Big Data

10

Miriam Trübner und Andreas Mühlichen

10.1 Der Hype um Big Data

Big Data ist zunächst einmal ein großes Versprechen, eine Gelddruckmaschine, Waffe im Krieg gegen den Terror und im Kampf gegen Verbrechen, Reformator von Verwaltungen, diagnostisches Wunderkind der Medizin, Stein der Weisen empirischer Wissenschaft und der kleine Helfer im Alltag einer ubiquitär vernetzten Welt. Big Data ist aber auch ein riesiger Hype, Allmachtsfantasie, gläserner Bürger, Kunde, Patient, Feind der Privatheit – und möglicherweise auch das drohende Ende konventioneller Umfrageforschung. Die positive Erwartungshaltung der empirischen Sozialforschung, in Echtzeit Daten über alles und jedes zu haben und endlich Antworten auch auf schwer erfassbare Fragen zu erhalten, ist dabei von der Angst begleitet, dass große Teile ihres etablierten Instrumentariums obsolet werden.

Mit diesem Beitrag soll ein Überblick über das Big-Data-Phänomen gegeben werden, der es erlaubt, zwischen Potential und Hype zu unterscheiden. Um dies zu beurteilen, wird im Folgenden Big Data in Differenz zur konventionellen Umfrageforschung (Reinecke, Kapitel 62 in diesem Band) betrachtet und die sich hieraus ergebenden Implikationen herausgearbeitet.

10.2 Was ist Big Data?

Auch wenn es (zumindest bisher) keine einheitliche Definition von Big Data gibt, sind wesentlicher Bestandteil vieler Denfinitionen drei Eigenschaften, die als die „drei V“ bezeichnet werden: Es handelt sich um in digitaler Form vorliegende Daten großer Masse („volume“), die schnell entstehen („velocity“) und sehr unterschiedlicher Art und Herkunft sein können („variety“) (Laney 2001). Je nach Fokus kommen aber auch verschiedene weitere „V“ in Betracht, die beispielsweise den Wert der Da-

N. Baur und J. Blasius (Hrsg.), *Handbuch Methoden der empirischen Sozialforschung*, https://doi.org/10.1007/978-3-658-37985-8_10

ten („value"), die Richtigkeit der Daten („veracity") oder die (In)Konsistenz der Daten über die Zeit („variability") beschreiben.

Bei aller Einigkeit über die unterschiedlichen „V" wird bei der praktischen Anwendung verschiedener Instanzen Big Data unterschiedlich gehandhabt: Wenn Betreiber sozialer Netzwerke (Schrape/Siri, Kapitel 92 in diesem Band) von der Analyse von Big Data sprechen, dann kann davon ausgegangen werden, dass in Echtzeit Daten mindestens im Gigabyte-Bereich automatisiert ausgewertet werden sollen. Sprechen empirische Sozialforscher hingegen von der Inklusion von Big Data in ihre Analysen, dann meinen sie damit meistens, dass sie zwar (in der Regel bestehende, eher konventionelle) Datensätze mit digitalen Daten (Thimm/Nehls, Kapitel 69 in diesem Band), wie sie unter Big Data erzeugt werden, füllen oder anreichern, dabei aber die Analyse nicht in Echtzeit läuft und oft auch nur ein Bruchteil der Daten zur Analyse verwendet wird (vgl. kritisch zu diesem Vorgehen auch Heiberger/Riebling 2016).

Die unter Big Data gefassten Daten lassen sich in (taxonomisch nicht immer ganz sauber abgetrennte) Subtypen klassifizieren:

