Vom Monopol auf Daten ist abzuraten

Big Data ist Big Business. Die Sammlung und die Verknüpfung von Informationen über Menschen im Netz bringen Milliarden ein. Und die Methoden werden immer abgefeimter und perfekter. Es ist Zeit, dagegen vorzugehen.

Von Thomas Hofmann und Bernhard Schölkopf

Big Data scheint mit einer Heftigkeit und Unaufhaltsamkeit über uns gekommen zu sein, die uns ratlos zurücklässt. Von technophiler Seite wird eine Anpassung unserer Werte empfohlen, am anderen Ende des Spektrums wird entschlossen zu Widerstand aufgerufen. Wir führen keine Debatte über Technik, heißt es, sondern eine politische Debatte. Doch vielleicht hat unsere Hilflosigkeit auch damit zu tun, dass wir die technischen und theoretischen Grundlagen nicht hinreichend bedenken. Gesetzgeberischer Eingriff und Regulation erfordert nicht nur Klarheit in den Zielen, sondern auch Verständnis der Entwicklungen und der zur Verfügung stehenden Instrumente.

Wir sind verstrickt in ein Geflecht von Datenerfassung, -verknüpfung und -nutzung, dessen Entstehung mit der Vermessbarkeit des Menschen zum Zwecke von Werbung und Konsumförderung zusammenhängt. Die Online-Welt hat zu einer flächendeckenden, quantitativen Instrumentierung geführt, die es erlaubt, die Effektivität von Werbung mit hoher Präzision zu messen und auf individueller Ebene zu optimieren. Diese Revolution der Werbebranche fand ihren maßgeblichen Antrieb in der Websuche. Eine Suchanfrage bekundet eine wache Aufmerksamkeit, einen Bedarf im Hier und Jetzt. Dieser Zugang stellt einen Wert dar, der umso größer ist, je besser die Nutzerintention erfasst wird. Es geht darum, mit hoher Genauigkeit abzuschätzen, welche Inhalte mit welcher Wahrscheinlichkeit an die Aufmerksamkeit des Benutzers anknüpfen, so dass sie zu einem weiterführenden Click führen. Suchmaschinen sind gigantische Systeme der Vorhersage, die für jede Anfrage in Echtzeit eine Auswahl aus einem Milliardenbestand von Links und Werbebotschaften präsentieren.

Wie ist eine solche quantifizierte Vorhersage über das unendliche Spektrum potentieller Suchanfragen überhaupt möglich? In den Anfängen beschränkte sich das Ranking auf Wortübereinstimmungen zwischen Anfragen und Websites, die man mit Hilfe eines Index ermittelt. Dann nahm man Eigenschaften hinzu, die sich aus der Verknüpfung des Webs ableiten, wie in Googles Page-Rank-Algorithmus, der die Popularität einer Website aus den eingehenden Links berechnet. Alle diese Inhalte und Daten waren öffentlich zugänglich. Das hat sich grundlegend geändert. Mittlerweile sind es die Clicks (oder auch „Click-Streams") der Nutzer, die gemeinsam mit anderen „Signalen" durch datengetriebene Lernmethoden die Such- und Werberesultate bestimmen. Jeder Click enthält wertvolles Feedback, durch das sich Suche und Werbung optimieren lassen. Diese nichtöffentlichen Daten werden von uns allen täglich erzeugt (aber meist nicht explizit an Dritte übertragen).

Selbst kleine Verbesserungen können bei Milliardenumsätzen zu beachtlichen Gewinnzuwächsen führen, und so hat sich eine Wissenschaft und Ingenieurskunst etabliert, die beständig darauf aus ist, die Optimierung ad extremum zu führen. Dies geht manchmal durch Verbesserung von Algorithmen, aber meist einfacher durch die Erschließung von reichhaltigeren Daten. Daten und nicht nur Algorithmen werden so zu einer Quelle von Innovation und Wertschöpfung. Demographie, vergangenes Kaufverhalten, persönliche Interessen, soziale Verbindungen: Alles ist willkommener Input in einer Welt, die Daten systematisch in Umsätze transformiert. Da diese Innovationsprozesse kontinuierlich sind und sich von Quartal zu Quartal immer wieder selbst überbieten sollen, herrscht eine Tendenz vor, Daten auf Verdacht zu speichern. Was heute noch nicht nutzbar ist, kann vielleicht morgen schon Verwendung finden. Diese Denkweise ist Webfirmen selbstverständlich.

