Die schiere Menge sagt noch nichts
Big Data in den Sozialwissenschaften

Alexandros Tokhi und Christian Rauh

Summary: What does big data mean for our understanding of social and political phenomena? Many observers believe that big data spells the end of social science methods and theories. In an age where information is available in an unprecedented scope and detail, the data effectively seems to speak for itself. This stance heralds big data as a kind of universal fix-all, ignoring, however, the relevance and insights of the social science approach. In effect, big data leads to the emergence of new methods for the automatic collection and processing of information. Yet it is only in combination with explicit theories and solid research designs that this new technology can help to find better answers for socially relevant questions.

Kurz gefasst: Was bedeutet Big Data für das Verständnis sozialer und politischer Phänomene? Viele Beobachter glauben, Big Data mache sozialwissenschaftliche Methoden und Theorien überflüssig. Im Zeitalter von Informationen in nie gekanntem Umfang und Detail sprächen Daten quasi für sich selbst. Diese Position überhöht Big Data zum Allheilmittel, ignoriert aber die Relevanz sozialwissenschaftlicher Erkenntnisansätze. Der Big-Data-Hype schafft tatsächlich neue Verfahren zur automatischen Gewinnung und Verarbeitung von Informationen. Aber nur in Kombination mit expliziten Theorien und soliden Forschungsdesigns können diese neuen Technologien dabei helfen, bessere Antworten auf gesellschaftlich relevante Fragen zu finden.

Der Begriff Big Data ist ein allgegenwärtiges Schlagwort geworden. Soziale Netzwerke, Smartphones sowie diverse Online-Anwendungen und Websites produzieren und veröffentlichen ständig Informationen in einem nie dagewesenen Umfang und einer ungekannten Detailgenauigkeit. Zusammen mit günstigem Speicherplatz und immer besseren Analysemethoden scheint diese Fülle an Information unser Verständnis sozialer Umwelten zu revolutionieren. Deshalb vermuten viele, dass sich dadurch auch der wissenschaftliche Erkenntnisprozess fundamental verändert. Ein Mehr an Daten allein, so eine häufig gehörte Behauptung, führe zu besseren Entscheidungen und zur Lösung gesellschaftlicher Herausforderungen. Ist die Big-Data-Revolution also das Ende der Sozialwissenschaften, wie wir sie kennen?

Wir behaupten: Das Gegenteil ist der Fall. Es gehört zu den Grundpfeilern der sozialwissenschaftlichen Ausbildung, dass Beobachtungen allein – egal, wie viele es sind – kaum zu verlässlichen Schlussfolgerungen führen, wenn die zugrunde liegenden Annahmen nicht sorgfältig spezifiziert und in fundierte Forschungsdesigns übersetzt werden. Erst wenn moderne Datenerhebungs- und –analyseverfahren mit diesen Grundprinzipien sozialwissenschaftlicher Forschung kombiniert werden, lassen sich aus Big Data neue Einsichten in gesellschaftlich relevante Fragen erwarten.

Die teilweise kühnen Aussagen über die Big-Data-Revolution fußen oft auf naiven Ansichten über Datenanalyse und auf unklaren Abgrenzungen, was genau Big Data eigentlich ist. So wird insbesondere von einer ganzen Reihe von Computerwissenschaftlern, Internetexperten und Datenjournalisten propagiert, dass Big Data das Ende der Theorie einläute. Diese Argumente setzen Big Data mit einem *„N=all"*-Ansatz gleich, in dem einfache Korrelationen auch die letzte Wahrheit über die Zusammenhänge der sozialen Welt offenbaren.

Für ausgebildete Sozialwissenschaftler ist aber klar: Daten über soziale Phänomene erzählen selbst sehr wenig, wenn nicht der Betrachter zusätzliche Annahmen ins Spiel bringt. Ob gewollt oder nicht, die Theorien des Betrachters bestimmen, was wir in einem gegebenen Datensatz beobachten, welche Aspekte oder Phänomene wir als relevant identifizieren und wie wir Scheinkorrelationen von kausal bedeutsamen Zusammenhängen unterscheiden. Man muss nicht tief graben, um die Relevanz von theoretischen Annahmen auch im Big-Data-Kontext zu unterstreichen. Die Suchroutinen von Google bauen beispielsweise auf der Annahme auf, dass mehr Backlinks, also Rückverweise von einer Website auf eine andere, Indikatoren für eine größere Bedeutung dieser Website sind. Dieser Annahme mag man zustimmen oder nicht, man sollte sich ihrer aber bewusst sein, wenn man die Resultate interpretiert.