1. *Metadaten* (Wenzig, Kapitel 92 in diesem Band): Wie beispielsweise Vorratsdaten, Paradaten (Felderer et al., Kapitel 29 in diesem Band) und Logdaten (Schmitz/Yanenko, Kapitel 70 in diesem Band). Ein Großteil dieser Daten wird automatisiert im Rahmen digitaler Prozesse erzeugt, oft ohne explizite Kenntnis des Nutzers über deren Existenz oder zumindest deren spezifischen Inhalt.
2. *Transaktionsdaten:* Dies sind z. B. Kreditkartenabrechnungen und (nicht notwendigerweise automatisiert erzeugte) administrative Daten (Salheiser, Hartmann/Lengerer, Kapitel 80 und 106 in diesem Band) wie Steuererklärungen.
3. *Nutzergenerierte Daten* (Thimm/Nehls, Kapitel 69 in diesem Band): Dies sind durch Nutzer digitaler Dienste erzeugte Daten, die vor allem im Rahmen von Web 2.0-Anwendungen anfallen, so in sozialen Netzwerken (Schrape/Siri, Kapitel 92 in diesem Band), bei Chats (Nam, Kapitel 91 in diesem Band), auf Youtube (Traue/Schünzel, Kapitel 93 in diesem Band) oder bei Twitter (Mayerl/Faas, Kapitel 90 in diesem Band).

Besonders vielversprechend für die empirische Sozialforschung erscheinen die Nicht-Reaktivität bei der Datenerzeugung, dies potentiell in Echtzeit, die Vielfältigkeit der Daten, die geringen Kosten und die digitale Datenverarbeitung, was aber zu einer Reihe methodischer Probleme führen kann. Viele der Besonderheiten und der Probleme, die dabei im Rahmen von Big Data als neu diskutiert werden, sind bereits bekannt, beispielsweise die Existenz von und die Methodendebatte zu prozessgenerierten Massendaten (Bick et al. 1984; siehe auch Baur 2009 über die Gefahr des Vergessens und die Notwendigkeit der Integration bekannter Ergebnisse in die aktuelle Debatte). Neu sind die Größenordnungen, um die es im Kontext von Big Data geht, sowie auch der Grad der Vernetztheit, in der diese Daten anfallen und verarbeitet werden können, was sich auf den gesamten Umgang mit Big Data auswirkt.

10.3 Operationalisierung und Messung

Im Kontext eines sich ausweitenden „Internet der Dinge" („Internet of Things", IoT), bei dem immer mehr Bestandteile physischer Umwelt an die Infrastruktur Internet angebunden und damit potentiell in die Produktion von Big Data eingebunden werden können – bis hin zum mit einem RFID-Funkchip („radio-frequency identification") ausgestatteten Joghurtbecher –, verspricht es mit Big Data möglich zu werden, objektiv erscheinende, weil nicht-reaktiv erhobene Daten in Echtzeit über den Großteil der Bevölkerung in zunehmend mehr Situationen erfassen zu können. Dabei ist zunächst einmal *unklar, was überhaupt gemessen wird.* Denn die Erzeugung der Daten findet nicht intentional mit spezifischem Bezug auf eine bestimmte Fragestellung statt, sondern sie ist eher motiviert durch technische Abläufe oder, im nutzergenerierten Fall, durch die Verwendung eines Dienstes.

Insbesondere dann, wenn es sich bei den Ausgangsdaten um Metadaten handelt, haben viele der so gewonnenen Daten zunächst nur einen sehr *geringen Informationsgehalt* – also beispielsweise eine große Anzahl Variablen mit Zeitstempeln sowie Längen- und Breitengrade im Rahmen einer GPS-gestützten Mobiltelefonanwendung (Kandt, Kapitel 199 in diesem Band). Um hieraus eine inhaltlich für weitere Analysen sinnvolle Variable zu entwickeln – beispielsweise ein Bewegungsprofil –, müssen solche Variablen miteinander verbunden werden (vgl. hierzu Abschnitt 10.6 unten). Dabei ist nicht immer von vornherein klar, welche Variablen sich eignen, um welche Konstrukte abzudecken bzw. welche Zusammenhänge überhaupt über die Variablen abgebildet werden können.

