Bedeutet 'Big Data' das Ende der sozialwissenschaftlichen Methodenforschung?

von Jochen Mayerl

Folgt man den Angaben von Google Trends zur Häufigkeit von Suchbegriffen in Google, so erlebt das Thema *Big Data* seit knapp fünf Jahren einen kometenhaften Aufstieg: Wurde der Begriff weltweit vor 2011 so gut wie nie als Suchbegriff in Google eingegeben, ist die Anzahl der Suchanfragen seit 2012 sehr stark angestiegen. Das gleiche Phänomen lässt sich auf der Ebene wissenschaftlicher *Peer-Review*-Artikel beobachten: Laut *Science Citation Index* lag die Anzahl an publizierten Artikeln zum Thema bis einschließlich 2011 unter 80 Veröffentlichungen pro Jahr. Seit 2012 hat die Anzahl an wissenschaftlichen Publikationen hingegen deutlich zugenommen (2012: 609 Artikel; 2013: 2528; 2014: 5212).[1] Interessanterweise stammen fast alle Publikationen aus den drei Bereichen '*Research Technology*', '*Social Science*' und '*Arts Humanities*'. *Big Data* besitzt demnach sowohl einen stark technologischen Innovationscharakter als auch eine hohe gesellschaftliche und sozial-ethische Relevanz. Darüber hinaus führt *Big Data*, wie ich nachfolgend verdeutlichen möchte, zu möglichen Neuausrichtungen in der Methodologie der Sozialwissenschaften.

Zunächst stellt sich jedoch die Frage nach einer Begriffsbestimmung von *Big Data*. Der Begriff verweist zum einen auf das immense *Ausmaß* an anfallenden sozio-technischen Daten, die Fragen der effizienten informationstechnischen Migration, Vernetzung, Speicherung und Verarbeitung aufwerfen. Daneben verweist *Big Data* aber auch auf eine *neue Qualität* von Daten mit einer deutlich höheren Auflösung als bislang, sowohl hinsichtlich der möglichen Analyseebene von individuellen bis hin zu kollektiven gesellschaftlichen Daten, als auch hinsichtlich ihrer räumlichen und zeitlichen Reichweite.[2] Typischerweise werden die Charakteristika von *Big Data* mit drei „V"s umschrieben: *Volume* (Datenmenge), *Velocity* (Geschwindigkeit der Datenverarbeitung bis hin zu Echtzeitverarbeitung) und *Variety* (Anzahl an unterschiedlichen Daten).

Zu solchen sozio-technischen Daten, die alle prinzipiell untereinander gekoppelt werden können, gehören sämtliche web-basierten Daten (*Social Media* etc.), prozessproduzierte Daten und administrative Daten. Demnach zählen zu *Big Data* alle digital gespeicherten Verhaltensspuren, sämtliche denkbare Zusatzinformationen aus Organisationen, politischen Institutionen, Vereinen, Nutzungsdaten des Mobilfunks, Verkehrsdaten, Konsum- und

Kreditkarteninformationen, Daten aller möglichen elektronischen Geräte, die mit ihrer Umgebung kommunizieren (zum Beispiel durch Übermittlung von GPS-Daten) sowie alle ‚smarten' Technologien (*Smart Home* etc.), mit deren Hilfe *Big Data* zu *Smart Data* avanciert.[3] Es sind dann nicht primär die Datenmengen, die diese Technologie so faszinierend machen, sondern die Art der Daten, deren Kopplung und Auswertung eine Vielzahl bislang ungeahnter Verwendungsmöglichkeiten eröffnet. Die Relevanz von *Big Data* erschöpft sich mithin nicht in der bloßen Sammlung von Daten immensen Ausmaßes, sondern impliziert auch und gerade die Möglichkeiten der Untersuchung von Zusammenhängen zwischen all diesen Informationseinheiten.[4]

Angesichts der genannten Entwicklungen ist es gewiss keine Übertreibung, wenn man behauptet, dass *Big Data* für die sozialwissenschaftlichen Methoden der Erhebung und Analyse sozialer Daten, aber auch für die sozialwissenschaftliche Methodologie eine große Herausforderung bedeutet. Tatsächlich stellt sich die Frage, ob mit *Big Data* das Ende der herkömmlichen sozialwissenschaftlichen Methodenforschung, insbesondere der Umfrageforschung, droht. Nachfolgend sollen vor diesem Hintergrund einige Überlegungen zu den *methodologischen Implikationen* und *methodischen Fallstricken* von *Big Data* angestellt werden.