Das Datensammeln und -verknüpfen betrifft nicht nur die Websuche, sondern auch das Surfen im Web, denn viele Sites sind Teil von Werbenetzwerken. Mit Hilfe von Cookies werden Nutzer über ihre Browser markiert. Sie bleiben über Monate identifizierbar, was es erlaubt, Interessenprofile aufzubauen und demographische Dimensionen zu erschließen. Diese Informationen werden Werbeagenturen verfügbar gemacht, die dann Werbung gezielt an bestimmte Nutzersegmente ausliefern können. Drittfirmen haben sich darauf spezialisiert, den Datenmix durch Zusatzdaten aufzuwerten, etwa indem sie Menschen identifizieren, die Anzeichen einer Kaufintention zeigen. Alle diese Systeme bilden ein Datenökosystem und sind programmatisch über APIs (Application Programming Interfaces) in Echtzeit miteinander verbunden.

So ist ein Schattenmarkt für Nutzerdaten entstanden. Daten von Nutzern werden verknüpft und ausgetauscht, ohne dass die Nutzer selbst dies verstehen oder nachverfolgen könnten. Man hat wohl AGBs zugestimmt, aber wusste man da, was man tat? Beliebt sind auch Remarketing-Listen, die es erlauben, Nutzer als einer Gruppe zugehörig zu „taggen". Diese Gruppe kann man dann gezielt mit Werbenachrichten versorgen, welche die Nutzer praktisch überall im Web erreichen können. Um die Datenprivatheit nicht zu verletzen, weisen die Listen eine Mindestgröße von hundert Nutzern auf, so dass der Einzelne seine digitale Nacktheit mit dem Feigenblatt einer Gruppe Gleichgetaggter abdecken kann. Vor ein paar Jahren war diese von der Branche im Wege der Selbstregulierung festgelegte Grenze noch die Zahl Fünfhundert. Privatsphäre sieht anders aus.

Doch heute beschränkt sich das große Datensammeln längst nicht mehr auf klassische Bildschirmcomputer. Eine Vielzahl von Sensoren in unseren Smartphones erlaubt es, eine wachsende Zahl von Aktivitäten zu erfassen. Mit „pervasive computing" und dem „Internet der Dinge" sind wir selbst Teil eines Datennetzes geworden, das die wirkliche und die Online-Welt umspannt und die Verknüpfbarkeit der Daten weiter entgrenzt: Daten darüber, welche Geschäfte ich wann aufsuche (Yelp), vor welchem Regal ich stehen bleibe (Hyper-Lokalisierung), welchen Menschen ich begegne (Bluetooth), welche Fortbewegungsmittel ich nutze (GPS), wann ich aufstehe (Kalender, Weckfunktion), wie ich mein Ei zum Frühstück verspeise (Google Glass). Es gibt keine naturgegebene Grenze, vor der die Datenerfassung haltmachen würde, weil alles schrittweise zur Vorhersagbarkeit und Beeinflussbarkeit menschlichen Verhaltens beiträgt.

All diesen Anwendungen von Big Data liegt ein Forschungsgebiet zugrunde, das im Zeitalter der Computer einen immensen Aufschwung erlebt hat: empirische Inferenz und maschinelles Lernen. Die aktuellen Entwicklungen sind komplex und verwenden Ergebnisse aus Informatik, Statistik, Numerik und anderen mathematischen Disziplinen, aber die Grundfragen sind einfach: Wie kann man aus Beobachtungsdaten auf zugrundeliegende Gesetzmäßigkeiten schließen?

Die Auswertung von Daten ist tief in unserem modernen Wissenschaftsbegriff verwurzelt. Exemplarisch möchte man an Kepler denken, der aus Tycho Brahes Beobachtungen erschloss, dass sich die Planetenpositionen am besten durch ein Modell erklären ließen, das den Planeten elliptische Umlaufbahnen um die Sonne zuwies und ihre Geschwindigkeiten und Umlaufzeiten einfachen Gesetzmäßigkeiten unterwarf.

Doch nicht nur Wissenschaftler betreiben empirische Inferenz. Die Verwertung sensorischer Daten wird täglich durch Inferenzprozesse ermöglicht. Ein Kind lernt in seinem Leben eine Vielzahl von Objekten zu erkennen und zu unterscheiden oder handgeschriebene Ziffern zu lesen. Strukturell haben diese Probleme vieles gemeinsam: In beiden Fällen gibt es Beobachtungsdaten, in beiden Fällen wird eine Gesetzmäßigkeit erschlossen. Keplers Inferenzprozess kulminiert in einem einfachen Satz von Gleichungen, der Einsicht vermittelt und gleichzeitig Voraussagen ermöglicht. Der Lernprozess bei der Objekterkennung hingegen resultiert in einem neuronalen Netzwerk oder – falls er in einem Computer realisiert wird – in einem komplexen Satz mathematischer Funktionen oder Gleichungen. Im Gegensatz zu den Keplerschen Gesetzen sind diese nicht leicht zu verstehen, ermöglichen aber gleichwohl Voraussagen. Die Inferenzprozesse, die der visuellen Wahrnehmung zugrunde liegen, laufen nach Hermann von Helmholtz unbewusst ab. Für den Neurowissenschaftler Horace Barlow ist das Gehirn „nichts als ein statistisches Entscheidungsorgan".