Ein weiteres Missverständnis ist es, die Verheißungen der Big-Data-Revolution mit der schieren Größe der Datensätze gleichzusetzen. Sozialwissenschaftler sind darauf trainiert, die Zahl und die Zusammensetzung empirischer Beobachtungen nur im Verhältnis zum tatsächlichen Auftreten des untersuchten Phänomens zu beurteilen. Der Big-Data-Hype weckt Assoziationen über Millionen von Beobachtungen, die soziale Netzwerke heute produzieren. Doch Individuen entscheiden selbst, ob und wie sie sich in diesen Netzwerken beteiligen – oder nutzen etwa alle Ihre Freunde Twitter? Wenn wir solche Selektionsprozesse ignorieren, wird auch ein noch so großer Datensatz zu verzerrten Schlussfolgerungen über gesellschaftliche Zusammenhänge führen.

In anderen, gesellschaftlich gleichermaßen relevanten Bereichen – etwa der weiter unten diskutierten Ratifikation internationaler Verträge – kann im Gegensatz dazu schon eine vergleichsweise kleine Zahl von Beobachtungen die Gesamtheit aller tatsächlich auftretenden Fälle gut repräsentieren. Es folgt: Ohne das soziale Phänomen genau zu bestimmen, gibt es von einem sozialwissenschaftlichen Standpunkt aus schlicht und einfach kein absolutes Kriterium, um Big Data und „small data" voneinander abzugrenzen. Das heißt aber auch, dass „große" und „kleine" Datenmengen den gleichen analytischen Herausforderungen gegenüberstehen, wenn es um die Repräsentativität und die Validität der aus den Daten gewonnenen Interpretationen geht. Wie sich aus Beobachtungsdaten gültige Schlussfolgerungen ziehen lassen, ist eine Kernfrage des sozialwissenschaftlichen Curriculums, und unsere Profession hat dazu ein umfassendes Instrumentarium entwickelt. Mit der schieren Menge der digital zur Verfügung stehenden Informationen nimmt die Bedeutung dieses Instrumentariums tatsächlich eher zu als ab. Aus unserer Sicht bedeutet Big Data daher nicht das Ende sozialwissenschaftlicher Grundprinzipien – das Gegenteil ist der Fall.

Wir sollten die Big-Data-Revolution also begleiten, können gleichzeitig aber auch enorm von ihr profitieren: Was wir in der Tat als „revolutionär" ansehen, ist die rasant wachsende Palette an Verfahren, um digitale Informationen automatisiert zu sammeln, zu verarbeiten und auszuwerten. In den verschiedenen sozialwissenschaftlichen Forschungsfeldern können uns diese Verfahren helfen, nur schwach strukturierte Datenquellen wie etwa Websites oder große Textkorpora systematisch in den Blick zu nehmen. Big Data stellt zeitsparende und oft auch frei zugängliche Mittel bereit, um bisher unsystematisierte empirische Quellen anzuzapfen. Kombiniert mit explizit spezifizierten Theorien und entsprechend soliden Forschungsdesigns können diese neuen Verfahren in der Tat zu einer besseren Beantwortung gesellschaftlich relevanter Fragen führen.

Denkt man an Big Data, sind Analysen zur inter- und supranationalen Politik wahrscheinlich nicht das erste Forschungsfeld, das einem in den Sinn kommt. Doch gerade deshalb sind zwei Beispiele aus unserer eigenen Forschung gut geeignet, um das sozialwissenschaftliche Potenzial entsprechender Datengewinnungs- und analyseverfahren zu veranschaulichen. Bei dem ersten Beispiel geht es um eine zentrale Frage der internationalen Beziehungen: Führen strenge rechtliche Verpflichtungen in internationalen Verträgen zu einer schnelleren Ratifikation durch einzelne Staaten, oder wirken sie eher abschreckend? Bisherige Studien zeigen widersprüchliche Resultate, vor allem weil sie nur kleine Stichproben internationaler Verträge untersucht haben.

