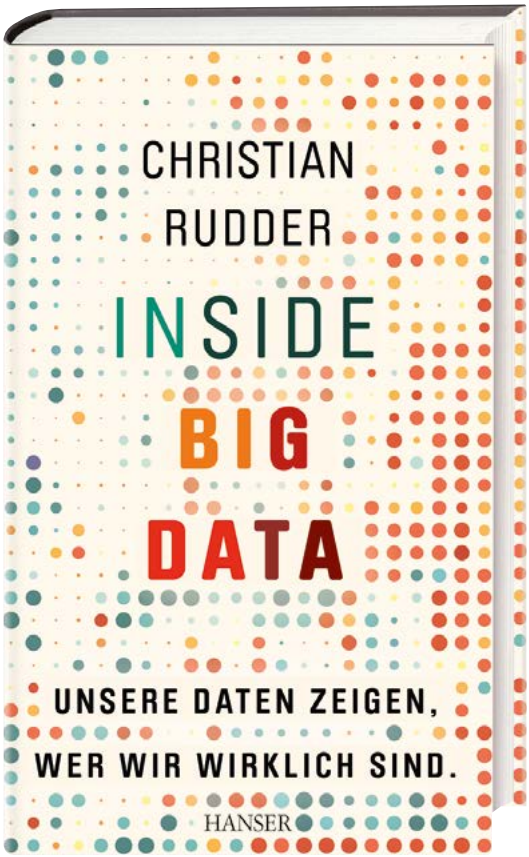
Leseprobe aus:

Christian Rudder
Inside Big Data



Mehr Informationen zum Buch finden Sie auf
www.hanser-literaturverlage.de

HANSER

Christian Rudder

Inside Big Data

Unsere Daten zeigen, wer wir wirklich sind

Aus dem Englischen von Kathleen Mallett

HANSER

Titel der Originalausgabe:
Dataclysm. Who We Are When We Think No One's Looking.
New York, Crown Publishers 2014

Tabelle auf Seite 178: C. J. Sorell: »Zipf's Law and Vocabulary«, in: C. A. Chapelle (Hg.): *The Encyclopedia of Applied Linguistics,* Oxford, Wiley-Blackwell 2012. Reprinted by permission of the author.

Tabelle auf Seite 246: Traits predicted by a Facebook user's »likes« adapted from Figure 2, »Prediction Accuracy of Classification of Dichotomous/Dichotomized Attributes Expressed by the AUC«, in: Michael Kosinskia, David Stillwell und Thore Graepel: »Private Traits and Attributes Are Predictable from Digital Records of Human Behavior«, in: *PNAS* 2013. Reprinted by permission of the Proceedings of the National Academy of Sciences of the United States of America.

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

1 2 3 4 5 20 19 18 17 16

www.hanser-literaturverlage.de
Bildbearbeitung: Christina Zeeb
Herstellung: Denise Jäkel
Umschlaggestaltung: Hauptmann & Kompanie Werbeagentur, Zürich
Satz: Kösel Media GmbH, Krugzell
Druck und Bindung: Friedrich Pustet, Regensburg
Printed in Germany
ISBN 978-3-446-44459-1
E-Book-ISBN 978-3-446-44460-7

Inside Big Data

Inhalt

Teil 2

Was uns auseinander-bringt

Teil 3

Was uns ausmacht

Einleitung

Sie haben bestimmt schon jede Menge über Big Data gehört: das enorme Potenzial, die bedrohlichen Folgen, das ach so paradigmenzerstörende *neue Paradigma*, das dieses Phänomen für die Menschheit und ihre geliebten Internetseiten bedeutet. Man wird ganz verwirrt dabei, wie von einem sehr stumpfen Gegenstand getroffen. Ich habe also nicht vor, mit noch mehr Hype oder atemloser Berichterstattung über das Datenphänomen aufzuwarten; vielmehr bringe ich Ihnen das, worum es geht: die Daten selbst, ganz ohne Phänomen. Ich möchte Ihnen zahlreiche Beispiele der tatsächlich gesammelten Informationen aus meinem eigenen Vorrat zeigen, den ich durch Glück, Arbeit, Herumprobieren und noch mehr Glück ansammeln konnte und der mir eine einmalige Ausgangslage für die Analyse verschafft.

Ich gehöre zu den Gründern der Webseite OkCupid, eines Internetpartnersuchdienstes, der im Laufe von zehn langen Jahren und ganz ohne Blase einer der weltweit größten geworden ist. Mit Freunden zusammen habe ich die Seite gestartet. Wir hatten alle einen Hang zur Mathematik, und der Erfolg unserer Seite gründet sich zum großen Teil darauf, dass wir diese Einstellung auf die Partnersuche anwandten; wir brachten Analytik und Methode in ein Feld, das traditionell die Domäne von Liebes»experten« und grinsenden Zauberern ist. Der Service funktioniert eigentlich ganz einfach – wie sich herausgestellt hat, braucht man an Mathematik, um den Vorgang einer Partnersuche zu modellieren, lediglich etwas nüchterne Arithmetik –, aber unser Ansatz stieß, warum auch immer, auf beträchtliche Resonanz, und alleine im laufenden Jahr werden zehn Millionen Nutzer mit unserer Seite auf Partnersuche gehen.[1]

