einsichten liefern kann.

DAS VIERTE V BRINGT DIE GEWINNE

Dank der Cloud – also im Netz verfügbaren Speichern und Rechenleistung, die man nach Belieben zuschalten kann, sofern man über eine Kreditkarte verfügt – stehen heute jedermann bereits eine Vielzahl von Verarbeitungsmethoden und schlüsselfertigen Plattformen zur Verfügung, um die drei Vs zu einem vierten V zu veredeln: Value, also dem primär monetären Wert, der sich aus Big Data gewinnen lässt. Das können kürzere Wege bei Fertigung und Auslieferung sein oder bessere und preisgünstigere Angebote für den einzelnen Verbraucher. Trotz aller vermeintlichen Vorteile sollte man jedoch bedenken, welche Gefahren in der Cloud lauern. Die Liste reicht von den technischen Risiken, seine Daten fern der eigenen Wohn- oder Arbeitsstätte zu speichern oder zu verarbeiten, bis zum unerkannten und unerlaubten Zugriff auf private Daten durch Dritte, mögen es Hacker, Konkurrenten oder Regierungsstellen sein. Je mehr Daten zirkulieren, desto größer ist die Wahrscheinlichkeit von Datenlecks und Datendiebstahl.

Wie lässt sich aus technischer Sicht aus Big Data Wert gewinnen? An erster Stelle sind hier Hadoop und ein Programm-Framework namens MapReduce zu nennen. Hadoop, benannt nach dem verschnupften Elefanten aus einem bekannten Kinderbuch, hat sich zu einem de-facto-Standard entwickelt, um große Datenmengen dezentral und schnell zu speichern und parallel zu bearbeiten. Es ging 2006 aus einem internen Forschungsprojekt der Firma Yahoo! hervor und wird inzwischen als Open-Source-Projekt unter dem Dach der Apache Foundation weitergeführt. Hadoop ist ein verteiltes Dateisystem, das es jedem Nutzer mit Netzanschluss erlaubt, enorme Datenmengen auf Gruppen oder Cluster von vielen Rechnern zu verteilen, um anschließend schneller auf sie zugreifen zu können.

Die eigentlichen Rechenaufgaben übernimmt dabei MapReduce. Dieses Framework entstand schon vor rund zehn Jahren im Hause des Suchriesen Google, um die parallele oder nebenläufige Berechnung großer Datenmengen in möglichst viele Häppchen auf möglichst viele Rechner aufzuteilen und Ergebnisse in Sekundenbruchteilen auszuspucken. Für Googles zentrale Rolle bei Big Data gibt es einen einfachen Grund: das gesamte Geschäftsmodell des Unternehmens basiert auf der Sammlung und Auswertung von Daten über seine Nutzer, um ihnen möglichst personalisierte Anzeigen zu servieren. Daraus ist ein weltweites Geschäft mit 38 Milliarden Dollar Jahresumsatz geworden, dessen Dienste aus dem Alltag fast nicht mehr wegzudenken sind. Gleichzeitig demonstriert Googles Dominanz wie kaum ein anderes Beispiel die Licht- und Schattenseiten der konstanten Datenerhebung. Jede Suchanfrage, jedes bei YouTube aufgerufene Video, jede bei Gmail versandte Nachricht bildet ein Puzzleteilchen, aus dem das Unternehmen die Identität, die Interessen und Intentionen von hunderten Millionen Menschen in aller Welt verfolgen, rekonstruieren und zu Geld machen kann.

WIE MAN PETABYTES BÄNDIGT

BIG DATA AUCH FÜR KLEINE NUTZBAR

Zurück zur Technologie: Auf der Basis der beiden frei erhältlichen Bausteine Hadoop und MapReduce haben sich inzwischen viele Erweiterungen und Werkzeuge entwickelt, die die unterschiedlichsten Software-Anbieter als schlüsselfertige Bündel offerieren. Das heißt, in der Cloud oder im Netz lässt sich Big Data nicht nur sammeln, sondern auch speichern und auswerten. Da auch ein Mittelständler so plötzlich Zugang zu leistungsfähigen Rechenzentren und der neuesten Software hat, sprechen Experten von einer Daten-Revolution, die weder große Anlaufinvestitionen noch eine kleine Armee von Informatikern erfordert. Oft genügt schon ein Browser auf dem Firmen-PC, um die vier Vs von Big Data für die eigenen Bedürfnisse zurechtzustutzen.

Der Wert der persönlichen Daten, richtig ausgewertet, ist enorm. Firmen, die ihre Kunden besser verstehen, können ihnen bessere Angebote unterbreiten oder sie zu mehr Einkäufen bewegen, ihre Angebotspalette und ihre Lagerhaltung optimieren. Der Wert von Big Data fällt dabei nicht nur Firmen, sondern auch Nutzern zu, sofern diese nichts gegen kontinuierliches Tracking haben. Ein Beispiel: Versicherungen bieten einzelnen Fahrern Sensoren für ihr Auto an, um anhand des tatsächlichen Fahrverhaltens individuelle Tarife zu berechnen. Der datengetriebene Handel steht erst am Anfang. Nach einer Studie der Boston Consulting Group waren persönliche Daten alleine in der Europäischen Union im Jahr 2011 rund 315 Milliarden Euro wert. Bis 2020 wird der Wert dieser Daten auf eine Billion Euro im Jahr steigen, in erster Linie aufgrund besser auf den einzelnen Nutzer zugeschnittener Produkte und Dienstleistungen.

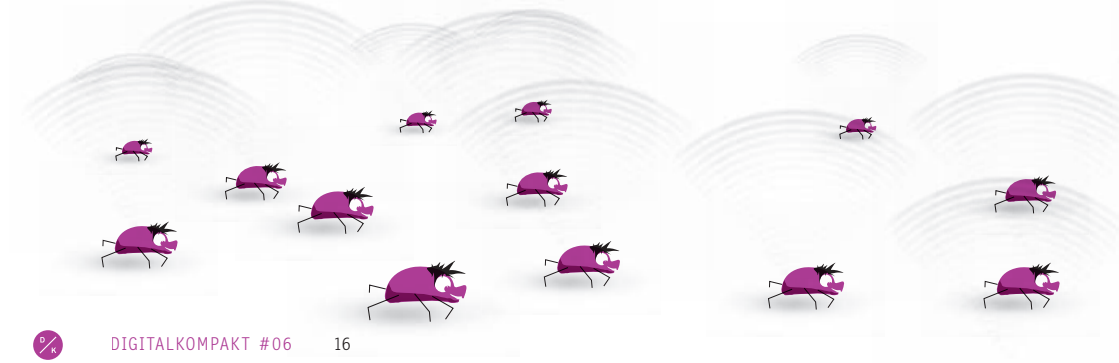
Gleichzeitig gibt es für Firmen wie auch den Einzelnen handfeste Gründe, seine wichtigsten Daten im eigenen Hause zu belassen, anstatt sie online zu speichern und zu bearbeiten. So behält man die Kontrolle über seine Daten, seien es Fotoalben eines ganzen Lebens, Geschäftsgeheimnisse oder die Akten einer Behörde. Ein netzbasierter Dienst kann gekapert werden oder gar abstürzen.

DIE CLOUD ALS GRUNDBEDINGUNG

Während sich die Fachwelt und Sicherheitsexperten über die beste Konfiguration für die Bändigung enormer Datenmengen streiten, lohnt sich festzuhalten: Big Data steht und fällt mit der Cloud und allen Geräten und Diensten, die an ihr hängen. Im sogenannten „Internet der Dinge“, das wir alle nutzen, entstehen nicht nur unerhört viele und vielfältige Datensätze. Sie werden dort auch von Geburt an gesammelt, gebündelt, gefiltert, nach Möglichkeit in die richtige, maschinenlesbare Form gebracht, damit sie sich abrufen, verwalten und befragen lassen. Zu diesem Zweck hat die EU trotz aller Bedenken zum Cloud-Computing eine große Cloud-Initiative angestoßen, die bis 2020 zu rund 2,5 Millionen neuen Arbeitsplätzen und einem jährlichen Zuwachs beim EU-Bruttoinlandsprodukt von rund 160 Milliarden Euro führen soll.

BIG DATA BEGINNT MIT SMALL DATA

Wenn Milliarden von Handys und preiswerten Sensoren jede Regung von Mensch und Maschine, von Prozessen und Produkten aufzeichnen und vermelden können, damit sie in einem Rechenzentrum ausgewertet werden können, bekommen Begriffe wie Selbstbewusstsein, Entdeckung und Entscheidungsfindung eine vollkommen neue Bedeutung. Ein neues Nervensystem für den Planeten entsteht.