Eine mögliche Vorgehensweise zur Entdeckung solcher Zusammenhänge ist ein eher explorativer Ansatz, der im Fall von Big Data aufgrund der sehr vielen, in ihrer Bedeutung sehr unklaren Daten aber schnell in einem eher wahllosen Durchpermutieren aller möglichen Korrelationen zu enden droht – und damit im Auffinden von *Scheinkorrelationen* (Kühnel/Dingelstedt, Kapitel 46 in diesem Band), die durch einen zufälligen gleichen zeitlichen Verlauf von zwei voneinander unabhängigen Variablen entstehen. Bei „Google Correlates" etwa ließen sich, bis zur Einstellung des Dienstes 2019, Korrelationen zwischen der Auftretungshäufigkeit verschiedener Suchbegriffe über Zeit darstellen, so war z. B. das Wort „Stuhl" fast perfekt mit „Traueranzeige" korreliert ($r = 0.975$). „Google Correlates" orientierte sich an der Häufigkeit der Suchen über die Zeit, in dem Fall haben nicht die gleichen Nutzer nach Stuhl und Traueranzeige gesucht, sondern die Suchhäufigkeiten verlaufen zufällig nach dem gleichen Muster, es gibt keine kausale Abhängigkeit (Diaz-Bone, Kühnel/Dingelstedt, Kapitel 5 und 46 in diesem Band). Die fälschliche Übertragung eines Zusammenhangs von Aggregatdaten der Makroebene (Graeff, Kapitel 102 in diesem Band) auf die Individualebene ist als *ökologischer Fehlschluss* bekannt. Dieses methodologische Grundproblem der Statistik erhält durch Big Data eine größere Bedeutung, weil es hier eben gerade Teil der Methode ist, sehr viele für sich unbedeutend erscheinende Daten miteinander zu etwas Sinnhaftem zu verbinden.

Während bei einer klassischen Umfrage (Reinecke, Kapitel 62 in diesem Band) viel Vorarbeit in die Operationalisierung (Stein, Kapitel 8 in diesem Band) gesteckt wird, kann bei Big Data in der Regel erst ex post versucht werden, aus den vorliegenden Daten solche zu generieren, die dann mit einem zu messenden Konstrukt korrespondieren – die Erhebung findet also vor der Konstruktion eines Messinstrumentes statt. Diese Situation ist vergleichbar mit der von Sekundäranalysen in der klassischen Sozialforschung (Schupp, Kapitel 85 in diesem Band), mit dem Unterschied, dass sehr viel mehr Variablen vorliegen können, bei denen zunächst recht unklar sein kann, wie sie überhaupt zusammenhängen bzw. was sie messen und ob eine solche Messung überhaupt valide und reliabel ist.

Ein Beispiel für die Problematik fragwürdiger *Validität* und *Reliabilität* (Krebs/Menold, Kapitel 35 in diesem Band) ist „Google Flu Trends", was mit Grippe in Zusammenhang stehende Google-Suchanfragen (Symptome, Medikamente etc.) verwendete, um die Verbreitung der Grippe in Echtzeit zu prognostizieren. Da Gesundheitssurveys aufgrund traditioneller Datenerhebungsmethoden wie telefonischen Befragungen (Hüfken, Kapitel 68 in diesem Band) oder persönlich-mündlichen Befragungen (Stocké, Kapitel 67 in diesem Band) erst mit einem gewissen zeitlichen Abstand zur Befragung Ergebnisse liefern können, hoffte man mit Echtzeitmessung möglichst direkt auf eine Ausbreitung von Krankheiten reagieren zu können. „Google Flu Trend" überschätzte jedoch die tatsächlichen Zahlen der Grippewelle deutlich. Dies liegt zum einen daran, dass nicht jeder grippebezogene Suchbegriff („Symptome", „Medikamente") darauf hindeutet, dass diese Person tatsächlich krank ist (Validitätsproblem). Zum anderen konnte die Überschätzung durch Google auf dessen eigenen Algorithmus zurückgeführt werden. Bei der Eingabe von Grippe-bezogenen Symptomen wie „Fieber" werden von Google Suchbegriffe vorgeschlagen, die sich zwar auf Grippe beziehen, wobei sich die Vorschläge des Algorithmus aber über Zeit unterscheiden dürften (Reliabilitätproblem) – was dann zu Verzerrungen in der Messung führt. Außerdem erhöht die erscheinende Auswahlliste die Wahrscheinlichkeit, den „Grippe"-Suchbegriff anzuklicken, was wiederum positiv in die Vorhersage der Verbreitung der Grippe-Welle eingeht (Lazer et al. 2014).