Ich weiß nur zu gut, dass Webseiten (und ihre Gründer) ungeheuer gerne mit großen Zahlen um sich werfen, und die meisten denkenden Menschen ignorieren dieses Gehabe inzwischen wahrscheinlich; man hört von Millionen dies und Milliarden jenes und weiß schon, dass es eigentlich nur Eigenlob mit einer Menge Nullen im Schlepp ist. Anders als Google, Facebook, Twitter und die anderen Quellen, deren Daten wir im vorliegenden Buch bevorzugt verwenden, ist OkCupid nicht besonders bekannt – wenn Sie und Ihre Freunde alle seit Jahren glücklich verheiratet sind, haben Sie wahrscheinlich nie von uns gehört. Ich habe also lange darüber nachgedacht, wie ich die Reichweite unserer Seite jemandem erklären kann, der sie nie besucht hat und sich für die Nutzerstatistik irgendeines Start-ups auch völlig zu Recht nicht besonders interessiert. Ich möchte es mit einem persönlichen Ansatz versuchen. Heute Abend haben etwa 30 000 Paare ein erstes Date, das durch OkCupid vermittelt worden ist.[2] Etwa 3000 von ihnen werden eine dauerhafte Beziehung eingehen. 200 davon werden heiraten, und viele werden natürlich Kinder bekommen. Genau jetzt, am heutigen Tag, leben und schmollen und trotzen Kinder – mürrische kleine Menschenwesen, die sich in dieser Sekunde weigern, ihre Schuhe anzuziehen –, die es ohne unsere versponnene HTML nie gegeben hätte.

Ich wiege mich nicht selbstgerecht im Glauben, dass wir irgendetwas perfektioniert hätten, und ich möchte betonen, dass ich zwar natürlich stolz auf die Seite bin, die meine Freunde und ich geschaffen haben, dass es mir aber nicht darauf ankommt, ob Sie Mitglied sind oder ein Nutzerkonto einrichten wollen oder was auch immer. Ich habe in meinem ganzen Leben noch kein Online-Date gehabt, meine Mitgründer übrigens auch nicht, und deswegen kann ich es gut verstehen, wenn Sie für Internetpartnersuche nichts übrig haben. Ich bin ganz bestimmt kein Technikmissionar, und ich will mit diesem Buch keine glitzernden digitalen Glasperlen für irgendjemandes wertvolle Insel eintauschen. Ich bin immer noch Abonnent analoger, gedruckter Zeitschriften. Am Wochenende bekomme ich die *Times* ins Haus. Twittern ist mir peinlich. Ich kann Sie also kaum dazu überreden, das Internet oder die sozialen Medien stärker zu respektieren oder daran zu »glauben«, als Sie es bereits tun (oder auch nicht). Denken Sie nur ruhig weiter so über das

Internetuniversum, wie Sie es vor der Lektüre dieses Buches getan haben. Aber wenn es etwas gibt, worüber ich Sie ernsthaft zum Nachdenken anregen möchte, dann sind es Ihre Ansichten und Annahmen über sich selbst. Denn darum geht es in diesem Buch. OkCupid war für mich nur der Grund, mich selbst mit dieser Frage zu beschäftigen.

Seit 2009 leite ich das OkCupid-Analyseteam, und es ist meine Aufgabe, etwas Sinnvolles aus den Daten zu machen, die unsere Nutzer erzeugen. Meine drei Mitgründer haben fast die ganze harte Arbeit auf sich genommen, die erforderlich war, um den Partnersuchdienst aufzubauen, während ich jahrelang nur mit Zahlen herumgespielt habe. Einige meiner Ergebnisse unterstützen unser Unternehmen – so ist es für einen Datingservice natürlich entscheidend, zu verstehen, dass und wie Männer und Frauen Sex und Schönheit verschieden bewerten. Aber ein ganzer Haufen meiner Erkenntnisse war nicht direkt praktisch verwertbar, sondern bloß interessant. Mit der Tatsache, dass rein statistisch Belle & Sebastian die am wenigsten schwarze Band der Welt ist, oder dass einen das Blitzlicht auf einem Foto durchschnittlich sieben Jahre älter wirken lässt, kann man zum Beispiel nur sehr wenig anfangen, außer *Aha!* zu sagen und sie vielleicht auf einer Cocktailparty zum Besten zu geben. Und mehr haben wir mit diesen Daten lange auch nicht gemacht; unsere Erkenntnisse brachten nicht mehr als gelegentliche lahme Pressemitteilungen hervor. Aber irgendwann lag uns genug Information zur Analyse vor, um größere Trends erkennen zu können, großräumige Muster, die sich aus den kleinen ergaben. Besser noch, ich merkte, wie ich Tabuthemen wie die Rassenfrage durch direkte Erhebungen untersuchen konnte. Ich musste also keine Fragebögen verteilen oder Studien mit begrenzten Teilnehmergruppen durchführen – das ist die alte Methode der Sozialpsychologie –, sondern ich konnte hingehen und mir anschauen, *was wirklich passiert*, wenn beispielsweise 100 000 männliche Weiße in privaten Kontakt zu 100 000 weiblichen Schwarzen treten. Diese Daten haben wir auf unseren Servern einfach so zur Verfügung. Eine unwiderstehliche Möglichkeit soziologischer Untersuchungen tat sich auf.