Die übliche Erwartung ist es, dass Staaten strenge Verträge möglichst lange ignorieren, um sich ihre Handlungsfreiheit zu bewahren. Wir argumentieren jedoch, dass diese zentrale Entscheidung von der Natur des jeweiligen Politikfelds eines Vertrages abhängt. Wenn strengere Vertragsregeln zu grenzübergreifenden Sachverhalten vor allem andere Staaten zu einer bestimmten Vorgehensweise (zum Beispiel der Reduzierung von Luftverschmutzung) anhalten, sollten einzelne Staaten eher bereit sein, Verträge mit höherer Bindungskraft zu ratifizieren. Wenn strenge Vertragsklauseln aber hauptsächlich die eigene Handlungsfreiheit beschränken, ohne für andere direkt bindend zu sein, sollten sie einzelne Staaten eher von der Ratifizierung abschrecken. Diese Logik ist im Bereich der Menschenrechtsnormen verbreitet, während die erste Annahme auf Bereiche zutrifft, in denen öffentliche Güter nur durch zwischenstaatliche Kooperation bereitgestellt werden können (wie zum Beispiel saubere Luft).

Aus der Forschungsfrage und den theoretischen Erwartungen folgen klare Erfordernisse an die Daten: Wir müssen zunächst die Bereitschaft einzelner Staaten zur Ratifizierung von Verträgen erfassen – hier gemessen als die Dauer bis zur Ratifikation. Gleichzeitig müssen wir die Themenfelder und die dazugehörigen Verträge voneinander abgrenzen und die gesamte Variation in der Verbindlichkeit vertraglicher Verpflichtungen abbilden. Wir sind damit schnell bei mehreren Tausend Beobachtungen, wenn wir zum Beispiel jeden der etwa 80 Verträge über Menschenrechte und Umweltschutz der vergangenen 50 Jahre für alle 193 UN-Mitgliedsstaaten der Welt betrachten.

Alexandros Tokhi ist wissenschaftlicher Mitarbeiter der Abteilung Global Governance. In seiner Forschung beschäftigt er sich mit internationalen Institutionen, Nichtverbreitungspolitik und autoritären Regimen. *[Foto: Udo Borchert]*

alexandros.tokhi@wzb.eu

Christian Rauh ist wissenschaftlicher Mitarbeiter der Abteilung Global Governance. Er forscht über die gesellschaftliche Politisierung europäischer und internationaler Politik. *[Foto: David Ausserhofer]*

christian.rauh@wzb.eu

Hier sind Big-Data-Methoden klar von Vorteil. Wir programmieren und implementieren daher einen Web-Scraping-Algorithmus in der Programmiersprache Python, um automatisch Daten von der United Nations Treaty Collection Database zu extrahieren und neu zu organisieren (Web-Scraping bedeutet das maschinengesteuerte Auslesen von Websiten). Innerhalb von 2,5 Minuten können wir so ca. 140.000 einzelne Beobachtungen sammeln. Die Geschwindigkeit ist jedoch nicht der einzige Vorteil. Um strenge Verpflichtungen herauszufiltern, nutzen wir die Tatsache aus, dass verschiedene Typen von Verträgen, die den gleichen Rechtsbereich regulieren, sich nur in ihrem Grad der Verbindlichkeit unterscheiden. Rahmenkonventionen bürden den Staaten weniger Verpflichtungen auf als ihre Fakultativprotokolle. Der Python-Algorithmus identifiziert den Vertragstyp und kodiert so unseren Verbindlichkeits-Indikator. Die statistische Analyse dieser Daten stützt unser Argument und schließt damit eine wichtige Lücke in der Forschung zu internationalen Institutionen.

Beim zweiten Beispiel aus unserer Forschung geht es darum, wie intensiv nationale Parlamente EU-Themen debattieren. Wenn Parlamente öffentlich sichtbare Auseinandersetzungen über die EU und ihre Politik führen, könnten sie demokratische Defizite supranationaler Entscheidungsfindung zumindest abschwächen. Während einige Beobachter argumentieren, dass die Anreize, EU-Themen zu politisieren, tatsächlich mit jeder Übertragung politischer Kompetenzen auf die supranationale Ebene gestiegen sind, behaupten andere, dass öffentliche Diskussionen über EU-Angelegenheiten eher von selektiven parteipolitischen Motiven getrieben sind. Die bisherige empirische Forschung hat sich jedoch nur auf ausgewählte Parlamentsdebatten beschränkt, in denen EU-Themen bereits auf der formellen Tagesordnung standen. So sind konsistente Vergleiche über die Zeit nur eingeschränkt möglich. Es wird ignoriert, dass die EU heute fast die gesamte Bandbreite der nationalen, im Parlament diskutierten Debatten beeinflussen kann.