Auch der Gegenstandsbereich dessen, was über Big Data operationalisierbar ist, unterscheidet sich typischerweise zu dem, was im Rahmen konventioneller Umfragedaten erreichbar ist. Standardisierte sozialwissenschaftliche Befragungen zielen in der Regel explizit darauf ab, über aufwändig operationalisierte Item-Batterien (Porst, Kapitel 73 in diesem Band) Einstellungen, Werte oder Persönlichkeitsmerkmale zu messen. Im Rahmen von Big Data hingegen sind entweder aus Metadaten in der Regel Variablen bezüglich des Verhaltens von Individuen ableitbar, oder es liegen im Fall von nutzergenerierten Einträgen (Mayerl/Faas und Nam, Kapitel 90 und 91 in diesem Band) nicht-standardisierte Meinungsäußerungen auf Social-Media-Plattformen (Schrape/Siri, Kapitel 92 in diesem Band) vor. Im ersten Fall ist dann fraglich, inwiefern überhaupt eine Messung jenseits des Verhaltens durch Metadaten möglich ist; im Fall der nutzergenerierten Daten bedarf es aufwändiger qualitativer Inhalt-

analysen (Mayring/Frenzl und Taddicken, Kapitel 43 und 107 in diesem Band), die entweder den Vorteil der Schnelligkeit von Big Data in Frage stellen oder aber voraussetzen, dass automatisierte Auswertungen mit Hilfe von Algorithmen, die natürliche Sprache „verstehen", diese Aufgabe zuverlässig übernehmen (für praktische Anwendungsbeispiele siehe Klochikhin/Boyd-Graber 2017). Dabei existiert dann immer noch das Problem, dass die qua Inhaltsanalyse gemessenen Konstrukte eben nicht standardisiert erhoben wurden, man also darauf angewiesen ist, dass ein Fall sich zu einem Thema überhaupt äußert (Mayerl 2015).

10.4 Auswahlverfahren

Eine der Eigenschaften, die Big Data so attraktiv macht, ist, dass Daten in so unüberschaubarer Menge vorliegen. Groß ist dabei nicht nur die Anzahl verschiedener Variablen oder, dank Echtzeitmessung, die Vielzahl von Messpunkten, sondern auch die Reichweite der Messung, die umso mehr Personen umfasst, je mehr Nutzer es von Diensten gibt und je mehr Geräte und Gegenstände potentiell ein Datum über eine Person anlegen. Auch wenn es dabei so erscheint, als ob über jeden Daten vorliegen, gestaltet sich die Situation aus methodischer Perspektive deutlich komplizierter.

10.4.1 Grundgesamtheit und Stichprobe

Bei der Erzeugung von Big Data steht nicht so sehr ein mehr oder weniger elaboriertes Sampling-Modell im Hintergrund, sondern die Situation ist eher vergleichbar mit Vollerhebungen, dann aber weitgehend undefinierter Grundgesamtheiten – die dann mit der üblicherweise schwer abgrenzbaren Nutzerschaft einer Technologie im Sinne eines Gerätes oder eines Dienstes korrespondieren. Dies bringt viele Probleme mit sich, insbesondere auch in Bezug auf mögliche Verzerrungen. Außerdem stellt sich die Frage, ob eine Vollerhebung nicht eine definierte Grundgesamtheit voraussetzt (vgl. ausführlicher zu Anforderungen und Begrifflichkeiten rund um Grundgesamtheit und Stichprobe von Häder/Häder, Kapitel 27 in diesem Band), was dann letztlich bedeutet, dass man es mit einer *willkürlichen Stichprobe* zu tun hat.