Ich stürzte mich hinein, und als sich die Entdeckungen häuften, machte ich es wie jeder, der mehr Ideen als interessierte Zuhörer hat:

Ich startete einen Weblog, um sie mit der Welt zu teilen. Aus diesem Blog ist das vorliegende Buch hervorgegangen – mit einer wichtigen Verbesserung: Für *Dataclysm* bin ich weit über den Rahmen von OkCupid hinausgegangen. Ich habe dabei eine Datensammlung zwischenmenschlicher Interaktion zusammengetragen, die wahrscheinlich größer als die jedes anderen Privatmannes ist – sie umfasst die meisten, wenn nicht alle wichtigsten Internetdatenquellen unserer Zeit. Im Folgenden spreche ich also nicht nur von den Gewohnheiten der Nutzer einer bestimmten Seite, sondern auch über menschliche Universalien.

Die öffentliche Diskussion zur Frage der Datenerhebung konzentriert sich auf zwei Bereiche: Schnüffelei durch Regierungen und Ausbeutung durch kommerzielle Unternehmen. Über den ersten Bereich weiß ich nicht mehr als Sie auch – nur das, was ich gelesen habe. Meines Wissens nach haben sich die US-Geheimdienste noch nie an einen Internetpartnersuchdienst gewandt, um seine Daten abzugreifen, und wenn sie nicht vorhaben, die Zurschaustellung durchtrainierter Bauchmuskeln ohne Gesicht oder die endlosen und offensichtlich unwahren Behauptungen junger Frauen aus Brooklyn, wie sehr sie Scotch mögen, zu einem Verbrechen zu erklären, wüsste ich auch nicht, warum. Im zweiten Bereich, wo es um die Verwandlung von Daten in Dollars geht, kenne ich mich besser aus. Als ich dieses Buch begann, waren die Computerblätter gerade sabberfeucht vor Gier, weil die ersten Facebook-Aktien ausgegeben wurden; diese Firma hatte schließlich die persönlichen Daten von so gut wie jedem zu Geld gemacht und wollte jetzt dieses Geld an der Börse zu noch mehr Geld machen. Eine Artikelüberschrift in der *Times* drei Tage vor dem Börsenstart sagt alles: »Facebook spinnt Daten zu Gold«. Es fehlte nur noch, dass Rumpelstilzchen einen Kommentar mit einer Kaufempfehlung dazuschrieb.

Als Gründer einer werbefinanzierten Webseite kann ich bestätigen, dass Daten tatsächlich nützlich sind, wenn man etwas verkaufen will. Jede einzelne Abteilung einer Webseite kann das gesamte Verhalten des Nutzers protokollieren – was er anklickt, was er eintippt, sogar, wie lange er wo verweilt – und daraus ein klares Bild seiner Gelüste gewinnen, und wie man sie befriedigt. Diese Macht mag zwar beeindruckend sein, aber ich will jetzt nicht von der geheimen Mission unserer Nation

anfangen, Leuten Deospray zu verkaufen, die ihren Freunde Updates über Deospray schicken. Ich führe vielmehr dieselben Protokolle des Nutzerverhaltens – die Klicks, Tasteneingaben und Millisekunden – einem ganz anderen Zweck zu. Wenn die zwei Hauptanwendungen für Big Data in den letzten Jahren Überwachung und Profit gewesen sind, dann, so kann man sagen, arbeite ich seit drei Jahren an einer dritten: der menschlichen Dimension.

Facebook weiß, dass Sie zu den vielen M&M-Fans gehören und schickt Ihnen entsprechende Werbung. Facebook weiß auch, wenn Sie mit Ihrem Freund Schluss machen, nach Texas ziehen, dann plötzlich auf vielen Bildern mit Ihrem Ex auftauchen und sich schließlich wieder mit ihm verabreden. Google weiß, wenn Sie sich nach einem neuen Auto umsehen, und kann Ihnen Marke und Typ ganz nach der Zielgruppe vorschlagen, der Sie zugerechnet werden. Sie sind ein abenteuerlustiger, umweltbewusster Typ B, männlich, Alter 25 bis 34 Jahre? Bitteschön, hier ist Ihr Subaru. Außerdem weiß Google auch noch, ob Sie schwul oder wütend oder einsam oder rassistisch sind oder sich Sorgen machen, ob Ihre Mutter womöglich Krebs hat. Twitter, Reddit, Tumblr, Instagram – all diese Firmen sind natürlich in erster Linie profitorientierte Unternehmen, aber gleich danach erheben sie auch demografische Daten bisher unbekannter Reichweite, Vollständigkeit und Bedeutsamkeit. Praktisch nebenbei zeigen uns digitale Daten jetzt realistisch, wie wir Menschen kämpfen, wie wir lieben, wie wir altern, wer wir sind und wie wir uns verändern. Man muss nur hinschauen: Aus ganz kurzer Distanz zeigen uns die Daten, wie sich Leute verhalten, wenn sie glauben, unbeobachtet zu sein. Im Folgenden zeige ich Ihnen, was ich gesehen habe. Und Deosprays sind mir dabei wirklich egal.