Dieses neuronale Netz wird alle Bereiche unseres Alltags – privat wie beruflich – berühren, von der Unterhaltung über Erziehung und Bildung, Forschung und Wissenschaft, bis zur öffentlichen Verwaltung, dem Gesundheitswesen und dem Umweltschutz. Wenn Milliarden an Sensoren permanent mit dem Netz verbunden sind, wird Datenerhebung und -speicherung zum Normalzustand. Aus dieser Rohmasse können viele Beteiligte Sinn stiften: Stadtplaner, Verwaltungsbehörden, Umweltschützer, normale Bürger, die sich ein Armaturenbrett ihres Stadtteils aufrufen können.

Big Data fängt dabei fast immer klein an, beim Nutzer oder einem Gerät. So machte die Firma EMC, die Speicherlösungen anbietet, gemeinsam mit dem Marktforscher IDC die Rechnung auf, dass im Jahr 2011 rund 1,8 Billionen Gigabyte digitaler Daten – oder 1,8 Zettabyte – angelegt wurden. Drei Viertel davon stammen von ganz normalen Verbrauchern. YouTube etwa streamt weltweit vier Milliarden Videos am Tag. Nutzer laden im Durchschnitt in jeder Sekunde eine Stunde Video auf YouTube hoch. Inzwischen sind es längst nicht mehr nur Amateurfilme über Haustiere. Alle diese Videos werden von Software ausgewertet, die die Tonspuren nach urheberrechtlich geschützter Musik oder Inhalten absucht, die etwa dem Jugendschutz unterliegen. Findet sie entsprechende Audio-Fingerabdrücke, werden die Filme gesperrt, Tantiemen-Zahlungen an Verlage und Künstler in die Wege geleitet, die populärsten Clips mit Werbung versehen und sogar Untertitel in mehreren Sprachen eingeblendet – ohne dass Menschen dabei Regie führen müssen.

Andere Firmen arbeiten bereits an Technologie, um Gesichter oder Logos zu erkennen, etwa um ein Kleidungsstück zum lebendigen Link zu machen, das man direkt beim Zuschauen in den Warenkorb legen könnte. Ähnliche Analyse-Methoden wenden Netzwerke wie Facebook oder Legos Online-Gemeinschaft an, um alle Nachrichten und Chats aus Gründen des Jugendschutzes zu überwachen: Maschinen „hören“ Menschen zu, ohne dass diese es wissen – und blockieren zuweilen völlig harmlose Inhalte, weil der Programmierer einen Fehler gemacht und falsche Regeln festgelegt hatte.

SENSOREN SIND ÜBERALL

Diese Liste ließe sich beliebig verlängern. Jedes Mal, wenn Sie ein Tablet in die Hand nehmen und bei einem e-Book die Seite umblättern oder eine Passage farbig markieren, wenn Sie auf dem Handy eine Adresse nachschlagen oder sich die beste Verkehrsverbindung anzeigen lassen, werden Sensoren aktiv. Sie messen Standort und Geschwindigkeit, fragen Tag und Uhrzeit ab, vergleichen Ihre Kontodaten mit den auf einem oder mehreren Servern hinterlegten Angaben und spielen unter Umständen neue Inhalte von einem Verlag oder Sender automatisch auf Ihr Gerät: eine Eilmeldung, eine neue Folge der Lieblingsserie oder eine aktualisierte Neufassung des Sachbuchs, bei dem Sie gerade in Kapitel 4 angekommen sind. Dass Sie auf dem Tablet die Passagen sehen können, die die meisten anderen Leser markiert haben, wäre ohne Big Data und die Infrastruktur der Cloud auch nicht möglich.

BIG DATA BEGINNT MIT SMALL DATA

AUCH ENTERTAINMENT WIRD INDIVIDUALISIERT

Selbst der einfache Akt der Unterhaltung ist längst kein einfacher Sendevorgang mehr, bei dem ein Nutzer wie Millionen andere vor seinem Fernseher sitzt und eine Sendung zu einer bestimmten Uhrzeit einschaltet. Stattdessen entscheiden Algorithmen, wer zu welcher Sendung welche Werbung gezeigt bekommt. Anbieter registrieren genau, welche Abonnenten welche Geschichten oder Folgen wie lange ansehen, und reichen diese kleinen Teile des gewaltigen Konsumpuzzles an andere Rechner weiter. Bald werden Unterhaltungszentren zum Wohnzimmer-Standard gehören, die mit Kameras und anderen Sensoren erkennen, welches Familienmitglied gerade wo im Zimmer ist und aufpasst. Alle diese Datenpunkte haben Einfluss darauf, wie lange ein „Aufmacher" auf der Homepage bleibt oder sogar, ob ein Ressort im nächsten Monat mehr oder weniger Budget erhält, um neue Inhalte zu produzieren. So droht Technologie, die angeblich für mehr Auswahl sorgt, auf Dauer die Meinungsvielfalt und Entscheidungsfreiheit einzuschränken.

Schon heute gibt es Redaktionen, in denen die Seitenabrufe und Erwähnungen eines Artikels oder Videos in Echtzeit gemessen und auf großen Displays angezeigt werden, um eine Art Wettbewerb unter den Reportern und Produzenten anzukurbeln. Suchmaschinen wie Google filtern die Ergebnisse anhand des Such- und Klickverhaltens jedes Nutzers, sodass dieselbe Anfrage zu ganz anderen Ergebnissen führen kann, je nachdem wer sie gerade eintippt.

JEDER LIEST NUR NOCH, WAS ER LESEN WILL

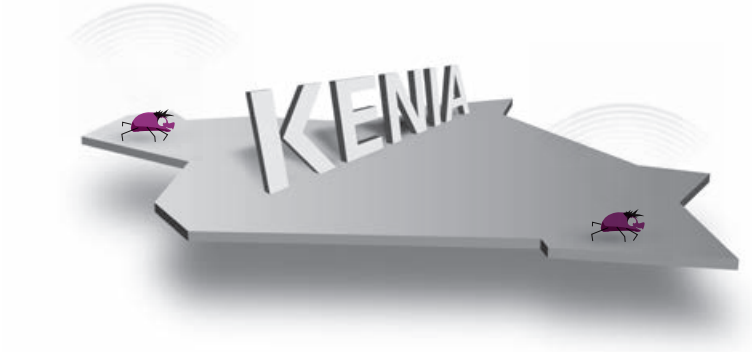
Hier kann unser Input an Small Data zu viel dynamische Personalisierung verursachen und zu einem gefährlichen Scheuklappen-Effekt führen, auf den der Internet-Aktivist Eli Pariser in seinem Buch „Die Filter Bubble" hingewiesen hat. Er kritisiert die Vorgehensweise von Datensammlern und -maklern wie Google, deren Algorithmen vordergründig dem Kundenerlebnis dienen, aber auf lange Sicht der Gesellschaft schaden. Jeder Verbraucher und Bürger baut sich Klick für Klick seine eigene Echo-Kammer, klagt Pariser, in der unpassende Neuigkeiten ausgefiltert werden, da sie nicht zu seinem oder ihrem Persönlichkeitsprofil passen. Auf der Strecke bleiben überraschende Entdeckungen und Kritikfähigkeit, die Grundlagen einer funktionierenden Demokratie.

Small Data hat auch erheblichen Einfluss auf die Art und Weise, wie Erziehung und Wissenschaft betrieben werden. Wenn sich das Erfassen und in einem zweiten Schritt die Interpretation von Daten zu einem erschwinglichen oder sogar kostenlosen „Volkssport" wandeln, entstehen im Idealfall neue Arten der basisdemokratischen Forschung.

In Kenia werteten Forscher die Bewegungsdaten von 15 Millionen Handys über ein ganzes Jahr aus und verknüpften jede Standortmeldung, jeden Anruf und jede SMS mit den Koordinaten. Daraus ermittelten sie das Reiseverhalten der Bürger und gleichen es mit der Verbreitung von Malaria-Fällen ab. Dank dieser Analyse konnten sie die Ausgangspunkte bestimmen, von denen infizierte Personen ins Umland reisen und die Parasiten weitertragen. Einen ähnlichen Vorsorge-Effekt erzielt Google, wenn es Suchanfragen zum Thema „Grippe“ bündelt und geografisch wie auch im Zeitverlauf auswertet, um eine Epidemie live zu verfolgen und Prognosen über ihre Verbreitung zu ermöglichen.

DER ZUGÄNGLICHE ZWILLINGSBRUDER

Big Data hat einen Zwillingsbruder namens „Open Data“. Seine Grundidee: Was rund um uns herum gesammelt und gespeichert wird, sollte auch allen „offen“ zugänglich sein, anstatt von Firmen oder Regierungen monopolisiert zu werden. Mit Open Data können Bürger, Behörden und Betriebe auf die weite Welt von Big Data ungehindert und kostenlos zugreifen und auf der Grundlage von frei verfügbaren Datensätzen neue Anwendungen entwickeln. So hätten weitere Bevölkerungskreise eine Chance zur Teilnahme an der datengetriebenen Welt, und Unternehmen wie Bürokraten wären dank stärkerer Transparenz mehr rechenschaftspflichtig als bisher.