Die Trennung zwischen unbewusster Inferenz bei Lebewesen und bewusster Inferenz beim Analysieren wissenschaftlicher Daten hat sich inzwischen überlebt. Im Zeitalter von Computern und Big Data wird sie ersetzt durch die Unterscheidung zwischen komplexen Modellen und einfachen Modellen. Komplexe Modelle werden automatisch aus Daten gelernt, ihr Ziel ist nicht Verständlichkeit, sondern nur noch die Tauglichkeit für Voraussagen. Wenn unsere Digitalkamera automatisch Gesichter erkennt und scharf stellt, ist es nebensächlich, dass diese Erkennung auf Funktionen mit Myriaden undurchschaubarer Rechenschritte beruht, solange sie nur – im Jargon des Forschungszweiges – „prädiktiv" ist und neue Beleuchtungsverhältnisse oder Kopfhaltungen „generalisiert". Einfache Modelle hingegen werden von Menschen erstellt und sind wesentlich für das reduktionistische Programm der Naturwissenschaften.

Komplexe Modelle liefern weit weniger Einsicht, werden daher oft als „Black Boxes" bezeichnet, sind aber dafür nicht nur in den Naturwissenschaften einsetzbar, sondern überall, wo es Beobachtungsdaten gibt. Sie sind Vorstufen dessen, was Isaac Asimov in seinen „Foundation"-Büchern als die künftige Psychologie bezeichnete: Sie versuchen, komplexe Prozesse wie die Handlungen von Menschen zwar nicht verständlich, aber zumindest statistisch vorhersagbar zu machen.

Prädiktive Modelle sind datengetrieben und dementsprechend datenhungrig. Man sollte aber bei Datenmengen zwischen der Anzahl der Beobachtungen und Dimensionalität differenzieren. Je mehr Messgrößen man verknüpft, desto höherdimensionaler die Daten und desto vielfältiger die möglichen Analysen. Wenn wir bei Tumorpatienten zusätzlich zur Histologie auf molekularbiologische Analysen des Tumorgewebes zurückgreifen, kann potentiell eine viel präzisere Diagnose erstellt werden. Werden die Daten höherdimensionaler, dann wächst jedoch die Anzahl der benötigten Daten (die Kardinalität) extrem.

In der Statistik wird vom „Fluch der Dimensionalität" gesprochen: Um eine Gesetzmäßigkeit hinreichend genau zu erschließen, sollte der Raum durch genügend Beobachtungen abgedeckt sein. Nun wächst das Volumen des Raums aber exponentiell mit seiner Dimensionalität. Wenn wir eine Messgröße hinzufügen, brauchen wir doppelt so viele Daten, bei zwei oder drei Messgrößen ist es schon ein Faktor vier beziehungsweise acht. Hohe Dimensionalität hat noch eine zweite, oft unerwünschte Nebenwirkung: Da das Volumen des Raums so schnell wächst, liegen die einzelnen Datenpunkte bei konstant gehaltener Kardinalität sehr bald weit auseinander und sind dann mit immer höherer Genauigkeit einzeln identifizierbar. Wenn man von einer Person nicht nur das Such- und Surfverhalten kennt, sondern auch die typischen Bewegungsmuster, gibt es schnell keine andere Person mehr, die in allen Aspekten mit ihr verwechselbar wäre. In solchen Datenräumen gibt es keine Anonymität.

Die Wirtschaftsfelder, die diese Revolution in der automatischen Konstruktion von Modellen am stärksten verinnerlicht haben, sind das Online-Marketing und die Finanzmärkte. In beiden Fällen führt schon eine gering erhöhte prädiktive Genauigkeit zu großen Gewinnen. Die Trivialität dieser Anwendungen wird manchmal beklagt; so bemerkte Jeff Hammerbacher, der 1982 geborene frühere Chefingenieur von Facebook: „Die besten Köpfe unserer Generation denken darüber nach, wie sie die Leute zum Klicken auf Anzeigen verleiten können." Aber dies sollte nicht darüber hinwegtäuschen, dass die empirische Inferenz mannigfache Anwendungen in den Wissenschaften hat, von der Suche nach Planeten außerhalb des Sonnensystems bis hin zur personalisierten Medizin.

Gleichzeitig wirft sie faszinierende wissenschaftliche Grundlagenprobleme auf: Wie kann man automatisch Beobachtungen und A-priori-Wissen über die erwartete Komplexität von Modellklassen kombinieren? Wie kann man Modelle konstruieren, die nicht nur statistische Abhängigkeiten finden, sondern auch kausale Aussagen machen? Laptops und Laptop-Rucksäcke werden bei Amazon oft gemeinsam gekauft – die Verkäufe sind also stark korreliert. Dies führte dazu, dass Amazon in der Vergangenheit beim Kauf eines Laptop-Rucksacks den Kunden empfahl, gleich noch einen Laptop zu kaufen. Die Kausalität ist aber umgekehrt: Der Laptop „verursacht" den Wunsch nach dem Rucksack, und nur in dieser Richtung ergibt die Empfehlung Sinn.