∞

Wenn Sie Erfahrung mit allgemein verständlichen Sachbüchern haben, dann werden Sie an Dataclysm vielleicht dies und das ungewöhnlich finden. Erstens die Verwendung der Farbe Rot. Zweitens, dass es darin um Menschengruppen und große Zahlen geht, woraus sich die seltsame

Erscheinung ergibt, dass in einem Buch über Menschen kaum welche vorkommen: Einzelpersonen werden so gut wie nicht erwähnt. Es gibt viele Grafiken, Statistiken und Tabellen, aber kaum Namen. Es ist ein beliebtes Stilmittel der Populärwissenschaft, große Ereignisse anhand von etwas Kleinem und Kuriosem zu schildern – die Geschichte der Welt aus der Sicht der Speiserübe, ein Fisch als Kriegsgrund, ein ganzer Regenbogen an der Zimmerwand, wenn man eine Taschenlampe im genau richtigen Winkel auf ein Prisma richtet. Ich mache es genau umgekehrt. Ich nehme etwas Großes – eine ungeheure Datenmenge über Handlungen, Gedanken und Äußerungen von Menschen, viele Terabytes an Daten – und filtere viele kleine Fakten heraus: Was das Netzwerk Ihrer Freunde über die Stabilität Ihrer Ehe denkt, wie Asiaten (oder Weiße, Schwarze und Latinos) sich am seltensten selbst beschreiben, wo und warum Schwule Ihre Neigung nicht öffentlich leben, wie sich die Schreibgewohnheiten in den vergangenen zehn Jahren geändert haben, und wie sich Wutausbrüche nicht geändert haben. Das Prinzip ist, unser Selbstverständnis auf Zahlen statt auf Beschreibungen zu gründen, oder vielleicht eher die Zahlen als Beschreibungen zu verstehen.

Dieser Ansatz entstand in den Schlackehalden der Statistik. *Dataclysm* ist eine Erweiterung der Arbeit meiner Kollegen und meiner selbst aus den letzten Jahren. Eine Internetpartnervermittlung möchte Menschen zusammenbringen, und um das seriös zuwege zu bringen, muss sie die Wünsche, Gewohnheiten und Abneigungen ihrer Nutzer kennen. Also macht man sich fleißig daran, einen Haufen detaillierter Daten zu sammeln und sie in allgemeine Theorien menschlichen Verhaltens zu verwandeln. Wer mit solchen Informationen arbeitet anstatt, zum Beispiel, für die Hochzeitsredaktion der Sonntagszeitung, bekommt mit der Zeit ein besonderes Verwandtschaftsgefühl zum bunt gemischten Durcheinander der Menschheit als Ganzes und nicht zu zwei Einzelpersonen. Man fängt an, die Menschen so zu begreifen, wie ein Chemiker vielleicht die umherwirbelnden Moleküle seiner Tinktur begreift und zu lieben beginnt.

Dennoch müssen natürlich alle Webseiten – und alle Informatiker – ihre Daten formalisieren. Mit allem, was keine Zahl ist, kann ein Algo-

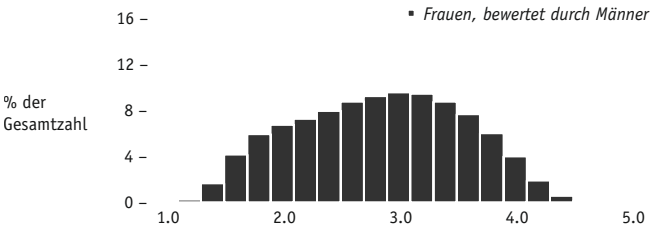
rithmus nichts anfangen; wenn ein Computer also etwas begreifen soll, muss man es ihm so weit wie möglich in Zahlen übersetzen. Webseiten und Apps stehen damit vor der Herausforderung, das Kontinuum menschlichen Verhaltens in kleine Schubladen mit den Etiketten 1, 2, 3 und so weiter zu füllen, ohne dass es auffällt; einen umfassenden, unteilbaren Vorgang – bei Facebook ist es Freundschaft, bei Reddit Gemeinschaft, bei Partnervermittlungen Liebe – in Stücke zu gliedern, mit denen ein Server umgehen kann. Gleichzeitig muss man aber so viel wie möglich das *Je ne sais quoi* dieses Vorgangs beibehalten, damit der Nutzer weiter glaubt, echtes Leben geboten zu bekommen. Es ist eine sehr empfindliche Illusion, dieses Internet; stellen Sie sich eine Möhre vor, die so sauber in Scheiben geschnitten ist, dass die Scheiben weiter zusammenkleben und aussehen wie eine ganze Möhre. Dieses Spannungsverhältnis – nämlich zwischen dem Kontinuum menschlicher Verhaltensweisen und der Fraktionierung der Datenbank – kann den Betrieb einer Webseite komplizieren, aber es steckt auch hinter meinem Ansatz. Die Näherungslösungen, wie sie uns die Technik für Sachen wie Lust und Freundschaft bietet, bieten eine ganz neuartige Gelegenheit, zeitlose Mysterien mit harten Zahlen festzunageln und Verhaltensweisen, die wir als »unquantifizierbar« abzutun gewohnt sind, neu zu sehen und besser zu verstehen. Die Näherungslösungen werden immer besser, weshalb die Menschen sie immer stärker als Teil ihres Lebens akzeptieren, und dadurch wiederum verbessert sich dieses Verständnis mit erstaunlicher Geschwindigkeit. Ich gebe Ihnen gleich ein kurzes Beispiel, aber zuerst möchte ich noch anmerken, dass wir damals eigentlich »Wir machen das Unteilbare teilbar« zum Motto von OkCupid hätten erklären sollen. Schade, zu spät.