BIG DATA BEGINNT MIT SMALL DATA

INPUT KOMMT VON ÜBERALL

Nichts ist zu trivial für Big Data: Dank eines Supercomputers in der Tasche wird jeder Bürger zu einem „Bürgerwissenschaftler" in seinem Viertel. In vielen Städten in aller Welt erproben Forscher wie der Italiener Carlo Ratti, ob deren Bewohner mit Small Data zu besserer Stadtplanung, Verkehrsführung und letztlich höherer Lebensqualität beitragen können. Die Forscher seines SENSEable City Lab, einer Einrichtung des Massachusetts Institute of Technology (MIT), instrumentieren Menschen, Fahrräder, Müllautos oder Bushaltestellen mit Luft- oder Lärm-Messgeräten und verbinden die Emissionswerte mit den Bewegungsdaten von zigtausenden Handys und Taxis. So werden plötzlich lebendige Stadtpläne sichtbar, die nicht nur eine Kommunalbehörde, sondern jeder einzelne Bewohner abrufen kann. Wenn er oder sie gerade an einem der neuralgischen Punkte steht und ein Smartphone in der Hand hat, schließt sich der unendliche Feedback-Kreislauf.

Selbst dort, wo noch keine oder nicht genug handfeste Daten existieren, können engagierte Bürger Schritt für Schritt, Bit für Bit, einen Teppich aus Big Data knüpfen, von dem der Rest der Gesellschaft profitieren kann. Das geht sogar ohne aktive Datensammlung, sondern einfach aufgrund der Tatsache, dass unsere Geräte eingeschaltet sind und als stumme Bewegungsmelder immer auf Empfang sind – mit allen negativen Folgen der lückenlosen Überwachung.

Millionen Smartphones mit dem Android-Betriebssystem liefern so anonyme Daten zum Verkehrsfluss, auf deren Grundlage Google Maps Staus erkennen und Routen berechnen kann. Mit Hilfe der passiven Teilnahme von 180.000 Nutzern konnte der israelische Navigationsanbieter Waze innerhalb weniger Monate den nach eigenen Angaben detailliertesten und aktuellsten Atlas des Landes anlegen. Wer mit dieser App zum ersten Mal eine Straße entlang fährt, schafft einen neuen Eintrag.

WER VERFOLGT MEIN SURF-VERHALTEN?

Nützlich ist es schon, wenn das Informationsangebot kontinuierlich steigt und auch die Qualität selbst feinmaschiger, lokaler Daten zunimmt. Aber zu welchem Preis? Mit Small Data steuern alle Menschen aktiv wie passiv zu einer permanenten Rasterfahndung durch Software bei – meist ohne zu wissen, bei welchen Diensten ihre Daten landen, wie sie weiterverarbeitet oder sogar weiterverkauft werden. Eine Webseite der TU Berlin ist eine praktische Kontrollinstanz, welche Webseiten den Nutzer im Alltag online verfolgen. Wer dort eine beliebige Web-Adresse eingibt, kann im Voraus sehen, wie viele Erst- und Drittanbieter auf seinem Rechner Cookies setzen wollen, um ihn künftig zu verfolgen und hochgradig personalisierte Werbung zu platzieren.

In die gleiche Richtung geht eine europaweite Initiative der Werbewirtschaft namens Youronlinechoices.com. Dort kann man auf einen Blick sehen, welche Werbenetze bereits Cookies auf einem Rechner hinterlegt haben, und sie entfernen.

Die Ausbeute an Small Data ist vielfältig. Das kann das Webprotokoll sein, mit dem ein soziales Netzwerk wie Facebook seine Nutzer quer durchs Internet verfolgt, um deren Verhaltensmuster anschließend an Dritte zu verkaufen. Das kann ebenso gut ein Supermarkt sein, der die Nutzer seines Bonusprogramms mit den neuesten Adressdaten der Post abgleicht. Oder ein Discounter wie die US-Marke Target, die aus allen intern wie extern verfügbaren Daten eine „Schwangerschafts-Prognose“ errechnet und seine Werbung danach steuert. Das Unternehmen ist oft besser informiert als die Frauen, die plötzlich Werbung für Windeln und Babypuder im Briefkasten vorfinden.

http://b-versio.verbraucher-sicher-online.de/jcookie/

BIG DATA BEGINNT MIT SMALL DATA

BEHÖRDEN UND VERWALTUNGEN ÖFFNEN SICH

Nicht immer geht es beim Auswerten von Small Data um die Gewinnmaximierung. Wenn Unternehmen und Behörden ihre Datenströme offenlegen, damit Bürger und andere interessierte Parteien daraus neue Anwendungen bauen können, spricht man von Open Data und Open Government, oder kurz: Open Gov. Der Kreativität sind dabei keine Grenzen gesetzt: Kriminalitätsstatistiken und selbst die Meldungen des Polizeiberichts vom Vorabend lassen sich mit geringer Verzögerung auf Webseiten und in mobilen Apps darstellen, ebenso die aktuelle Verfügbarkeit von Carsharing-Angeboten im Vergleich zu den Abfahrtszeiten des öffentlichen Nahverkehrs. Grundbucheinträge einer Kommune, gekoppelt mit Bewertungen der örtlichen Schulen und Kindergärten, können Familien dabei helfen, ein für sie geeignetes und erschwingliches Domizil zu identifizieren, während sie durch ein neues Stadtviertel schlendern.

Während sich deutsche Behörden noch vergleichsweise schwer damit tun, ihre Daten nicht nur offenzulegen, sondern auch für den automatischen Zugriff von kommerziellen Diensten und Apps vorzubereiten, preschen Städte wie San Francisco vor. Die Technologie-Hochburg war eine der ersten Städte, die bereits 2009 Richtlinien zum Umgang mit Big Data verabschiedete. Sie hat das Gesetz sogar als quelloffenen Text ins Netz gestellt, damit andere Städte die Paragraphen kopieren und schneller umsetzen können. Immer häufiger werden auch Kommunen den Posten des Daten-Managers oder Chief Data Officer schaffen.

DER GANZE MENSCH WIRD ÖFFENTLICH

Über die Gelegenheiten für Körperschaften sollte man den eigenen Körper nicht vergessen. Wer sich kontinuierlich selbst (und vielleicht mit anderen) misst, kann sich zur Avantgarde für ein „Quantifiziertes Ich“ zählen. Technologen und Gesundheitsfanatiker sind die Pioniere dieses aus den USA stammenden Trends, bei dem es darum geht, alle nur möglichen Daten über den eigenen Körper und das eigene Leben zu sammeln, auszuwerten und zu teilen. Chancen dazu bieten sich reichlich – von Apps auf dem Handy oder Accessoires, die sportliche Aktivitäten und Vitaldaten wie den Puls messen, bis zu Software, die die Zahl der versandten und beantworteten E-Mails zählt und anhand der Telefonverbindungsdaten berechnet, wie oft wir mit unseren Bekannten korrespondieren und wer gerade auf der Beliebtheitsskala oben steht.

Ob diese Art der vernetzten Nabelschau langfristig etwas Positives bewirkt, sei dahingestellt. Sie öffnet etwa der genetischen Diskriminierung durch Arbeitgeber, Versicherungen oder selbst ganz normale Hersteller von Verbrauchsgütern Tür und Tor. Wenn ein Unternehmen durch Recherchen in sozialen Medien oder anderen Datenquellen ermitteln kann, wer mit hoher Wahrscheinlichkeit an einem bestimmten Leiden erkranken wird, könnten diese Datensätze früher oder später ge- und missbraucht werden. Schon jetzt erproben erste Firmen in den USA und Großbritannien vernetzte Pflaster und Mikronadeln. Sie mögen wie technische Wunderwerke der Miniaturisierung gefeiert werden, aber diese Sensoren sind erste Vorboten der Big-Data-Landnahme an und sogar in unserem Körper, die intimste Datenströme wie Temperatur, Sauerstoffsättigung und andere Blutwerte kontinuierlich erheben und obendrein drahtlos übermitteln. Die Hersteller dieser Geräte planen, diese Daten Dritten zugänglich zu machen. Welche Mitspracherechte der Einzelne dabei hat, ist eine noch ungeklärte Frage. Bereits heute stehen chronisch Kranke vor dem Problem, dass in ihrem Körper eingepflanzte Medizintechnik, wie beispielsweise ein Defibrillator bei Herzpatienten, beständig Daten sammelt. Diese werden zwar dem Arzt und dem Hersteller zugänglich gemacht, aber nicht dem Patienten selber, dessen Körper die Daten generiert. Über diesen Streitpunkt der Eigentümerschaft sind bereits Klagen entbrannt.