Aufgrund der Schwierigkeit der *Definition der Grundgesamtheit* und der vorliegenden Größe der Datenmengen werden leicht vorschnelle Rückschlüsse auf eine vermeintlich größere Grundgesamtheit gemacht, also auch auf solche Individuen, welche die Plattform oder Dienste gar nicht verwenden. Es wäre also ein Fehlschluss, von prozessgenerierten Daten beispielsweise von Dating-Webseiten Rückschlüsse auf die Partnersuche und menschliches Handeln von einer Grundgesamtheit außerhalb dieser Dating-Plattformen zu ziehen. In diesem Fall sind die Mitglieder der Dating-Plattform die Grundgesamtheit, unter der eine Vollerhebung vorgenommen wurde,

aber diese Plattform muss keinesfalls repräsentativ für alle Partnersuchenden sein (Schmitz et al. 2009).

Zudem muss beachtet werden, dass im Bereich nutzergenerierter Daten *Selbstselektion* ein gewichtiges Problem darstellt. Möchte man beispielsweise die Einstellung zum Klimawandel unter Twitter-Nutzern (Mayerl/Faas, Kapitel 90 in diesem Band) analysieren, ist zu beachten, dass gerade diejenigen, die eine starke Meinung zu diesem Thema haben, am häufigsten darüber twittern, während meinungslose Personen sehr wahrscheinlich keine Tweets zu diesem Thema senden. Die aus diesen Daten abgeleiteten Analysen sind dadurch verzerrt und können nicht verallgemeinert werden.

Diese Probleme sind aber nicht erst seit Big Data relevant, sondern existieren seit den Anfängen der empirischen Sozialforschung. Ein altbekanntes Beispiel für falsche Schlussfolgerungen aus großen Datenmengen ist die Wahlumfrage des „Literary Digest" aus dem Jahre 1936. Dabei wurden insgesamt 10 Millionen Amerikaner (Bruttostichprobe), deren Adressen aus Abonnentenlisten des „Literary Digests" und Listen über Telefonanschlüsse generiert worden sind, angeschrieben und über die anstehende Präsidentschaftswahl befragt. Auf der Grundlage von tatsächlich realisierten 2,4 Millionen Befragungsteilnehmern (Nettostichprobe) wurde fälschlicherweise der Sieg Landons statt Roosevelts prognostiziert. Grund dafür war zum einen, dass sich die Bruttostichprobe, die angeschrieben wurde, systematisch von der wahlberechtigten Bevölkerung der USA unterschied und zum anderen, dass ein stärkeres Interesse am Thema der Befragung – der Präsidentschaftswahl – bei den Landon-Befürwortern vorhanden war und diese somit eher teilnahmen und sich in der Nettostichprobe befanden (Bryson 1976). Die Diskussion darüber, dass *viele* Daten nicht mit *guten* Daten gleichzusetzen sind, hielt später vor allem bei der Einführung der Online-Befragung und Access-Panels (Wagner-Schelewsky/Hering, Kapitel 70 in diesem Band) erneut Einzug – und es erscheint notwendig, sie auch für Big Data wieder zu führen.

10.4.2 Nonresponse

Wie trügerisch die aufgrund der Datenmenge empfundene Vollständigkeit bei Big Data ist, zeigt zusätzlich die Nonresponse-Problematik (Engel/Schmidt, Kapitel 29 in diesem Band), die durch Big Data vermeintlich umgangen wird. Denn auch bei Metadaten bieten sich Möglichkeiten, sich der Datenaufzeichnung zu verweigern – was als Entsprechung von klassischer Nonresponse für den Big Data-Kontext angesehen werden kann.

Ein Beispiel hierfür wäre das Setzen des „do-not-track-http-header"-Feldes in einem Browser. Unabhängig davon, ob sich die angefragten Seiten tatsächlich an diese Anweisung halten oder nicht, kann hiermit ein Nutzer einer besuchten Seite signalisieren, dass der Verwendung der beim Surfen anfallenden Daten zu Analysezwecken widersprochen wird – und damit z. B. Web-Analytics (die Analyse von Webseiten-