Im Internet wird dauernd bewertet. Ob die »Up/down«-Votes bei Reddit, die Kundenrezensionen bei Amazon oder sogar der »Gefällt mir«-Button bei Facebook – ständig möchte irgendeine Webseite von Ihnen, dass Sie abstimmen, denn dadurch wird etwas Fließendes und Idiosynkratisches – Ihre Meinung – so formalisiert, dass auch ein Rechner sie versteht. Auf den Webseiten von Partnervermittlungen sollen die Nutzer einander gegenseitig bewerten, weil sich so erste Eindrücke wie

Er hat schöne Augen
Hmmm, sieht ganz gut aus, aber ich mag keine Rothaarigen
Iiih, wie eklig

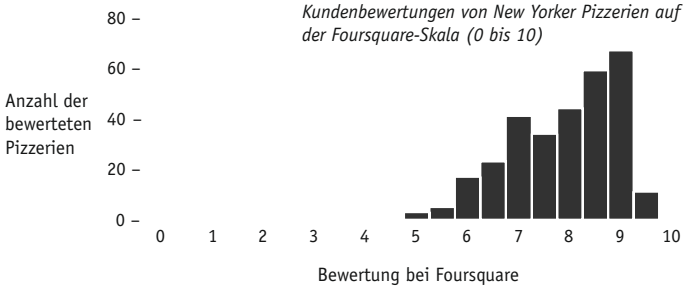
in einfache Zahlen verwandeln, zum Beispiel Fünf, Drei und Eins auf einer Fünf-Punkte-Skala. Webseiten sammeln Milliarden solcher Mikrobewertungen, Schnappschüsse von dem Eindruck, den ein Mensch von einem anderen hat. Zusammengenommen bilden diese winzigen Gedankensplitter eine enorme Erkenntnisquelle dafür, wie solche Meinungen entstehen.

Das Einfachste, was man mit Mensch-zu-Mensch-Bewertungen tun kann, ist, sie zusammenzuzählen. Man stellt zusammen, wie viele Menschen, einen Stern, zwei Sterne und so weiter erzielt haben, und vergleicht die Summen. Hier folgt eine solche Aufstellung für die durchschnittliche Bewertung heterosexueller Frauen durch heterosexuelle Männer. Die Kurve hat folgende Form:

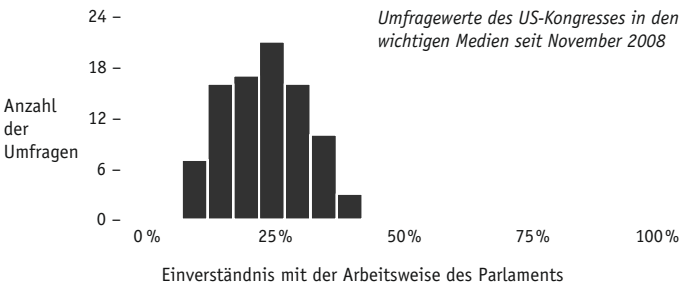


51 Millionen Bewertungen lassen sich zu dieser einfachen Reihe von Rechtecken kondensieren. Im Wesentlichen ist das die gesammelte männliche Ansicht zu weiblicher Schönheit auf OkCupid. Sie bildet aus all den winzigen Geschichten (was ein Mann von einer Frau hält, millionenfach wiederholt) und all den Anekdoten (von denen jede einzelne hier als Aufhänger dienen könnte, wenn es ein entsprechendes Buch wäre) ein verständliches Ganzes. Eine solche Sichtweise auf die Menschen ist ungefähr so, als wenn man sich die Erde aus dem Weltall anschaut; die Einzelheiten gehen verloren, aber man sieht Vertrautes auf eine völlig neue Weise.

Was also sagt uns diese Grafik? Die Grundform – eine sogenannte Glockenkurve – nimmt man leicht für normal, weil das in den Lehrbüchern so steht, aber die Bewertungen hätten auch durchaus stark nach der einen oder der anderen Seite ausschlagen können. Wenn es um persönliche Vorlieben geht, passiert das ziemlich oft. Nehmen wir die Bewertungen von Pizzabäckereien bei Foursquare, die extrem positiv ausfallen:[3]

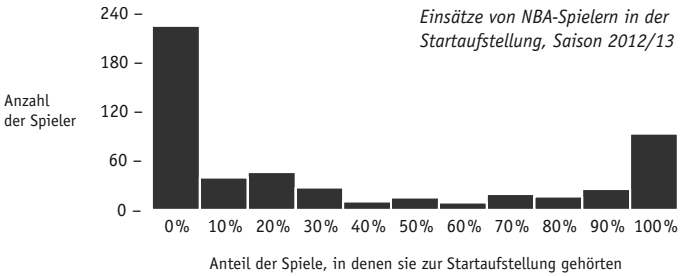


Oder nehmen wir eine Beliebtheitsstatistik für den US-Kongress, die, weil Politiker moralisch gesehen das Gegenteil von Pizza sind, in die andere Richtung tendiert:[4]

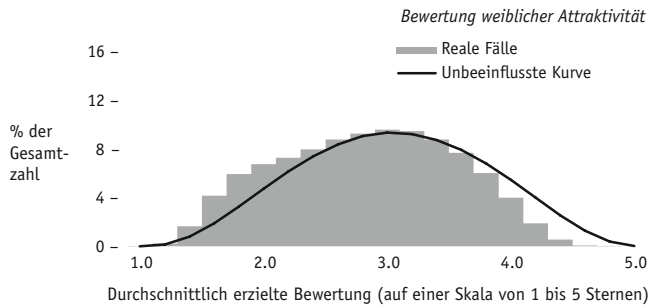


Außerdem ist die Bewertung von Frauen durch Männer in unserem Beispiel *unimodal*. Das bedeutet, dass die Werte der Frauen sich um eine einzige Spitze gruppieren. Auch das kann man leicht für selbstverständlich halten, aber in vielen Fällen zeigen Statistiken mehrere solche Modi oder »typische« Werte. Wenn man die Einsatzzahlen von Spielern der Basketballspitzenliga NBA in der Saison 2012/13 grafisch