1) *Theorie- versus datengetriebene Erkenntnisgewinnung*: *Big Data*-Ansätze sind stark datenzentriert, folgen also einer Vorgehensweise, die „*data-driven*" und nicht „*theory-driven*" ist. Mitunter argumentieren manche Fachvertreter, dass schon die schiere Datenmenge und -tiefe theoretische Modelle über menschliches Handeln und gesellschaftliche Prozesse und Dynamiken überflüssig mache, da sie ohnehin zu grob skaliert seien, zu stark verallgemeinerten und eine Abstraktion der sozio-technischen Realität darstellten, die nur zu Ungenauigkeiten in der empirischen Forschung führe und stets nur modellhafte Ergebnisse liefern könne.[5] In einem „*data-driven*"-Ansatz ersetzt die große Datenmasse und -vielfalt sowie deren Durchforstung mittels hochkomplexer Algorithmen demnach die klassische deduktiv-nomologische Logik, der zufolge wissenschaftliche Erkenntnis immer zunächst mit einem Problem und einer allgemeingültig formulierten Hypothese beginnt, und die empirische Beobachtung dann nachrangig erfolgt, um die theoretische Aussage vorläufig zu akzeptieren oder zu falsifizieren.[6] *Big Data*-Ansätze folgen hingegen einer *induktiv-explorativen* Logik, ähnlich wie *Data Mining*-Verfahren. Aus wissenschaftstheoretischer Perspektive lässt sich dieser Diskurs recht leicht auflösen, wenn man explorative *Big Data*-Analysen als eine Vorstufe im wissenschaftlichen Erkenntnisprozess begreift: *Big Data* liefert eine schier grenzenlose Datenbasis und ist insofern „*data-driven*". Aber auch mittels *Big Data* gewonnene Beobachtungen werden in wissenschaftlichen Arbeiten immer zur problemorientierten Formulierung allgemeingültiger sozialer Mechanismen und Gesetzmäßigkeiten führen, die ihrerseits dann wieder anhand empirischer Beobachtungen überprüft werden. Genau an diesem Punkt setzt soziologische Theoriebildung ein; sie stellt die allgemeinen Thesen auf, die dann mit Hilfe sozio-technischer Daten überprüft werden. In dieser Hinsicht ändert sich also nichts für die (sozial-)wissenschaftliche Methodologie des kausalen Erklärens, Verstehens und

Prognostizierens von sozialen Phänomenen. Wissenschaft sucht nach möglichst einfachen Antworten und Lösungen, wenn sie dem Grundsatz der Logik der abnehmenden Abstraktion folgt: „Modelliere so einfach wie möglich und so realistisch wie nötig.“[7] Es sind diejenigen wissenschaftlichen Forschungsprogramme, die dasselbe mit weniger Annahmen erklären können, die sich durchsetzen. An all diesen wissenschaftstheoretischen Fundamenten vermag *Big Data* nicht zu rütteln, sondern fügt sich, wie aufgezeigt, als exploratives Entdeckungsinstrument nahtlos in den seiner Logik nach unveränderten Forschungsprozess ein. *Big Data* verhilft dann höchstens zu einer *deskriptiven Erklärung*, kann also Zusammenhänge in der Datenmatrix beschreiben, aber eben nicht dazu beitragen, Antworten auf die Frage nach dem „Warum“ derselben zu geben. Die Frage nach den allgemeinen sozialen Kausalmechanismen, die hinter einem statistischen Zusammenhang stehen und ihn bewirken, wird auch in Zukunft nicht ohne Theorie auskommen!