besuchern und deren Eigenschaften) nicht durchgeführt werden sollen. Auch durch andere *technische Schutzmechanismen* wie „Browser-Plugins", die Cookies restriktiv verwalten, oder „Tor-Netzwerke", die zur Anonymisierung eingesetzt werden können, lassen sich Rückschlüsse auf Nutzer von Internetdiensten oder -inhalten erschweren und sie können damit zu einer Nonresponse führen. Inwiefern hierüber für Nutzer tatsächlich ein zuverlässiges Mittel informationeller Kontrolle (Mühlichen, Kapitel 23 in diesem Band) erreichbar wird, ist in vielen Umständen fraglich und zuweilen gibt es bei technischen Schutzmechanismen ein Wettrüsten zwischen Instanzen, die informationelle Selbstbestimmung schützen, und solchen, die Daten auswerten wollen. Dennoch können, auch wenn dabei immer noch *Metadaten* (Wenzig, Kapitel 101 in diesem Band) anfallen (z. B. dass eine Seite besucht wurde), diese Daten gegebenenfalls nicht mehr mit anderen kombiniert (Cielebak/Rässler, Kapitel 31 in diesem Band) und personenbezogen verarbeitet werden, was den theoretisch erreichbaren Informationsgehalt einer Big-Data-Analyse entsprechend reduziert.

10.5 Datengewinnung

Wie bereits erläutert, werden die Daten im Fall von Big Data typischerweise nicht mit einer spezifischen Fragestellung zu einem Gegenstand vom Forscher selbst erhoben, sondern im Rahmen eines technischen Prozesses oder durch die Nutzung eines Dienstes für einen gänzlich anderen Zweck erzeugt; die Daten selbst liegen also bereits vor, und zwar bei Dritten. Eine wesentliche Frage ist deshalb, wie denn diese vorliegenden Daten überhaupt sinnvoll abgerufen und in einen für die eigene Nutzung brauchbaren Datensatz überführt werden können:

1. Die Vielfältigkeit und Größe der prozessgenerierten Daten bieten verschiedene Möglichkeiten des Datenzugangs. Zum einen besteht die Möglichkeit, per *copy & paste* Daten aus einer vorher bestimmten Quelle aus dem Netz zu filtern – etwa Informationen auf Webseiten (Schünzel/Traue, Kapitel 88 in diesem Band) – und diese in gängige Software für quantitative (wie SPSS, Stata, R; vgl. Lück/Landrock, Kapitel 33 in diesem Band) oder qualitative (z. B. MAXQDA oder Atlas.ti) Datenanalyse zu übertragen (Kuckartz/Rädiker, Kapitel 32 in diesem Band). Dieses Prozedere ist aber zu ineffektiv und fehleranfällig, um echte Big-Data-Analysen sinnhaft durchführen zu können. Es eignet sich aber als Notlösung, um einen bestehenden (konventionellen) Datensatz in einer Tabelle um Eigenschaften aus einem Big-Data-Datensatz zu erweitern.
2. Wenn sich Daten ausschließlich auf Webseiten befinden, bietet sich *Web Scraping* an. Nicht jede Webseite erlaubt Web Scraping, oder nur in eingeschränktem Maße (für eine Einführung vergleiche Neylon 2017).
3. Als Alternative bieten sich Schnittstellen zur Anwendungsprogrammierung, sogenannte *APIs* (application programming interface), an, die von Betreibern einer

Plattform angeboten werden, um Externen zu ermöglichen, von ihnen vorher festgelegte Daten auszulesen. Um mittels API Zugriff auf Daten zu erlangen, bedarf es Kenntnisse von Programmiersprachen wie beispielsweise Python. Auch wenn über eine API ein professioneller Zugriff auf einen Big-Data-Bestand möglich ist, verschärft sich damit das Stichprobenproblem von Big-Data-Analysen: Denn bei der Abfrage entscheidet allein der Anbieter, welche Daten er wann freigibt. Dabei ist nicht unbedingt ersichtlich, wie vollständig die abgerufenen Daten sind und welche versteckten Restriktionen durch den Anbieter gesetzt sein mögen. Man hat also möglicherweise keine Informationen darüber, ob bestimmte Typen von Fällen systematisch ausgelassen werden (denkbar wären zum Beispiel bestimmte Nutzertypen oder Premiumnutzer eines Dienstes), oder inwiefern sich das, worauf ein Zugriff gewährt wird, über die Zeit verändert, was eine unbekannte systematische Verzerrung der Stichprobe nach sich zieht. Auch wenn, wie oben diskutiert, im Fall von Big Data anstelle eines Samplingprozesses eher die Idee einer Vollerhebung tritt, bei der aber bereits unklar ist, welcher spezifischen Population, kommt nun das folgenschwere Problem hinzu, dass es bereits durch die Abfrage der Daten – wie über den Weg einer API – von außen nicht beurteilbare Lücken in den vermeintlich vollständigen Daten geben kann und es sich letztendlich um eine willkürliche Stichprobe handelt (Morstatter et al. 2014).