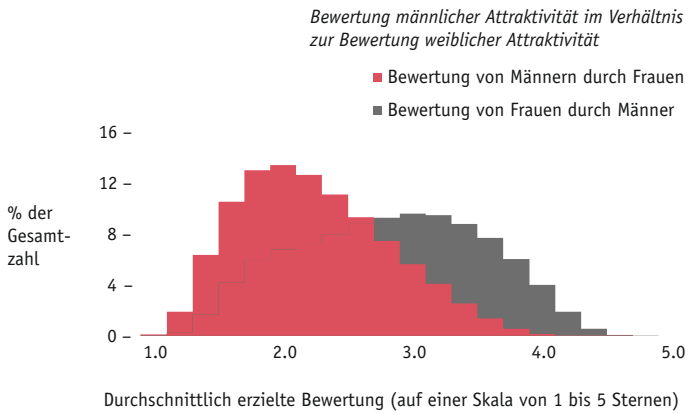
darstellt, drängen sie sich alle an den beiden Enden, während die Mitte fast leer bleibt:[5]



Hier sagen uns die Daten, dass der Trainer jeweils einen Spieler entweder für gut genug hält, von Anfang an zum Einsatz zu kommen, oder eben nicht. Also gehört der Betreffende zur Startaufstellung oder nicht, und zwar in jedem Spiel. Das ist ein deutlich binäres System. Ebenso könnten in unserem Bewertungsbeispiel die Frauen von den Männern auch einfach als entweder »schön« oder »hässlich« eingestuft worden sein; es wäre ja vorstellbar, dass körperliche Schönheit etwas ist, das man hat oder auch nicht. Aber die Kurve im ersten Beispiel sagt eben etwas anderes aus. Datenanalyse besteht sehr oft darin, dass man seine Resultate mit solchen Was-wäre-wenn-Hypothesen vergleicht. Manchmal ist ein eindeutiges Ergebnis eigentlich erstaunlich, wenn man unendlich viele Alternativen hätte. Unsere Kurve kommt sogar einer sogenannten *symmetrischen Betaverteilung* recht nahe, einer Kurve, die oft als Modell für grundlegende, unbeeinflusste Entscheidungen dient. Wir legen sie zur Verdeutlichung über die tatsächliche Bewertungskurve für Schönheit:



Unsere realen Daten weichen nur leicht (um sechs Prozent) von diesem abstrakten Ideal ab, was bedeutet, dass die Grafik männlichen Begehrens so ziemlich dem entspricht, was wir ohne alle Daten geraten hätten; sie ist damit eines dieser Lehrbuchbeispiele, die ich eben so leichthin abgetan habe.[6] Die Kurve ist also berechenbar, durchschnittlich – vielleicht sogar langweilig. Na und? Nun, manchmal ist auch Langweiligkeit etwas Besonderes, und in diesem speziellen Zusammenhang ist das der Fall: Sie impliziert, dass die Männer, von denen die Bewertungen stammen, als Einzelpersonen ebenfalls berechenbar, langweilig und vor allem unbeeinflusst sind. Wenn man jetzt an die Supermodels, die Nacktfotos, die Covergirls, die Videospielheldinnen à la Lara Croft, die aufreizenden Werbemädels und vor allem an die mit Photoshop manipulierten Bilder denkt, die all diese Männer täglich zu sehen bekommen, ist, so finde ich, die Tatsache, dass die männliche Bewertung weiblicher Schönheit immer noch völlig normal bleibt, ein kleines Wunder. Man sollte doch erwarten, dass Männer inzwischen absolut unrealistische Anforderungen an weibliche Schönheit stellen, aber diese einfache Grafik zeigt uns, dass es nicht so ist. Jedenfalls sind die Männer bei ihren Bewertungen viel nachsichtiger als die Frauen. Deren Bewertung fällt so aus:



Die rote Kurve erreicht ihren Höchstwert schon nach einem knappen Viertel der Skala; nur jeder sechste Mann gilt, absolut gesehen, als

»überdurchschnittlich« attraktiv. Sex-Appeal wird normalerweise nicht so quantifiziert, also möchte ich es gerne vertrauter ausdrücken: Wenn diese Kurve Intelligenz statt Attraktivität bewertete, dann lebten wir in einer Welt, in der Frauen 58 Prozent aller Männer für geistig behindert hielten.[7]

Nun sind die männlichen Nutzer von OkCupid nicht hässlicher als andere Männer – und zwar nachweislich. Ich habe zu diesem Zweck eine Zufallsauswahl unserer Kunden mit einer Zufallsauswahl aus einem sozialen Netzwerk vergleichen lassen und dieselben Werte für beide Gruppen von Männern erhalten – und entsprechende Kurven erhält man, wie sich herausgestellt hat, auch für die Daten jedes anderen Internetpartnersuchdienstes, den ich kenne: Tinder, Match.com, DateHookup, also für Webseiten, bei denen zusammengenommen fast die Hälfte aller US-Singles registriert ist.[8] Daraus müssen wir schlussfolgern, dass Männer und Frauen, wenn es um Sex geht, unterschiedlich denken. In *Harper's Bazaar* hieß es zutreffend einmal: »Frauen bedauern später immer den Sex, auf den sie sich eingelassen haben, Männer den Sex, den sie nicht gekriegt haben.«[9] In den Datenkurven sieht man genau, wie das funktioniert. Und natürlich, so möchte ich hinzufügen, müssen die Männer in diesen Beispielen wirklich voller Bedauern sein.