Das Problem wird sich in Zukunft noch verschärfen, denn allgegenwärtige und preiswerte Hardware und Software erlauben es zum ersten Mal, ein fast lückenloses Protokoll des Lebens zu erstellen, aufzubewahren und sich darin nach Belieben umzusehen. Viele dieser Alltags-Datensätze sind schlicht und einfach „digitale Abgase“, die der Verkehr im Netz erzeugt, und nicht der genaueren Beachtung wert.

Einige dieser Sammlungen können für Wissenschaftler durchaus von Interesse sein, etwa wenn sie neuen Volkskrankheiten wie Fettleibigkeit und Diabetes auf der Spur sind, oder testen wollen, wie sich bestimmte Parameter auf das Verhalten einzelner Bevölkerungsgruppen auswirken. Während sie früher mit Flugblättern und Kleinanzeigen nach Probanden suchen mussten oder theoretische Modelle im Rechner durchspielten, können sie jetzt auf einen ständig wachsenden Fundus frischer und vielfältigster Daten zugreifen.

DAS EINMALEINS DER DATEN

Nicht umsonst wird das Netz oft mit einem Sammelsurium von Röhren verglichen, durch die Kubikmeter oder auch nur kleine Rinnsale an Daten schwappen. Big-Data-Dienste funktionieren dabei wie eine Art intelligenter Klempner, um die richtigen Röhren anzuzapfen und Ströme zu bündeln. Aber anders als in der physischen Welt sind diese Verbindungen dynamischer Art und äußerst flexibel.

Große und kleine Unternehmen sammeln und analysieren Big Data bereits rund um die Uhr, rund um die Welt. Dabei haben sie die Wahl unter mehreren Alternativen. Eigene Daten fließen in ihre internen Systeme ein, externe Daten werden aus dem Web eingesaugt, beispielsweise Tweets, Social-Media-Einträge oder andere, öffentlich zugängliche Quellen. Meist funktioniert das über sogenannte APIs oder Programmierschnittstellen, über die ein Dienst oder Programm mit anderen kommuniziert. Wer sich eine Zugangsberechtigung besorgt, kann in Echtzeit oder in vordefinierten Intervallen auf Datenströme zugreifen.

Je nach Datenaufkommen und Bedürfnissen sind diese Verbindungen ein Live-Stream oder die häppchenweise Übermittlung von größeren Datensätzen, die nur hin und wieder abgerufen bzw. in einem Zwischenspeicher abgelegt werden. Dem folgt als zweiter Schritt die Bearbeitung der Rohmasse. Unstrukturierte oder halb-strukturierte Daten müssen bereinigt und normalisiert werden, damit sie dieselbe Sprache sprechen wie ihre Datenkollegen aus anderen Quellen. Wer nicht genügend Daten hat, kann sie heute mit ein paar Klicks von einem der vielen neuen Daten-Marktplätze besorgen oder kaufen. Das sind große Online-Speicher von allen nur erdenklichen Datensätzen: Wetterberichte und Wettervorhersagen bis auf die Postleitzahl genau, geografische Informationen, Satellitenbilder, wirtschaftliche Indikatoren, öffentlich verfügbare Daten von Kommunen und Staaten, anonymisierte Datensätze von Webseiten, Finanzdienstleistern oder Mobilfunkbetreibern.

DATEN ALS DIENSTLEISTUNG

Die Daten müssen nun bereinigt, übersetzt und verarbeitet werden, und zwar – wir erinnern uns an die vier Vs – möglichst schnell. Mehr und mehr dieser Prozesse laufen heute als Dienstleistung großer Anbieter ab, die Speicherung, Bearbeitung und auch die Auswertung großer Datensätze offerieren. Amazon hat so durch Amazon Web Services einen Weg gefunden, seine Rechenzentren besser auszulasten. Auf seinen Servern kann jedermann Daten speichern und sie bearbeiten. In Amazons Simple Storage System (S3) sind bislang mehr als eine Billion Objekte abgelegt, und in Spitzenzeiten greifen Rechner irgendwo in der Welt 650.000 Mal in der Sekunde auf diese Dateien zu. Ebenso bieten Microsoft Azure und Googles Cloud Platform Speicherplatz, Rechenleistung und Analyse im Netz als Abonnement an. Man bezahlt für die genutzte Kapazität und die einzelnen Anfragen an seine Datensätze – vergleichbar einem Anruf bei der Auskunft, die entweder in den eigenen Telefonbüchern nachschlägt, die man vorher dort hinterlegt hat, oder beliebig viele externe Telefonbücher wälzt.

BIG DATA IST BIG BUSINESS

Schon jetzt hat Big Data das Wirtschaftsleben revolutioniert und zu einem endlosen Feedback-Kreislauf gemacht. Daten sind zu einem neuen Produktionsfaktor geworden, der gleichberechtigt neben Kapital, Ressourcen und Arbeitskraft steht.

EIN PAAR BEISPIELE AUS DER PRAXIS:

Beispiel 1

Finanzhäuser setzen Hochleistungscomputer und spezielle Software für sogenannten Hochfrequenzhandel ein. Nach Expertenschätzungen wickeln Algorithmen, die sich Millisekunden Zeitvorsprung zu Nutze machen, in Deutschland rund 40 Prozent aller Börsenaufträge ab, in den USA sogar knapp mehr als die Hälfte. Wer Computer handeln lässt, geht jedoch das Risiko wilder Kursschwankungen ein, die mit den Fundamentalwerten eines Unternehmens nichts mehr zu tun haben.

Beispiel 2

Das Kreditkartennetzwerk MasterCard verarbeitet 34 Milliarden Transaktionen im Jahr. In wenigen Sekunden werden eine ganze Reihe von Entscheidungen getroffen: Ob der Kauf verdächtig ist, weil er nicht zum Konsumverhalten, Standort oder Reiseverlauf eines Kunden passt, ob er das Kreditlimit überschreitet etc. Diesen Datenfundus vermietet MasterCard an Werbekunden, die die Transaktionsdaten nach Kunden und deren Einkaufsverhalten sortieren können. Für Datenschützer ist das ein Albtraum, da diese Prozesse weder ein informiertes Einverständnis der Nutzer erfordern noch transparent sind.

Beispiel 3

Die Firma Climate Corp., von zwei ehemaligen Google-Managern gegründet, bietet US-Landwirten eine Ernteversicherung an. Dazu wertet ihre Software alle sechs Stunden aktuelle Wetterdaten in 22 verschiedenen Kategorien von 1,5 Millionen Wetterstationen aus und kombiniert die Daten mit Bodenmessungen. In ihrem Modell spielt Climate Corp. rund 10.000 verschiedene Szenarien mit 34 Billionen Simulationen in den kommenden zwei Jahren durch, um die Versicherungsprämie für einen einzelnen Landwirt zu berechnen.

Beispiel 4

Xerox, der einstige Pionier bei Kopiergeräten, vertraut einem Algorithmus, um die besten Bewerber für die fast 50.000 Stellen in seinen Call Centern auszusieben. Die Software fand heraus, dass die Personalabteilung nicht nach der Erfahrung fragen sollte, sondern nur nach dem Charakter des Kandidaten. Wer zu kreativ denkt, wirft schneller das Handtuch. Nach einem halben Jahr dieser absichtlich „entmenschlichten“ Big-Data-Personalpolitik ist die Fluktuationsrate um 20 Prozent gefallen.

BIG DATA IST BIG BUSINESS

Kaum jemand hat die volkswirtschaftlichen Vorteile von Big Data besser dargelegt als das McKinsey Global Institute in einer Studie mit dem Titel „Big Data: The next frontier for innovation, competition, and productivity." Die Berater identifizieren darin fünf positive Effekte der Daten-Sintflut:

- **# Big Data schafft mehr Transparenz, was Unternehmen hilft, den Überblick zu bewahren und schneller bessere Entscheidungen zu treffen.**
- **# Big Data erlaubt mehr Planspiele und Simulationen, da Unternehmen auf unerhört großen Datenmengen sitzen und sie zeitnah auswerten können.**
- **# Big Data verbessert den Zugang zum einzelnen Kunden, sodass Produkte und Dienstleistungen auf eine Person zugeschnitten werden können.**
- **# Big Data unterstützt Firmen dank Analysewerkzeugen, Simulationen und Prognosen bei der Entscheidungsfindung.**
- **# Big Data sorgt für die Entstehung neuer Geschäftsmodelle, Produkte und Dienstleistungen – entweder von etablierten Unternehmen oder vollkommen neuen Firmen.**

Das mag abstrakt klingen, doch ein Grundsatzpapier des Bundesverbandes Informationswirtschaft, Telekommunikation und neue Medien (BITKOM) listet eine Handvoll von Beispielen für den Einsatz von Big Data für Unternehmen auf. Marketing und Vertrieb können mit vielen, intelligent ausgewerteten Daten die Produkte und Dienstleistungen ihrer Firma besser auf den Kunden abstimmen, da man erstmals jeden einzelnen Verbraucher kennenlernen und verfolgen kann. Akademiker und Forschungs- und Entwicklungsabteilungen in Unternehmen profitieren ebenfalls von Big Data.