10.6 Datenbereinigung

Im Rahmen der Diskussion der Messmodelle bei Big Data wurde deutlich, dass Daten, die in diesem Rahmen erzeugt werden, jeweils für sich genommen oft nur einen sehr geringen Informationsgehalt haben – beispielsweise den Zeitpunkt, zu dem eine spezifische Webseite aufgerufen wurde. Erst wenn mehrere solcher Daten miteinander kombiniert werden, lassen sich höherwertige Informationen daraus ablesen.

Ein großes Problem beim Zusammenführen verschiedener Datensätze aus unterschiedlichen Quellen (Cielebak/Rässler, Kapitel 31 in diesem Band) ist aber, einzelne Fälle zu identifizieren und in diesem Sinne doppelte sowie falsche Fälle zu bereinigen. Typische Probleme dabei sind:

- *Mehrfachzählung derselben Person:* Hier werden beispielsweise Menschen, die wiederholt das Gleiche in Google-Suchanfragen eingeben, mehrfach gezählt. Dieses Problem existiert auch bei Seitenbesuchen, wenn sich bei mehreren Zugriffen der gleichen Person die IP ändert (dynamische IPs), diese mehrere Konten verwendet oder von mehreren Geräten aus auf Apps, Webseiten, Dienste oder ähnliches zugreift. Diese mehrfach im Datensatz vorhandenen identischen Fälle führen dann zu verzerrten Ergebnissen.
- *Mehrfachnutzung desselben Accounts:* Des Weiteren kann kaum kontrolliert werden, ob andere Personen einen fremden Account nutzen, wie es beispielsweise

beim Kauf auf Amazon unter dem Konto eines anderen Familienmitgliedes oder beim Nutzen einer anderen Payback-Karte der Fall ist. Ereignisse werden so einer Person zugeschrieben, auf die diese gar nicht zutreffen.

- *Nichtmenschlicher Nutzer:* Insbesondere sich wiederholende Aufgaben im Netz werden oft nicht durch Nutzer, sondern mit Hilfe sogenannter Bots (Programme, die automatisiert bestimmte Aktivitäten durchführen und dabei als Nutzer auftreten können) durchgeführt, wie sie beispielsweise bei Chats zur Kundeninformation auf Webseiten eingesetzt werden. Diese dürfen bei Analysen nicht mit echten Nutzern verwechselt werden.
- *Fake-Accounts:* Eine mögliche Quelle massiver Verzerrungen entsteht durch Fake-Accounts, mit deren Hilfe entweder reale Personen oder sogenannte Social Bots gezielt Fehlinformationen zu lancieren versuchen. Dabei kann es sich beispielsweise um gefälschte Kommentare oder Produkt-Bewertungen handeln oder um weit gestreute, gegebenenfalls durch einen Automatismus erzeugte Informationen in sozialen Medien, die eingesetzt werden, um einen falschen Eindruck der Wichtigkeit bzw. der Echtheit dieser zu vermitteln. Gehen diese Daten unreflektiert in Big-Data-Analysen ein, muss auch hier mit Verzerrungen der Ergebnisse gerechnet werden.