2) *Wissenschaftliche Standards*: Genügt *Big Data* den sozialwissenschaftlichen Standards der Gültigkeit (Validität), Zuverlässigkeit (Reliabilität), Wertfreiheit und intersubjektiven Nachprüfbarkeit der Messdaten? In vielen Fällen ist das fraglich, vor allem, wenn *Big Data*-Analysen mit Daten arbeiten, deren Ursprung und Qualität in den meisten Fällen ungeklärt ist: Die im Rahmen von *Big Data* verwendeten Quelldaten werden in der Regel nicht nach wissenschaftlichen Standards erhoben und auch nicht primär für wissenschaftliche Zwecke erzeugt, weshalb sie für gewöhnlich auch nicht den strengen Ansprüchen an Datenqualität genügen, die in wissenschaftlichen Analysen standardmäßig vorausgesetzt werden.[8] Ist die Datenqualität also ungeklärt, so stellen sich auch Fragen zur Qualität der Ergebnisse von *Big Data*-Analysen. Hinzu kommt, dass wissenschaftliche Erkenntnisgewinnung notwendigerweise die Möglichkeit der *Replikation* von Analyseergebnissen erfordert. Ist der freie Zugang zu den *Big Data*-Archiven jedoch für die Forschung nicht möglich, etwa weil Unternehmen wie Google der Öffentlichkeit keinen Einblick in die Datengenerierung erlauben, so fehlt den generierten Ergebnissen ein wesentliches Merkmal wissenschaftlicher Erkenntnisgewinnung: die intersubjektive Nachprüfbarkeit und Wiederholbarkeit empirischer Resultate. Umgekehrt kann aber auch argumentiert werden, dass sozialwissenschaftliche Daten nunmehr einer breiten Öffentlichkeit zugänglich gemacht werden können und dadurch mehr Transparenz hergestellt wird.[9] Entscheidend ist jedoch in jedem Fall, wie frei sich der Zugang zu den Datenquellen (hinsichtlich der zugrundeliegenden Algorithmen, die die Daten letztlich generieren) wirklich gestaltet.

3) *Indikatoren-Problem: Big Data*-Analysen stehen vor dem Problem, dass sie selten in der Lage sind, *multiple* Indikatoren für *individuelle* (!) Einstellungen, Wertorientierungen, Wahrnehmungen, Bewertungen, Emotionen, persönlichkeitsbezogene Charakteristika etc. liefern zu können – und das zudem mit akzeptabler Messgüte hinsichtlich der Kriterien Reliabilität und Validität. Ausschlaggebend dafür sind zwei Gründe: Zum einen müssen *Big Data*-Analysen häufig mit sehr indirekten Indikatoren Vorlieb nehmen, wenn auf Individualebene primär Verhaltensdaten vorliegen (zum Beispiel Konsumverhalten, digitale

Verhaltensspuren etc.). Aber von einer Verhaltensweise kann nicht fehlerfrei auf deren Ursachen rückgeschlossen werden. Gerade die Ursachen sozialen Verhaltens aber sind es, die die Sozialwissenschaften interessieren! Zum anderen liegen zwar gerade im Bereich *Social Media* auch individuelle Daten über schriftliche Meinungsäußerungen vor, diese Daten sind jedoch schwer vergleichbar, da sie nicht standardisiert sind, weshalb sie zur Auswertung aufwändigen, nicht-automatisierten Inhaltsanalysen unterzogen werden müssen. In Umfragen kann man Personen hingegen in standardisierter Form direkt nach ihren Werthaltungen, ihren Gefühlen, ihren Ängsten, ihren Verhaltensabsichten, ihren Wahrnehmungen, ihren Erwartungen und ihren Meinungen zu allen möglichen Themen befragen. Die Anwendung mehrerer Indikatoren pro theoretischem Konstrukt ermöglicht dabei eine im Vergleich deutlich höhere Zuverlässigkeit der Messungen.