Eine Betaverteilung ist die grafische Darstellung einer Datenmenge, die sich als Ergebnisliste einer langen Reihe Münzwürfe begreifen lässt – sie bezeichnet die einander überschneidenden Wahrscheinlichkeitsverteilungen zahlreicher voneinander unabhängiger binärer (zweiwertiger) Ergebnisse.[10] Die männliche Münze ist dabei »fair«, denn ihr Wurf ergibt etwa genauso oft Kopf (was ich mit »positiv« gleichsetze) wie Zahl. Die weibliche Münze in unseren Datenbeispielen ist dagegen gezinkt: Sie fällt nur bei jedem vierten Wurf so, dass sie Kopf zeigt. Zahlreiche Naturvorgänge, zum Beispiel das Wetter, können mit einer Betakurve dargestellt werden, und dank des obsessiven Archivierungsdrangs eines bestimmten Wetternerds konnte ich tatsächlich unsere Mensch-zu-Mensch-Bewertung mit historischen Klimamustern vergleichen. Die männliche Bewertung kommt der Funktion sehr nahe, mit der sich die Bewölkung über New York darstellen lässt, während die weib-

liche Psyche in derselben Analogie eher einem regnerischen, bedeckten Ort wie Seattle entspricht.

Diesem Ansatz werden wir im ersten der drei übergreifenden Themenbereiche des vorliegenden Buchs folgen: den Daten von Menschen, die in Kontakt zueinander treten. Die körperliche Attraktivität, ihre Veränderung und ihre Entstehung, wird dabei unser Ausgangspunkt sein. Wir werden sehen, warum eine Frau, mathematisch gesehen, mit 21 eine alte Jungfer ist, und wie wichtig eine auffällige Tätowierung ist. Aber dann geht es schnell um viel mehr als nur fleischliche Verbindungen. Wir werden sehen, was uns Tweets über die moderne Kommunikation sagen, und was Freundschaften auf Facebook über die Stabilität einer Ehe verraten. Profilbilder sind sowohl ein Segen als auch ein Fluch des Internets: Fast jedes Internetangebot (Facebook, Jobvermittlungen und, natürlich, die Partnersuchdienste) wird durch sie zu einer Schönheitskonkurrenz. Wir sehen uns daher einmal an, was passiert, wenn OkCupid sie einen Tag lang ausblendet und auf das Beste hofft. Die Liebe ist nicht blind, obwohl wir sehen werden, dass das vielleicht besser wäre.

In Teil zwei sehen wir uns dann die Teilungslinien der Menschheit an. Beginnen werden wir mit jener allzu unübersehbaren in verschiedene Hautfarben. Dieses Thema können wir jetzt zum ersten Mal auf der Ebene der Einzelperson angehen. Unsere einmaligen Daten enthüllen Einstellungen, zu denen sich die meisten Menschen niemals öffentlich bekennen würden, und wir werden sehen, dass Rassenvorurteile nicht nur ausgeprägt, sondern auch hartnäckig sind. Sie wiederholen sich fast *verbatim* (gut, eigentlich *numeratim*) auf jeder untersuchten Webseite. Rassismus kann auch eine Sache des Einzelnen sein – ein Mann, sein Vorurteil und seine Tastatur. Wir werden sehen, was Google Search über das meistgehasste Wort in den USA zu sagen hat – und was dieses Wort über die USA aussagt. Dann befassen wir uns mit den Unterteilungen in Typen körperlicher Schönheit, und zwar anhand einer Datenbank, die viele Tausend Mal mehr aussagt als alles bisher Dagewesene. Hässlich zu sein hat einen hohen sozialen Preis, der jetzt endlich quantifizierbar wird. Danach schauen wir uns an, was Twitter über unsere Neigung zu Wutausbrüchen sagt. Mit diesem sozialen Dienst kann

man minütlich in Kontakt bleiben, aber er kann einen auch genauso schnell auseinandertreiben. Die sich gegenseitig aufschaukelnde Wut, die durch ihn ermöglicht wird, verleiht jener Urform menschlicher Versammlungen, dem Mob, eine neue Qualität der Gewalt. Wir werden sehen, ob Twitter auch eine neue Qualität von Verständnis bewirken kann.

Im dritten Teil werden wir uns dann, nachdem wir die Wechselwirkung zwischen zwei Menschen, im Guten oder im Schlechten, betrachtet haben, mit der Einzelperson befassen. Wir erforschen, wie sich die ethnische, sexuelle und politische Identität ausdrückt, wobei wir uns auf die Wörter, Bilder und kulturellen Marker konzentrieren, mit denen Menschen sich selbst darstellen. Hier sind fünf typische Wortfolgen aus dem Sprachgebrauch weiblicher Weißer:

my blue eyes
red hair and
four wheeling
country girl
love to be outside

(meine blauen Augen
rote Haare und
Vierradantrieb
Mädchen vom Lande
bin gerne draußen)

Ist das nun ein Haiku der Country-Sängerin Carrie Underwood oder eine Datenmenge? Die Entscheidung liegt bei Ihnen! Wir schauen uns den öffentlichen Sprachgebrauch der Menschen an, aber auch, wie sie sich im Privatleben ausdrücken und verhalten, mit besonderem Augenmerk auf Fälle, in denen sich Bezeichnungen und Handlungen unterscheiden: Männliche Bisexuelle zum Beispiel stellen unsere Ideen einer klar definierten Identität infrage. Dann nehmen wir uns einen weiten Quellenbereich vor – Twitter, Facebook, Reddit, sogar Craigslist –, mit dessen Hilfe wir uns selbst zu Hause über die Schulter schauen können,

sowohl physisch wie auch in anderer Hinsicht. Schließen werden wir dann mit der Frage, die sich ganz natürlich aus einem Buch wie dem vorliegenden ergibt: Wie kann man in einer Welt, in der solche Überwachung möglich ist, seine Privatsphäre wahren?