GENAUERE ANALYSEN, SCHNELLERE ABLÄUFE

Wer Sensordaten und Feedback über soziale Medien erhält, kann schneller Hypothesen testen, Fehler finden und das Innovationstempo anziehen. Stammen die Inputs aus der Fertigung oder aus dem laufenden Betrieb beim Kunden, lassen sich die Herstellung optimieren und Probleme identifizieren. Wenn beispielsweise die Sensoren an einem Düsentriebwerk ungewöhnliche Temperaturen oder Vibrationen messen und schon aus der Luft weitermelden, können die Wartungsarbeiten optimiert werden und die Daten in die Entwicklung der nächsten Generation einfließen.

Ähnlich positive Effekte erwarten Experten für Logistik und Warenwirtschaft. Die Tatsache, dass ein Kurierdienst oder eine weltumspannende Spedition jeden Laster, jeden Container und jedes noch so kleine Paket live verfolgen kann und diese Daten mit dem Absender und Empfänger teilt, hilft bei der Optimierung der Routenplanung. Das spart Zeit und Ressourcen – von Diesel bis zu Überstunden. Wer einen solchen Lieferwagen fährt, wird sich allerdings über die ständige Bespitzelung und den daraus resultierenden Zeitdruck sorgen. Buchhaltung und Controlling schließlich haben ebenfalls brennendes Interesse, diesen Datenozean anzuzapfen. Sie sind in der Lage, Prognosen zu entwickeln, Risikomodelle durchzuspielen und Betrugsfälle schneller zu erkennen.

GROSSE CHANCEN FÜR ALLE BRANCHEN

Von diesen Chancen können fast alle Industrien und Branchen profitieren, vom Einzelhandel und Maschinenbau über Pharmafirmen bis zum Gesundheitswesen und dem öffentlichen Dienst. Verlage und Werbeagenturen sind bereits dabei, sich Programme zunutze zu machen, die ursprünglich für den Börsenhandel entwickelt wurden, um heute damit Online-Annoncen in Millisekunden zu platzieren. Große Supermarktketten wie Tesco in Großbritannien oder Wal-Mart in den USA sind so vernetzt, dass sie ihre Zulieferer virtuell ins Ladenregal blicken lassen.

Insgesamt, schätzt McKinsey, kann der Einsatz von Big Data die Marge eines Einzelhändlers um bis zu 60 Prozent steigern. Im Gesundheitswesen der USA veranschlagen die Berater den Mehrwert dank Big Data auf mehr als 300 Milliarden Dollar im Jahr. Für den öffentlichen Dienst in der Europäischen Union schließlich seien mit den vier Vs Effizienzsteigerungen im Wert von 100 Milliarden Euro möglich, ohne gestiegene Steuereinnahmen, weniger Rechnungsirrtümer und Betrugsfälle mit einzubeziehen.

Die Ökonomen Erik Brynjolfsson und Andrew McAfee von der Sloan School of Management in Massachusetts ermittelten in einer Studie von 179 Großunternehmen, dass datengetriebenes Management Produktivitätsgewinne von fünf bis sechs Prozent freisetzt. Unternehmen beziehen diese neue Art der Wertschöpfung zunehmend in ihr Kalkül ein. Ein Fünftel aller britischen Großunternehmen gaben in einer Umfrage an, ihre Daten bereits als Aktivposten in der Bilanz zu führen.

BIG DATA IST BIG BUSINESS

AUCH DER STAAT INVESTIERT IN BIG DATA

Neben den oben beschriebenen kommerziellen Anbietern, die das Versprechen von mehr Effizienz, Wachstum und Gewinn mit ihrer Hardware und Software einlösen wollen, versuchen auch staatliche Initiativen, die Verbreitung und Verwendung von Big Data voranzutreiben. In den USA investiert die Regierung rund 200 Millionen Dollar in eine landesweite „Big Data Research and Development Initiative", an der sechs Ministerien und Behörden beteiligt sind. In eine ähnliche Richtung zielt das THESEUS-Projekt des Bundesministeriums für Wirtschaft und Technologie. Das seit 2006/7 laufende Forschungsprogramm bringt 60 Partner aus Wissenschaft und Wirtschaft zusammen, die gemeinsam den Zugang zu Informationen vereinfachen, Daten zu neuem Wissen vernetzen und die Grundlage für die Entwicklung neuer Dienstleistungen im Netz schaffen wollen.

Sechs Partner aus der Wirtschaft und dem Hochschulbereich unter Führung der TU Berlin haben einen cloudbasierten Marktplatz für Informationen und Analysen (MIA) geschaffen, der sich auf das deutschsprachige Web und andere Datenquellen konzentriert. „Stellen Sie sich vor, Sie könnten für ein paar Euro im Monat auf den Datenbestand des deutschsprachigen Webs zugreifen. Welche Webanwendung würden Sie bauen?" fragt MIA in einer Präsentation. „Ein nachhaltiger, cloudbasierter Informationsmarktplatz ermöglicht insbesondere innovativen Startups und KMUs in Deutschland und Europa Teilhabe an der Informationsökonomie."

Dazu bedarf es neuer Fähigkeiten, neuer Studiengänge und neuer Berufsbilder wie dem des „Datenwissenschaftlers". Die Unternehmensberatung McKinsey schätzt, dass alleine in den USA in den kommenden fünf Jahren zwischen 140.000 und 190.000 Arbeitnehmer mit gut ausgebildeten Analysekenntnissen gefragt sein werden, sowie weitere anderthalb Millionen Manager, die zumindest ein grundlegendes Verständnis von Big Data haben, um ihrer Arbeit nachzugehen. Erste Hochschulen bieten Studiengänge in dieser neuen Disziplin an, während immer mehr Unternehmen ihren Mitarbeitern Zugang zu modernen Analysewerkzeugen gewähren. Wer damit aufwächst, seinen Fotostream und seine Facebook-Freunde zu pflegen oder sportliche Aktivitäten im Netz zu teilen, wird sich auf dem neuen Armaturenbrett der Big-Data-Arbeitswelt gut zurechtfinden. Vorausgesetzt, das Elternhaus, Schulen und andere Teile der Gesellschaft haben das Fundament für moderne Medienkompetenz sowie ein gesundes Misstrauen gegenüber Big Data gelegt.

GEFAHREN UND NACHTEILE FÜR DEN NUTZER, ODER: DIE ETHIK DER DATEN

Der Weg zur buchstäblich „selbst-bewussten" Volkswirtschaft und datengetriebenen Gesellschaft ist natürlich nicht nur mit goldenen Geschäftsideen und Diensten gepflastert. Wenn immer mehr Lebensbereiche von großen Datenmengen getrieben werden und Algorithmen Menschen Entscheidungen abnehmen, tun sich eine ganze Reihe schwieriger Fragen auf.

? **Wem gehören die Daten, die Menschen und ihre Geräte erzeugen?**

? **Wer hat das Recht, diese Daten zu sammeln, zu bündeln und auszuwerten?**

? **Wo werden sie gelagert und wie werden sie übermittelt?**

? **Wie hat ein „Datensubjekt" die Gelegenheit, sie einzusehen und ihre Korrektur oder Löschung zu verlangen?**

? **Wer wird an der Umwandlung und Veredelung von Small Data zu Big Data verdienen?**

? **Wer sorgt dafür, dass Datenschutz und Privatsphäre gebührende Beachtung finden?**

? **Wie können sich die Teilnehmer in einer globalen Wirtschaft auf miteinander vereinbare Regeln und Gesetze einigen?**

? **Wer behält die Software im Auge, damit sie die Entscheidungsfreiheit der Menschen nicht beschneidet?**

? **Wie verändert das ständige Vernetztsein den Menschen und seine Kultur?**

Über diese Fragen denken Kommissare und andere Beamte vor allem der Europäischen Union laut nach und liefern sich mit Firmen, die an Daten verdienen, hitzige Debatten vor und hinter den Kulissen. Der Philosoph und Unternehmensberater Kord Davis ist einer der ersten, der sich in einem Buch über die „Ethik von Big Data" Gedanken gemacht hat. Der stete Datenstrom, den wir alle erzeugen und oft ohne unser Wissen ins Netz pumpen, schafft einen neuen Gesellschaftsvertrag, argumentiert Davis. Er wirft Fragen nach der Vertraulichkeit und der Vertrauenswürdigkeit der Daten auf und was ihre Verwendung durch Dritte für die Identität und Reputation des Einzelnen bedeuten. Diese Fragen sind in Europa ein großes Thema, aber sie drohen in den USA, wo die meisten Big-Data-Anbieter angesiedelt sind, aufgrund der Begeisterung für technische Innovation übersehen zu werden. Meist entspinnt sich eine Debatte in den Medien, unter Aufsichtsbehörden und Gesetzgebern erst dann, wenn Datenlecks intime Details von Tausenden oder Millionen Verbrauchern, Angestellten oder Patienten ins Netz spülen. Der Jurastudent Max Schrems etwa startete seine Kampagne „Europe vs. Facebook" aus eigenem Antrieb und bewegte bislang rund 40.000 EU-Bürger dazu, vom sozialen Netzwerk Einblick in die über sie erhobenen Daten zu verlangen.