4) *Repräsentativität*: Wenn bei *Big Data* die Trumpfkarte der schieren Größe ausgespielt wird, so darf das nicht darüber hinwegtäuschen, dass die Datenauswahl hochgradig selektiv bleibt. Gerade die Temporalität von *Social Media* macht deutlich, dass über entsprechende Portale gewonnene Daten keine bevölkerungsrepräsentativen Aussagen zulassen. Es gibt (noch?) kein *Social Media*-Portal, an dem alle (!) Menschen teilnehmen. Natürlich ist auch in der Umfrageforschung festzustellen, dass immer mehr soziologische Umfragen auf bevölkerungsweite Zufallsstichproben verzichten, sondern zum Beispiel auf selbstrekrutierten Web-Umfragen oder Schneeballverfahren in *Social Medi*a-Netzwerken wie Facebook etc. beruhen. Solche Stichprobenverfahren erreichen jedoch keine bevölkerungsweite Repräsentativität. Allerdings ließe sich dem entgegenhalten, dass auch in bevölkerungsweiten Zufallsstichproben derzeit nur noch eine Ausschöpfungsquote von ca. 20-30% erreicht wird, sodass diese Stichproben ebenfalls systematisch verzerrt sind. Der große Unterschied ist jedoch, dass Zufallsstichproben zumindest prinzipiell repräsentativ sein *können*, dass für jedes Element in der Population eine Auswahlwahrscheinlichkeit angebbar ist und dass Stichprobenfehler berechnet werden können. Das ist auch der Grund, warum 2.000 Befragte einer Zufallsstichprobe 20.000 Befragten einer nicht-zufälligen *Social Media*-Umfrage vorzuziehen sind, wenn Aussagen über die gesamte Population getroffen werden sollen. *Big Data*-Analysen auf Basis von *Social Media*-Daten sind *immer* selektiv und nicht bevölkerungsrepräsentativ. Das liegt in erster Linie daran, dass individuelle *Social Media*-Daten hochgradig durch *Selbstrekrutierung* entstehen. Anders könnte die Situation natürlich aussehen, wenn in Zukunft mittels *Big Data*-Vernetzung ein Ausmaß an Daten über alle Individuen und gesellschaftlichen sowie technischen Vorgänge in der Welt entstünde, das groß genug wäre, tatsächlich jeden und alles zu erfassen und Fragen der Stichprobenbildung somit schlicht überflüssig machte.[10] Aber realistisch ist eine solche Annahme aus heutiger Sicht nicht – insbesondere nicht aufgrund der nachfolgend diskutierten Einschränkungen.

5) *Systematischer Bias in den Messdaten*: Viele Daten aus dem Web, etwa aus dem Bereich der *Social Media*, sind sehr stark systematisch verzerrt. Wenn bekannte Musikgruppen oder große Unternehmen ihre Anzahl an Freunden in Facebook künstlich durch Anheuerung erhöhen, wenn individuelle Internetprofile in *Social*

Media zum *Impression Management* genutzt werden, um sich ein zweites, idealisiertes Ich zu schaffen, oder wenn in Bewertungsportalen zu Produkten oder Urlaubsreisen Schein-Bewertungen eingetragen werden, dann sind *Big Data*-Analysen, die solche Daten verwenden, ebenfalls systematisch verzerrt. Grundsätzlich können alle möglichen *ideologisch verzerrten Daten* Teil einer *Big Data*-Analyse werden. Auch die Abhängigkeit von automatisierten Algorithmen kann zu verzerrenden Ergebnissen führen, wenn beispielsweise in Suchmaschinen die Trefferanzeige von vorherigen Suchanfragen abhängt und die Trefferanzeigen durch bestimmte Methoden so verändert werden können, dass die Rangfolge der Treffer beeinflusst wird.

6) *Ethik und Datenschutz*: *Big Data-Verfahren* stehen vor großen Problemen hinsichtlich ihrer ethisch-sozialen Akzeptanz sowie möglichen Verstößen gegen Datenschutzrichtlinien – gerade und insbesondere auf Individualebene. So lange in einem Rechtsstaat datenschutzrechtliche Restriktionen festgelegt und persönliche Daten in welcher Form auch immer geschützt werden, können Daten im Rahmen von *Big Data* nicht in beliebiger Form erhoben, auf Personenebene verarbeitet und aus verschiedenen Datenquellen verknüpft werden. Hinzu kommt die datenschutzrechtliche Frage, ob öffentlich zugängliche Daten (etwa aus dem Bereich *Social Media*) überhaupt zu wissenschaftlichen Zwecken eingesetzt werden dürfen. Ebenso betrifft *Big Data* neue Dimensionen sozialer Kontrolle und sozial-ethische Fragen des Schutzes der Privatheit, aber auch Fragen nach Selbstbestimmung, Eigentumsrechten und Kommerzialisierung der Daten. Wem gehören zum Beispiel personalisierte und standortbezogene Daten, die mit *smarten* Technologien (beispielsweise *Smart Watches* oder *Smart Cars*) erhoben werden? Wer darf auf diese zugreifen und dürfen solche Daten Dritten kommerziell angeboten werden? Die Forschung mit *Big Data* erfordert demnach eine eingehende Auseinandersetzung mit solchen ethischen und datenschutzrechtlichen Fragen.