Firmen wie Google, Amazon, Facebook & Co sind von Datenhunger und Phantasie beim Verknüpfen von Daten getrieben. Reichtum an Daten führt zu besseren Online-Diensten und zu mehr Nutzern, was den Datenreichtum weiter erhöht. Ein solcher Daten-Kapitalismus wird von diesen Protagonisten in Perfektion praktiziert und hat ihnen eine beinahe konkurrenzlose globale Dominanz eingebracht. Im Falle von Google liegt dies nicht an Algorithmen im Eigentum des Unternehmens, deren Veröffentlichung manchmal verlangt wird, sondern an der in Umfang und Reichhaltigkeit einzigartigen Datenbasis. Allein mit einem cleveren Algorithmus – man denke an das Page-Rank der neunziger Jahre – wäre es heute schwierig, erfolgreich zu sein.

Die Situation würde sich erst grundlegend ändern, wenn diese Daten im Firmeneigentum, die zum überwiegenden Teil durch das Tun und Zutun der Nutzer entstanden sind, aus den Datensilos der Firmen entlassen würden. Diese und andere kollektive Datenbestände könnten als Trainingsdaten die Grundlage neuer Dienstleistungen sein, deren Entwicklung dann nicht allein in der Hand weniger globaler (amerikanischer) Megakonzerne liegen würde. Die meisten dieser Daten haben die Benutzer niemals willentlich an Internetfirmen übertragen, sie bilden ein kollektives Gut, das aber nicht als Gemeingut verfügbar ist. Die Firmen entscheiden, wie und ob die Gemeinschaft von der Nutzung dieser Daten profitiert. Sie pochen auf ihr Eigentum an „ihren" Daten. Vorausschauend versuchen die Internetunternehmen daher, ihre Nutzer dazu zu bewegen, sich ausdrücklich zu registrieren und eingeloggt zu bleiben. Damit werden Daten aus unklaren Besitzverhältnissen den Nutzungsbedingungen der Konzerne unterstellt.

Daten schaffen Werte, aber Daten schaffen auch Verwertungsmonopole. Dies ist ein reales Innovationshemmnis. Gleichzeitig sind Daten ein Einfallstor in unsere Privatsphäre. Zur Lösung beider Probleme sollten wir die Hoheit über unsere Daten zurückverlangen. Daten, die von mir oder durch meine Mithilfe erzeugt wurden, sollten dauerhaft unter meiner Kontrolle bleiben – sie sollten anderen immer nur geliehen sein. Dafür reicht es nicht, wenn sich Internetfirmen nach Gutdünken AGBs und „Privacy Dashboards" ausdenken und wir uns mühsam von Hand hindurchklicken. Im Sinne der Nutzer müssen stattdessen programmatische Schnittstellen definiert werden, auf die wir mit anderen Programmen effizient zugreifen können. Das, was in der Industrie zwischen Firmen Usus ist, muss auch die eigentlichen Urheber der Daten einschließen: uns alle.

Diese standardisierten Schnittstellen müssen alle Daten abdecken, die wir bei der jeweiligen Firma erzeugt haben, sei es durch Anfragen, Klicks, Page Views oder das Ausfüllen von Textformularen. Die Schnittstellen müssen es ermöglichen, Daten zu löschen, sie mit einem Verfallsdatum zu versehen oder sie anderen Diensten zur Verfügung zu stellen. Diese Normierung ist nicht im unmittelbaren Geschäftsinteresse der Platzhirsche, wäre aber in der technischen Umsetzung kaum anspruchsvoller als das Tagesgeschäft dieser Unternehmen. Der Mehrwert für unsere Gesellschaft wäre immens: Eine solche Umstellung der Gesetze und Programme eröffnet die Aussicht auf ein neues Ökosystem von Firmen und Produkten. Die Privatsphäre gewönne zurück. Die Wertschöpfung aus Daten würde transparent gemacht und damit der Regelung zum Nutzen der Allgemeinheit zugänglich.

Thomas Hofmann ist Professor für Datenanalytik und Informatik an der ETH Zürich. Er war fast acht Jahre bei Google als Entwicklungsleiter tätig.

Bernhard Schölkopf ist einer der Gründungsdirektoren des neuen Max-Planck-Instituts für Intelligente Systeme (Tübingen/Stuttgart) und arbeitete in der Forschungsabteilung von Microsoft.

Frankfurter Allgemeine Zeitung,
29.01.2015, Feuilleton, Seite 14