Deswegen bietet Big Data als „neues Öl" eine weitere Parallele: Der Run auf den neuen Rohstoff wird fast unweigerlich zu Sicherheitspannen, „Umweltverschmutzung" und Katastrophen führen, bis strengere Regelungen eingeführt werden. Gesetze und kulturelle Normen haben mit den technischen Möglichkeiten nicht Schritt gehalten. Das belegen die Debatten um Datenerhebung und personalisierte Online-Werbung in den USA, wo regelmäßig Firmen abgemahnt oder zu Bußgeldern verurteilt werden, weil sie Kundendaten ungefragt und unerlaubt erheben.

GEFAHREN UND NACHTEILE FÜR DEN NUTZER, ODER: DIE ETHIK DER DATEN

DISKUSSION UM DEN DATENSCHUTZ

Gleiches gilt für die neuen Datenschutzrichtlinien der EU, die weitaus strenger sind als vergleichbare Regelungen in den USA. Hier wird Big Data in naher Zukunft für hitzige Auseinandersetzungen sorgen, da US-Unternehmen auch bei europäischen Kunden ungehindert Daten einsammeln wollen, möglichst ohne sich an die enger gefassten Vorschriften zu halten. Selbst so grundlegende Konzepte wie das Brief- oder Fernmeldegeheimnis müssen neu definiert werden, wenn E-Mails oder biometrische Daten durch Dutzende von Servern fließen, deren Betreiber sie zu Zwecken der Prozessoptimierung analysieren. Selbst wenn ein Dritter nicht wirklich mitliest, die Auswertung der Metadaten genügt, um individuelle Informationen preiszugeben. Wem gehören diese Daten und wie können Nutzer sicherstellen, dass sie nicht in die Hände unbefugter Dritter gelangen oder kontrollieren, dass ihre Daten tatsächlich auf Verlangen gelöscht werden? Das etwa ist eines der Probleme des in der EU anvisierten Rechts darauf, vergessen zu werden. Es ist keineswegs klar, ob ein solches Ansinnen technisch überhaupt machbar ist, wenn sich die Datenspuren jedes Nutzers in alle Winkel des Webs zerstreuen und mit anderen Datensätzen neue Kombinationen eingehen. Denken Sie nur an die Metadaten eines Bildes, auf dem Sie abgebildet sind und dem andere Nutzer Kommentare hinzugefügt haben. Wem gehört dieses Bild und die damit verbundenen Informationen? Wer soll prüfen, ob ein eventueller Löschungsanspruch berechtigt ist und wer soll ihn durchsetzen und kontrollieren?

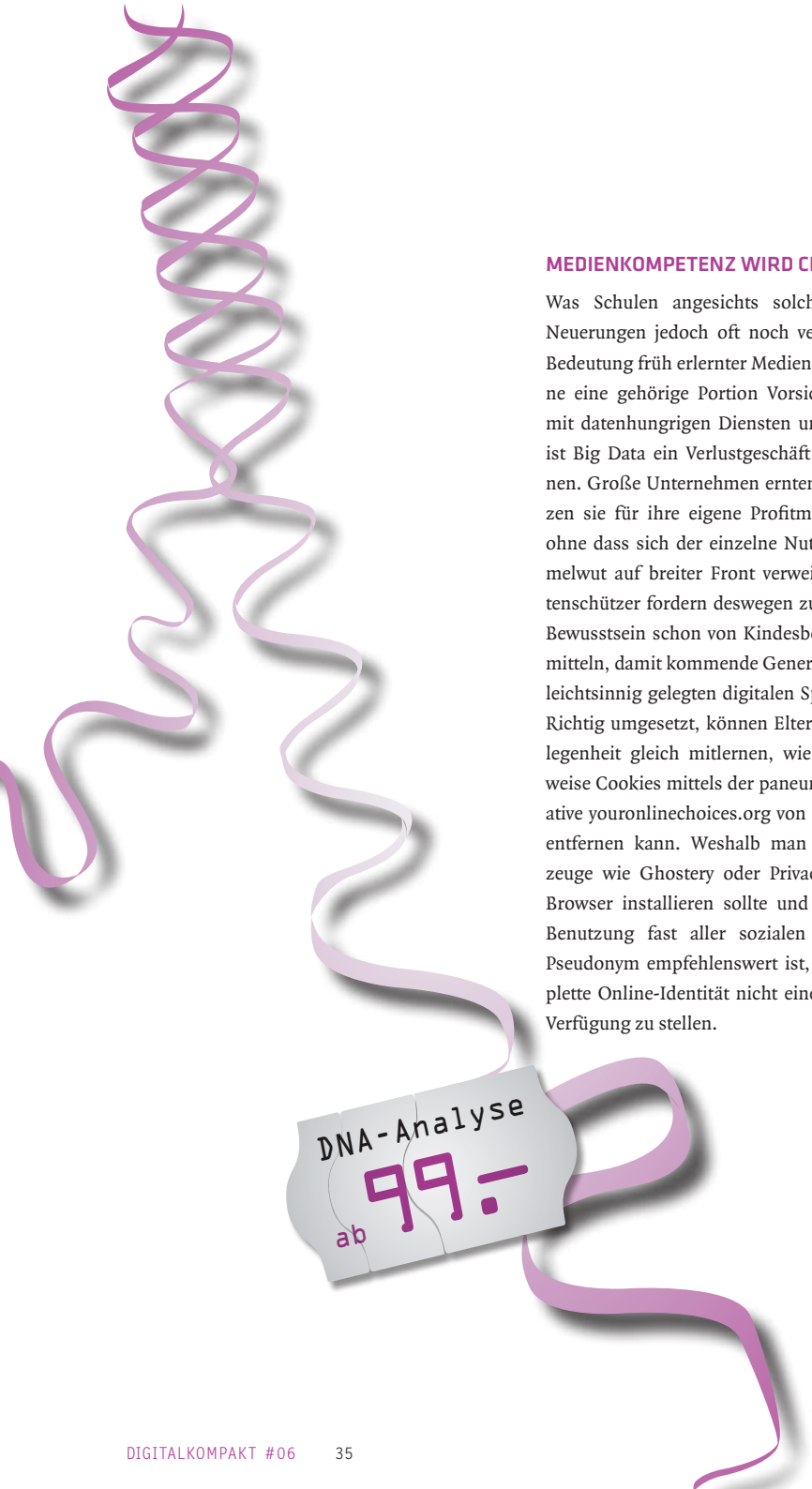
DIE NEUEN DIENSTE SIND SCHON DA

Hinzu kommt, dass die Abwägung zwischen Wohl und Wehe von Big Data dabei je nach Gesellschaft anders ausfällt. In den USA ist Privatheit kein Grundrecht, sondern ein kommerzielles Gut, das man einklagen kann. Diese Einstellung erklärt viele der neuen Dienste, die auf Big Data basieren und Millionen von Verbrauchern in ihren Bann ziehen. Ohne Big Data gäbe es kein Google Streetview und keine automatische Gesichtserkennung bei sozialen Netzwerken – beides bequeme Erfindungen, die zugleich erhebliche Konsequenzen für die Privatheit haben. Ohne Big Data könnten neue Firmen wie 23andme.com in Kalifornien keine DNA-Analyse für Otto Normalverbraucher für 99 Dollar anbieten. Damit soll man angeblich Gesundheitsrisiken besser überblicken und genetisch entfernte Verwandte finden können. Das sieht auf den ersten Blick revolutionär aus. Doch wer seine Speichelprobe an solche Firmen einschickt, stellt sein gesamtes Erbgut einem Unternehmen zu Verfügung – ohne sichergehen zu können, was die Firma langfristig damit anstellen will. Sollte ein solcher Dienst zur Gen-Analyse etwa gehackt oder verkauft werden, können brisante persönliche Daten ins Internet gelangen oder von anderen kommerziellen Anbietern ausgewertet werden. Das wirft erhebliche Risiken der sozialen oder genetischen Diskriminierung auf.