7) *Datenprobleme auf Individualebene*: *Big Data* verleitet schnell zu der Annahme, dass ein „gläserner Mensch“ entsteht, der keine Geheimnisse mehr hat. Gleichwohl besteht zum jetzigen Zeitpunkt eines der größten Probleme von *Big Data* eben darin, die Daten aus unterschiedlichsten Quellen zweifelsfrei einzelnen Individuen zuordnen zu können − sei es aus datenschutzrechtlichen Gründen oder aus Gründen der fehlenden Informationen in den einzelnen Datensätzen. In Online-Datenquellen sind es beispielsweise zumeist Accounts, denen einzelne Vorgänge zugeordnet werden können, aber nicht reale Personen. Problematisch ist das besonders dann, wenn Personen mehrere Accounts bei einem Anbieter haben oder sie sich einen Account mit mehreren anderen Personen teilen, was der Datenbasis jedoch nicht entnommen werden kann. Wenn *Big Data*-Analysen ihre größten Probleme also auf der Individualebene haben, dann verlieren sie für viele sozialwissenschaftliche Fragen an Wert. Denn liefern *Big Data*-Analysen keine gesicherten Individualdaten, so entsteht schnell die Gefahr eines *ökologischen Fehlschlusses* (von gesellschaftlichen Makro-Daten kann nicht fehlerfrei auf die Individualebene geschlossen werden) oder eines *kollektivistischen Fehlschlusses* (von einer Makro-Situation kann nicht kausal auf eine andere Makro-Situation

geschlossen werden). In der Soziologie hat sich in diesem Zusammenhang der Methodologische Individualismus gegen den Kollektivismus weitestgehend durchgesetzt: Eine soziologische Erklärung muss demnach Verbindungen zwischen der Makro- und der Mikro-Ebene herstellen können und auf der Individualebene eine Gesetzmäßigkeit benennen können, sonst ist es keine vollständige ursächlich erklärende soziologische Erklärung.[11] Dazu sind sowohl eine gute theoretische Argumentation als auch für entsprechende empirische Analysen geeignete Individualdaten notwendig – und an beidem scheint es *Big Data* bis auf weiteres zu mangeln.

Big Data kann in Zukunft auch bedeuten, dass Umfragedaten schlicht in das große Datennetz integriert werden − Umfragedaten sind dann wie alle anderen sozio-technischen Daten ein Puzzleteil in einer gigantischen Rechenoperation. Aber selbst dann müssten weiterhin Umfragen durchgeführt werden, um die *Big Data*-Analysen mit aktuellen Umfragedaten zu füttern.

Die Sozialwissenschaften müssen demnach keinen Ausverkauf durch *Big Data* befürchten, weder auf methodologischer Ebene noch auf Ebene der Erhebung sozial relevanter Daten. Vielmehr sollten die Sozialwissenschaften *Big Data* als einen wertvollen neuen zusätzlichen (!) Datenzugang zur sozio-technischen Realität begreifen und als solchen nutzen. Wenn Informationen aus der digitalen und physischen Welt zusammenkommen, dann können die betreffenden Daten geschickt als Kontextinformationen in soziologischen Analysen genutzt werden. In der sozialwissenschaftlichen Methodenforschung sind solche Informationen als sogenannte „nicht-reaktive Daten" und „Paradaten" längst bekannt. Jede Handlung hinterlässt Spuren, digital wie physisch, und diese Spuren können für sozialwissenschaftliche Untersuchungen fruchtbar eingesetzt werden. Und wenn es mit den *Big Data*-Technologien möglich werden sollte, diese Zusatzinformationen in ungeahnter Datenfülle zugänglich zu machen, und das dazu noch auf allen möglichen Aggregatebenen zwischen einzelnen Individuen, Organisationen und gesellschaftlichen Prozessen, dann kann darin eine große Chance für die Sozialwissenschaften liegen. Insbesondere in Hinblick auf Untersuchungen von Zeitreihen besitzt *Big Data* ein sehr großes Potenzial. Freilich ist derzeit noch offen, ob solche Daten zukünftig auch auf Individualebene zugeordnet werden können. Aber auf Kontextebene sind hier sicherlich sehr große Datenschätze für soziologische Analysen in Aussicht. Grundsätzlich ist aber auch zu sagen, dass die Art der verwendeten Daten stets geeignet sein muss zur Beantwortung der jeweiligen Forschungsfrage: „in some cases, small is best".[12]