Das Internet ist, wie sich dabei herausstellen wird, ein lebendiger, brutaler, liebevoller, nachsichtiger, trügerischer, sinnlicher und hasserfüllter Ort. Ganz klar, es besteht ja auch aus Menschen. Als ich diese Informationen in ihrer Fülle zusammengetragen habe, wurde mir allerdings auch klar, dass es eben nicht aus *allen* Menschen besteht: Wenn Sie weder Rechner noch Smartphone benutzen, dann kommen Sie in diesem Buch nicht vor. Dieses Problem kann ich nur nennen und umgehen, aber nicht lösen.

Immerhin aber ist die Reichweite von Anbietern wie Twitter und Facebook bereits enorm, und sogar die eingeschränkteren Daten meines eigenen Partnersuchdienstes sind erstaunlich umfassend. Wenn Sie kein Freund sozialer Netzwerke und ähnlicher Dienste sind, dann ist Ihnen das vielleicht gar nicht klar. 87 Prozent der US-Bürger sind im Netz aktiv,[11] und zwar quer über alle demografischen Grenzen hinweg.[12] Stadt- wie Landbewohner, Reiche und Arme, Schwarze, Asiaten, Weiße und Latinos – alle sind miteinander verbunden. Bei den sehr Alten und in den bildungsfernen Schichten ist die Internetakzeptanz geringer (um die 60 Prozent), und deshalb habe ich meine »Altersgrenze« schon weit vor diese Gruppe gelegt – bei 50 Jahren – und ignoriere den Bildungsstand der Nutzer ganz einfach vollkommen. Über ein Drittel aller Amerikaner geht *täglich* auf Facebook.[13] Dieses soziale Netzwerk hat weltweit 1,3 Milliarden Nutzer. Wenn man davon ausgeht, dass etwa ein Viertel der Weltbevölkerung unter 14 Jahre alt ist,* bedeutet das, dass etwa ein Viertel aller Menschen über diesem Alter einen Account bei Facebook haben. Die Partnersuchwebseiten, die wir hier betrachten, haben in den letzten drei Jahren ungefähr 55 Millionen US-Bürger registriert – wie gesagt, das ist ein Account für jeden zweiten Single in diesem Land. Demografisch ein besonders interessanter Fall ist Twitter. Dieses Unternehmen ist eine vor Geld triefende Hightech-Erfolgs-

* 14 Jahre ist das Mindestalter für einen Facebook-Account

geschichte und gentrifiziert gerade im Alleingang einen großen Teil San Franciscos. Der Dienst selbst, den die Firma anbietet, ist allerdings absolut populistisch, sowohl was die Zugänglichkeit der Plattform als auch was die Universalität der Nutzer angeht. Es gibt zum Beispiel keinen signifikanten Geschlechtsunterschied bei der Intensität der Nutzung. Bloße Highschool-Abgänger tweeten genauso oft wie Doktoranden, Latinos ebenso viel wie Weiße und Schwarze doppelt so viel.[14] Und dann gibt es natürlich noch Google. Die 87 Prozent der US-Bürger im Internet entsprechen 87 Prozent Google-Nutzern unter den US-Bürgern.

Diese enormen Zahlen belegen nicht, dass ich irgendwie ein vollständiges Bild von irgendetwas bieten könnte, aber zumindest weisen sie auf ein solches Bild voraus. Und natürlich sollte das Vollkommene auch nicht der Feind des Besseren sein. Die Datenmengen, mit denen wir arbeiten werden, umfassen mehrere Tausend Mal so viele Menschen wie eine klassische Meinungsumfrage; das ist offensichtlich. Weniger offensichtlich ist, dass sie auch viel aussagekräftiger sind als die meisten sozialpsychologischen Studien.

In der Sozialpsychologie gibt es nämlich ein schon lange bekanntes, aber nur selten öffentlich erwähntes Problem: Fast alle ihre grundlegenden Thesen gründen sich auf Untersuchungen an kleinen Gruppen von Collegestudierenden. Als ich noch Student war, verdiente ich mir einmal 25 Dollar oder so dazu, indem ich im Massachusetts General Hospital eine Stunde lang ein leicht radioaktives Markierungsgas einatmete und dann Denksportaufgaben löste, während mein Gehirn gescannt wurde. Es tut nicht weh, versicherte man mir. Die Strahlung entspricht etwa einem Jahr auf einem Linienflug. Keine große Sache, sagten die Forscher. Was sie nicht sagten – und was mir damals nicht klar war –, war, dass ich, wie ich da so mit einem leichten Kater in einer Art Computertomograf lag, Wörter ablas und einen Fußschalter betätigte, als Stellvertreter des männlichen Durchschnittsmenschen diente. Auch ein Kommilitone, den ich kannte, nahm an der Studie teil, und er war ein weißes College-Kid genau wie ich. Ich wette, die meisten Teilnehmer waren Studenten, und damit natürlich alles andere als repräsentativ für den Mann als solchen.