Zugleich ist auch wahr, dass sich dank Big Data Bildungseinrichtungen besser denn je auf den einzelnen Schüler oder Studenten einstellen können, beim Lehrplan, bei Stipendien, bei der Nachhilfe. Denn so könnten Verlage und Dozenten plötzlich sehen, welche Texte ihre Schüler und Studenten wann wie lange lesen oder welche Fehler sie am häufigsten machen.

MEDIENKOMPETENZ WIRD CHEFSACHE

Was Schulen angesichts solcher technischen Neuerungen jedoch oft noch vergessen, ist die Bedeutung früh erlernter Medienkompetenz. Ohne eine gehörige Portion Vorsicht im Umgang mit datenhungrigen Diensten und Programmen ist Big Data ein Verlustgeschäft für den Einzelnen. Große Unternehmen ernten Daten und setzen sie für ihre eigene Profitmaximierung ein, ohne dass sich der einzelne Nutzer dieser Sammelwut auf breiter Front verweigern kann. Datenschützer fordern deswegen zu Recht, Online-Bewusstsein schon von Kindesbeinen an zu vermitteln, damit kommende Generationen nicht an leichtsinnig gelegten digitalen Spuren scheitern. Richtig umgesetzt, können Eltern bei dieser Gelegenheit gleich mitlernen, wie man beispielsweise Cookies mittels der paneuropäischen Initiative youronlinechoices.org von seinem Rechner entfernen kann. Weshalb man Blocking-Werkzeuge wie Ghostery oder PrivacyFix in seinem Browser installieren sollte und warum bei der Benutzung fast aller sozialen Netzwerke ein Pseudonym empfehlenswert ist, um seine komplette Online-Identität nicht einem Anbieter zur Verfügung zu stellen.



GEFAHREN UND NACHTEILE FÜR DEN NUTZER, ODER: DIE ETHIK DER DATEN

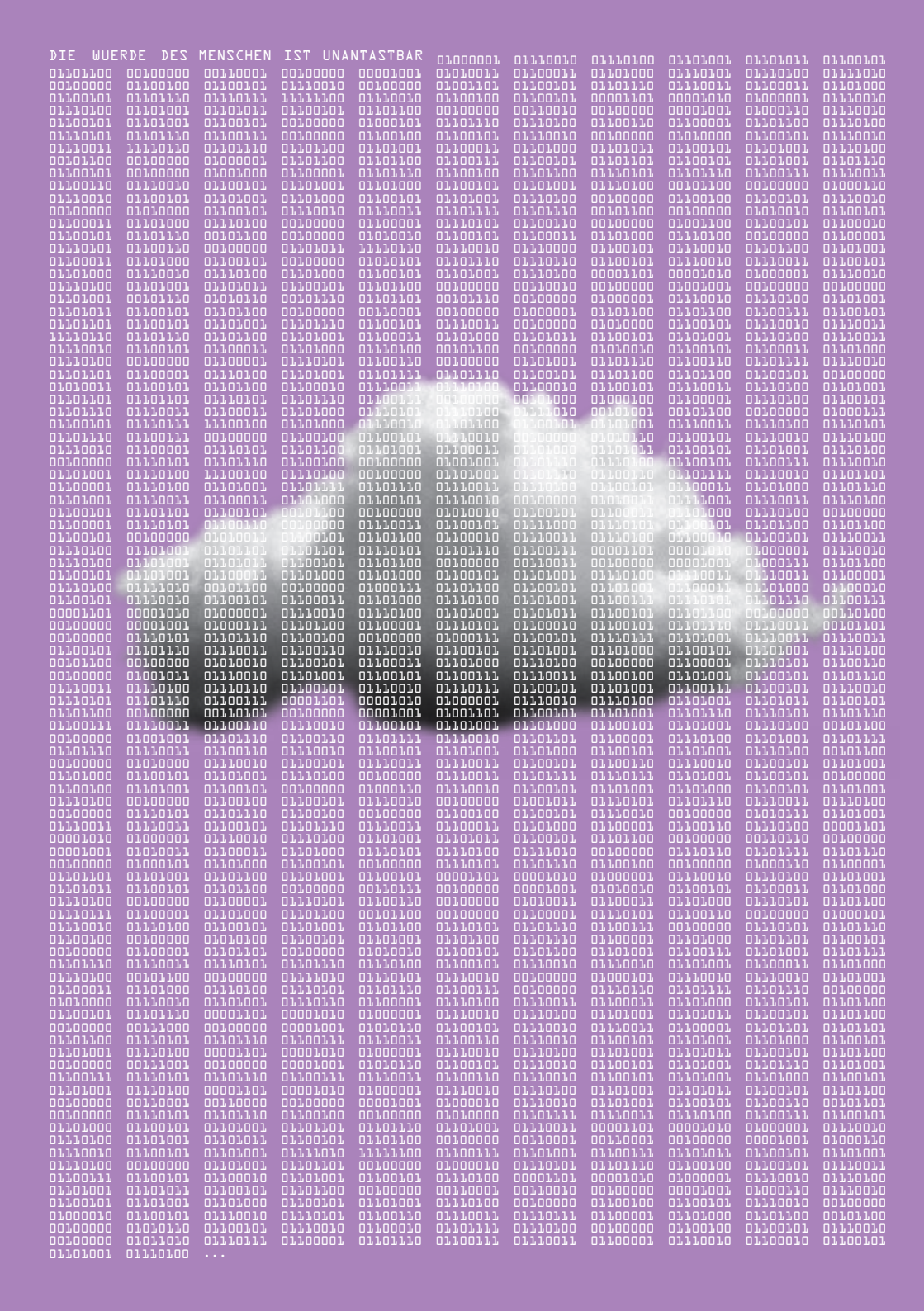
HOFFNUNGEN UND ÄNGSTE RUND UM BIG DATA

Schlimmstenfalls wird Big Data eine Welt des maschinenlesbaren, gläsernen Menschen schaffen, die das düstere Bild vom „Big Brother" durch etwas noch Bedenklicheres ersetzt. Uns droht eine Welt der „Little Brothers", in der jeder des Anderen Aufseher wird. In diesem Panoptikum besitzt jeder Gefangene ein Smartphone. Je nach gesellschaftlichem Hintergrund, Einkommen und Gesundheitszustand sieht jeder von uns andere Preise im Ladenregal oder Webshop, werden ihm bestimmte Suchergebnisse vorenthalten oder sogar ein Arbeits- oder Studienplatz verweigert. Selbst Werbung auf der Grundlage des individuellen Erbguts ist längst kein Hirngespinst mehr.

Als die US-Stiftung Pew Research im Sommer 2012 mehr als 1.000 Technologen und prominente Internetnutzer über ihre Hoffnungen und Ängste rund um Big Data befragte, stimmten gerade einmal 53 Prozent folgender Aussage zu: „Die Analyse großer Datensätze durch Menschen und Maschinen wird die gesellschaftliche, politische und wirtschaftliche Intelligenz bis 2020 steigern. Unter dem Strich ist der Aufstieg von Big Data ein großes Plus für fast alle Bereiche der Gesellschaft." Demgegenüber gaben 38 Prozent zu Protokoll, Big Data werde mehr Probleme schaffen als lösen: „Die Existenz großer Datensätze wird falsches Vertrauen in unsere Prognosemöglichkeiten erzeugen und viele Menschen zu bedeutenden und schmerzlichen Fehlern verleiten. Big Data wird von einflussreichen Menschen und Institutionen missbraucht werden, die eigennützige Ziele verfolgen."

Wenn sich schon Informatiker, Ökonomen und Technologie-Unternehmer uneins sind, ist es wichtig, rechtzeitig eine Debatte über Risiken und Nebenwirkungen von Big Data zu führen, an der sich die gesamte Gesellschaft beteiligt. Sonst schaffen Programmierer, Ingenieure und Unternehmer, die von Big Data profitieren, vollendete Tatsachen, mit denen sich alle Nutzer arrangieren müssen. Das schon erwähnte EU-weit anvisierte „Recht, vergessen zu werden" zeigt eindringlich auf, welche ethischen und juristischen Dilemmas auf dem Weg in die Big-Data-Welt warten.

Die ersten 12 Grundrechte des Grundgesetzes im Binärcode >>


DIE WUERDE DES MENSCHEN IST UNANTASTBAR

LEBEN IN DER BIG-DATA-WELT

Aufhalten lässt sich der Wandel zum datengetriebenen Leben nicht. Menschen passen ihren Lebenswandel den ständig datenhungrigen Geräten an, die ihr Leben einfacher machen sollen, aber bemerken erst später, welchen hohen Preis sie für den Komfort der immer verfügbaren Dienste bezahlen. Nämlich die Gefahr, quer durchs Netz und von ihnen unbekannten Anbietern verfolgt, umworben und ausgespäht zu werden. Auf der Strecke bleibt oft die Gelegenheit, informierte Entscheidungen zu treffen und die Einspeisung der eigenen Daten zu verweigern.