Die Nutzung von *Big Data* bedeutet für die empirische Sozialforschung auch neue Herausforderungen hinsichtlich der *interdisziplinären* Zusammenarbeit. Denn *Big Data* macht es für eine umfassende Analyse und Bewertung der Daten erforderlich, deren Ursprung und somit deren zugrundeliegenden Algorithmus zu verstehen. Damit werden interdisziplinäre Kooperationen insbesondere mit Informatik und Informationstechnik notwendig.[13]

Dass *Big Data* auch fundamentale gesellschaftliche und sozial-ethische Fragen

aufwirft, steht außer Frage und bietet Stoff für eine ganz andere und wichtige soziologische sowie gesellschaftspolitische Debatte. *Big Data* hat in dieser Hinsicht vermutlich ein größeres Potenzial, das zukünftige gesellschaftliche Zusammenleben nachhaltig zu prägen und massiv zu verändern, als die sozialwissenschaftliche Methodenforschung in Zukunft überflüssig zu machen oder radikal umzugestalten. Die Sozialwissenschaften sind jedenfalls gut beraten, beide Entwicklungen kritisch zu begleiten.

Dieser Beitrag ist Teil eines Themenschwerpunkts zu Big Data. Weitere Texte finden Sie hier.

Fußnoten

1 Die jährlichen Publikationszahlen ergeben sich aus einer themengebundenen Schlagwortsuche nach dem Begriff „Big Data“ (mit Anführungszeichen) jeweils innerhalb eines einjährigen Zeitraumes von 2008 bis 2015 (letzter Zugriffszeitpunkt: 16.12.2015).

2 Vgl. Sandra González-Bailón, Social Science in the Era of Big Data, in: Policy and Internet 5 (2013), 2, S. 147–160, online unter: papers.ssrn.com/sol3/papers.cfm.

3 Siehe dazu u.a. Elke Theobald / Ulrich Föhl, Big Data wird zu Smart Data. Big Data in der Marktforschung, in: Joachim Dorschel (Hrsg.), Praxishandbuch Big Data, Wiesbaden 2015, S. 112–123.

4 Vgl. Viktor Mayer-Schönberger / Kenneth Cukier, Big Data. Die Revolution, die unser Leben verändern wird, München 2008.

5 Chris Anderson, The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, in: Wired magazine (2008), 6, online unter: www.wired.com/2008/06/pb-theory/.

6 Vgl. dazu den klassischen Text von Karl R. Popper, Logik der Forschung [1935], 10. Aufl., Tübingen 1994.

7 Hartmut Esser, Soziologie. Allgemeine Grundlagen, 2. Aufl., Frankfurt am Main / New York 1996, S. 140.

8 David Lazer et al., The Parable of Google Flu: Traps in Big Data Analysis, in: Science 343 (2014), 6176, S. 1203–1205, online unter:

scholar.harvard.edu/files/gking/files/0314policyforumff.pdf.

9 Danah Boyd / Kate Crawford, Critical Questions for Big Data, in: Information, Communication & Society 15 (2012), 5, S. 662–679, online unter: www.tandfonline.com/doi/pdf/10.1080/1369118X.2012.678878.

10 Mike Savage / Roger Burrows, The Coming Crisis of Empirical Sociology, in: Sociology 41 (2007), 5, S.885–899.

11 Vgl. Esser, Soziologie, S. 98ff.

12 Boyd/Crawford, Critical Questions for Big Data, S. 670.

13 Vgl. González-Bailón, Social Science in the Era of Big Data, S. 158.