Um das abzubilden sind systematische Informationen über den parlamentarischen Stellenwert von EU-Themen nötig: Sie müssen über einen langen Zeitraum hinweg unterschiedliche Problemfelder einbeziehen sowie verschiedene Ebenen der EU-Behörden und die unterschiedlichen Konstellationen des nationalen parteipolitischen Wettbewerbs erfassen. Deshalb haben wir auf dem Dokumentenserver des Deutschen Bundestags automatisch alle Plenarprotokolle aus dem Zeitraum von 1991 bis 2013 ausgelesen. Mithilfe regulärer Ausdrücke teilen wir diese Texte in mehr als 148.000 individuelle Reden der Bundestagsabgeordneten auf und ordnen sie einzelnen Parteien zu. Schließlich nutzen wir flexible Wörterbücher und einen in der R-Umgebung implementierten Text-Mining-Algorithmus, um in jeder Rede alle wörtlichen Bezugnahmen auf das politische System, die Entscheidungsfindung und einzelne Politikbereiche der EU auszuzählen.

Die so gewonnenen Daten zeigen, dass die EU tatsächlich ein relevanterer Bezugsrahmen in allen Debatten des Deutschen Bundestags geworden ist. Die konkreten Muster lassen sich am besten über die mit jeder Vertragsrevision zunehmenden Kompetenzen auf EU-Ebene erklären. Dieses Ergebnis bleibt auch dann robust, wenn wir parteipolitische Differenzen und andere Kontrollvariablen berücksichtigen. Ob dies auch für andere Parlamente zutrifft, bleibt offen; gegenwärtig weiten wir diese Datengewinnungsstrategie deshalb auch auf Parlamente anderer EU-Mitgliedstaaten aus.

Big Data ist, was wir daraus machen

Big Data allein ist kaum ein Allheilmittel für alle Herausforderungen, denen sich moderne Gesellschaften gegenübersehen. Wenn entsprechende Analysen nicht durch klar spezifizierte Theorien gestützt werden, die Daten kontextualisieren und ihnen Bedeutung verleihen, starren wir bestenfalls auf riesige Zahlenberge. Schlimmstenfalls leiten wir politische Empfehlungen aus zweifelhaften Korrelationen und verzerrten Stichproben ab. Genau deshalb sollten sich die Sozialwissenschaften in der gegenwärtigen Debatte über Big Data Gehör verschaffen. Big Data ist nicht das Ende der Theorie. Vielmehr brauchen wir gerade jetzt das

Instrumentarium der Sozialwissenschaften, um die enorme Flut digitaler Informationen kritisch zu reflektieren und ihr Sinn zu geben.

Dazu müssen sich die Sozialwissenschaften selbst den Big-Data-Technologien öffnen. Einerseits müssen wir die Funktionsweise von Algorithmen ausreichend verstehen, um ihre Effekte und Aussagekraft bewerten zu können. Andererseits können wir Technologien wie Web Scraping, Pattern Recognition, Machine Learning oder Text Mining in unser methodisches Handwerkszeug aufnehmen und so Zeit und Kosten sparen, um die Fragen zu beantworten, die wir für relevant halten. Big Data wird die Sozialwissenschaften nicht transformieren, aber wir können sowohl zu den sich entwickelnden Technologien beitragen, als auch von ihnen profitieren.

Literatur

Anderson, Chris: „The End of Theory: The Data Deluge Makes the Scientific Method Obsolete". In: Wired Magazine, 06.23.2008.

Cukier, Kenneth Neil/Mayer–Schoenberger, Viktor: „The Rise of Big Data". In: Foreign Affairs, 2013, May/June. Online: http://www.foreignaffairs.com/articles/2013–04–03/rise–big–data (Stand 16.11.2015).

Dai, Xinyuan/Tokhi, Alexandros: „Depth, Participation, and International Human Rights Law". Paper presented at the American Political Science Association. San Francisco: September 3–6, 2015.

Rauh, Christian: „Communicating Supranational Governance? The Salience of EU Affairs in the German Bundestag, 1991–2013". In: European Union Politics, 2015, Vol. 16, No. 1, pp. 116–138.

Simon, Munzert/Rubba, Christian/ Meißner, Peter/Nyhuis, Dominic: Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. Chichester: Wiley 2015.