Die nächste Handy-Generation etwa wird gesprochene Fragen verstehen, auch wenn das Gerät ausgeschaltet auf dem Nachttisch liegt. Es kann menschliche Sprache an einen Server schicken, der sie in Sekunden interpretiert, Programme auf dem Gerät ansteuert und es mit Daten aus dem Web füttert. Das Handy wird wissen, wie lange wir am Nachmittag nach Hause unterwegs sein werden und uns rechtzeitig warnen, dass wir den Elternabend verpassen. Jede Handlung oder jede unterlassene Handlung schaffen neue Datensätze, die umgehend in den Datenteppich eingeflochten werden – vom Tritt aufs Gaspedal, wenn die Ampel schon tiefgelb ist, bis zur Schlagzeile, über der wir mit der Maus länger als üblich verharren. Automatische Spionage wird so zum serienmäßigen Angebot der Unterhaltungselektronik, der sich ein Nutzer nur dann effektiv widersetzen kann, wenn er auf die Benutzung moderner Technologie verzichtet.

Neben einigen positiv zu bewertenden Aspekten wirft der Erfolg von Big Data neue Probleme für Bürger und Verbraucher auf. Wenn wir – oft ungefragt – zur Komponente in einem maschinenlesbaren System reduziert werden, dann beschneidet das zutiefst menschliche Werte wie Privatheit, kritische Meinungsbildung und den freien Willen.

Es bleibt aber auch die Aufgabe jedes Einzelnen, seine Daten wie ein wertvolles Gut für sich zu behalten anstatt sie jeder Seite und jedem Dienst zu übereignen. Jeder Nutzer sollte Auskunft darüber verlangen, wer seine Daten für welche Zwecke einsammeln und auswerten will. Das kann man heute oft nur mit Hilfsmitteln für den Browser oder eigens installierter Software tun bzw. indem man sich neuen Apps verweigert.

Doch die Innovation gebiert bereits neue Technologien als Antwort auf die ungelösten Fragen von Big Data. Schon bald werden neue Dienste als Gewährsleute oder Privatheits-Makler des Big-Data-Zeitalters auftreten. Mit ihnen wird ein Nutzer seine Daten verbergen, vernichten oder – so er es will – zu seinen Konditionen verkaufen. Die Industrie, vor allem Werbetreibende, werden mit Empörung reagieren und neue Wege finden, wertvolle Daten zu sammeln, wenn sie nicht mehr freiwillig fließen. Gesetzgeber, Verbraucherschützer und Behörden sind deswegen gefragt, die Rechte des Verbrauchers auf informationelle Selbstbestimmung zeitgemäß zu definieren und zu verbriefen. Dazu gehört es gerade in Europa auch, die Niederlassungen weltweit operierender Firmen in die Pflicht zu nehmen, die ihre großen Gewinne aus vielen Einheiten kleiner Daten von EU-Bürgern schöpfen. Dazu gehört außerdem, den gesellschaftlichen Diskurs zu fördern, auf Chancen und Risiken hinzuweisen, die Rechte des Verbrauchers deutlich und die Problemlage öffentlich zu machen. Digitale Selbstverteidigung gepaart mit mehr und früh beginnender Ausbildung zur Medienkompetenz und zeitgemäßen Regelungen zum Datenschutz – alle diese Elemente zusammen sind nötig, um Big Data zu einem Phänomen der Zukunft zu machen, das allen Menschen greifbaren Nutzen bietet anstatt sie auf reine Rechengrößen zu reduzieren.

DER AUTOR DIESER AUSGABE

Steffan Heuer ist US-Korrespondent des Wirtschaftsmagazins brandeins und berichtet aus San Francisco über Innovation und Humankapital in der Wissensgesellschaft. Er ist Ko-Autor des Buchs „Mich kriegt ihr nicht! Gebrauchsanweisung zur digitalen Selbstverteidigung“, das im Februar 2013 im Murmann-Verlag erscheinen wird.

Steffan Heuer (©Bildnachweis: Thomas Kern)

GLOSSAR

Algorithmus Eine Abfolge von Handlungsschritten, um ein Problem zu lösen. Nach einem persischen Gelehrten benannt, bilden Algorithmen die Grundlage für die Lösung einer einfachen Aufgabe (Erst die erste Seite lesen, dann die zweite...) bis hin zu hochkomplexen Computerprogrammen, um große Datensätze zu analysieren und ständig neues Feedback zu berücksichtigen.

API Application Programming Interface oder Programmier-Schnittstelle, die es einem Stück Software erlaubt, bei entsprechender Zugangsberechtigung selbstständig Daten von einem anderen Programm abzurufen.

Big Data Sammelbegriff für Datenmengen, die sich mit herkömmlicher Hardware und Software nicht mehr bewältigen lassen, wobei diese Grenze je nach Definition willkürlich gezogen wird und nicht an der Dateigröße festgemacht werden kann.

Business Intelligence Datenanalyse, um Entscheidungsprozesse in einem Unternehmen zu unterstützen. BI wird zunehmend als Cloud-Dienst angeboten, bei dem sowohl die Speicherung und Aufbereitung von Daten als auch deren Bearbeitung auf Servern im Internet erfolgt. Lediglich die Visualisierung oder die Manipulation der Ergebnisse läuft auf einem örtlichen Rechner.

Mash-Up Eine dynamische Ad-hoc-Verbindung mehrerer Dienste und Datenquellen, um einen neuen Dienst zu schaffen. Ein Beispiel wäre die Verknüpfung von Immobilienanzeigen in einer Stadt mit Google Maps und den Verbrechensstatistiken der Polizei. Mash-Ups erfolgen über APIs.

Small Data Die einzelnen, kleinen Datensätze, die Geräte, Sensoren, Webdienste oder Menschen erzeugen. Sie werden zu größeren Datensätzen gebündelt oder als steter Strom in Systeme eingespeist, wo sie sich zu Big Data vereinigen.

Strukturierte, semistrukturierte und unstrukturierte Daten Daten sind auf Lateinisch nichts anderes als „gegebene Zeichen“ mit einem gewissen Informationsgehalt. Je nach Art und Quelle werden sie als strukturiert, semistrukturiert oder unstrukturiert klassifiziert.

Strukturierte Daten sind dabei die ordentlichsten Mitglieder dieser Familie. Sie haben klar definierte Eigenschaften und sind einem festen Feld in einer Datenbank oder einer Tabelle zugeordnet.

Semistrukturierte Daten haben zwar feste Eigenschaften, anhand derer sie identifiziert werden können. Aber sie sind halbe Freigeister, da sie nicht länger einer bestimmen Zelle in einer Tabelle zugeordnet sind.

Unstrukturierte Daten, das wichtigste Rohmaterial der Big-Data-Welt, sind die Enfants Terribles der Informatik. Es handelt sich um beliebige Objekte wie Texte, PDFs, E-Mails, Bild- oder Videodateien. Ihnen muss Software (oder ein menschlicher Bearbeiter) erst die Eigenschaften zuweisen, um sie weiter verarbeiten zu können.

Vier Vs Die vier Kriterien für Big-Data-Anwendungen in Anlehnung an die Anfangsbuchstaben der englischen Begriffe Volume, Velocity, Variety und Value. Dabei geht es um die Datenmenge, die Geschwindigkeit, mit der Daten eingehen und verarbeitet werden, die Vielfalt der Datensätze und schließlich den Wert, der sich aus ihrer Aufbereitung und Analyse gewinnen lässt.

IMPRESSUM

Herausgeber
Landesanstalt für Medien
Nordrhein-Westfalen (LfM)
Zollhof 2
40221 Düsseldorf
Tel.: 0211. 77 00 7-0
Fax: 0211. 72 71 70
www.lfm-nrw.de
info@lfm-nrw.de

Verantwortlich für den Inhalt
Dr. Thomas Bauer,
Leiter Projektinitiative NRW digital

Autor
Steffan Heuer

Redaktion
Dr. Dörte Hein, Sabrina Nennstiel, David Gerl (LfM)

Gestaltung, Fotografie und Illustration
Fritjof Wild, serviervorschlag.de

Bildnachweis
S. 02,04,08 © tom-fotolia.com
S. 07 © electriceye-fotolia.com
S. 27 © tiero-fotolia.com
S. 39 © Thomas Kern

Druck
Börje Halm

Landesanstalt für Medien
Nordrhein-Westfalen (LfM)
Zollhof 2
40221 Düsseldorf
Postfach 10 34 43
40025 Düsseldorf

Telefon
02 11 / 7 70 07-0
Telefax
02 11 / 72 71 70
E-Mail
info@lfm-nrw.de
Internet
http://www.lfm-nrw.de