Ist ein Datensatz erstellt, wird er einer Plausibilitätsprüfung (Lück/Landrock, Kapitel 33 in diesem Band) unterzogen, um falsche und sich widersprechende Einträge zu identifizieren. Wie schwierig dies im Rahmen automatisiert erzeugter Big Data werden kann, wird beispielsweise bei der Frage nach Bewegungsprofilen (Kandt, Kapitel 119 in diesem Band) deutlich: Soll untersucht werden, wo sich eine Person aufgehalten hat, so können hierzu verschiedene Geräte zum gleichen Zeitpunkt unterschiedliche Daten liefern. Nehmen wir an, ein Pendler bewegt sich mit seinem Auto zwischen zwei Städten und hat sein Mobiltelefon bei sich. Hier könnten mindestens drei unterschiedliche Daten zum Standort vorliegen: Erstens meldet sich das Telefon bei den Funkmasten des Mobiltelefonnetzes an, die, je nach Standort, unterschiedlich große Funkzellenbereiche abdecken und so Lokalisierungen ermöglichen, die im ländlichen Bereich ungenauer sind als im städtischen. Zweitens können gerade in Ballungsgebieten zusätzlich über sogenannte WLAN-basierte Ortung relativ präzise Standortdaten generiert werden. Schließlich liefert das Satellitennavigationssystem des Autos als dritte Quelle für die Standortinformation die genausten Lokalisierungsdaten – zumindest bis zum Park & Ride, wenn ein Wechsel in den ÖPNV stattfindet. Bevor nun die eigentlich interessierende Analyse eines Bewegungsprofils stattfinden kann, muss zuerst aus diesen sehr unterschiedlichen Ortsdaten eine möglichst genaue und mit wenigen Fehlern behaftete Variable erstellt werden, die eine solche Analyse zulässt. Dabei müssen dann auch widersprüchliche Informationen plausibel ausgefiltert werden, wenn beispielsweise das Auto die Information eines Ortes liefert, während sich das Mobiltelefon weiterbewegt, um hieraus zu schließen, dass der Fahrer in ein anderes Verkehrsmittel umgestiegen und nicht an zwei Orten

gleichzeitig ist. Oder es muss eine Fallbereinigung stattfinden, weil z. B. davon ausgegangen wird, dass das Auto oder das Mobiltelefon an eine andere Person übergeben wurde.

Auch das Nichtvorhandensein von Daten muss in diesem Rahmen berücksichtigt werden: So kann ein fehlender Eintrag beispielsweise von den GPS-Daten (Kandt, Kapitel 119 in diesem Band) eines Mobiltelefons darauf hinweisen, dass ein Gebäude betreten wurde, innerhalb dessen GPS typischerweise nicht mehr funktioniert. In diesem Fall wäre darauf zu schließen, dass die Person an dem Standort verblieben ist – dann handelt es sich um ein sinnvolles Datum zum Standort. Das Ausbleiben einer GPS-Positionierung kann aber auch bedeuten, dass das Telefon deaktiviert wurde und trotz eines Standortwechsels nun keine Daten liefert – in diesem Fall wäre das Ausbleiben von Daten mit fehlenden Werten gleichzusetzen.

Das Auffinden solcher sich widersprechender oder falscher Einträge ist aufgrund der Masse an Daten bei Big Data manuell nicht möglich. Eine Plausibilitätsprüfung wird zunehmend automatisiert erfolgen müssen, um bestimmte Muster als (potentielle) Fehler zu identifizieren. Je nach Umfang mag es dann noch möglich sein, die so identifizierten Fälle einzeln zu prüfen, je komplexer und umfangreicher die Daten aber sind, umso schwieriger wird sich dies umsetzen lassen.

10.7 Analyseverfahren

Während die Ergebnisse von Big-Data-Analysen in Bezug auf beispielsweise automatisierte Autos, Spracherkennung oder personalisierte Werbung längst im Alltag angekommen sind, ist den meisten Nutzern unklar, mit welchen Verfahren diese ausgewertet werden und wie solche Prognosen und Mustererkennungen überhaupt funktionieren.

Wenn konventionelle Datensätze um relativ wenige Fragmente aus einer Big-Data-Anwendung ergänzt werden, kann mit den klassischen qualitativen oder quantitativen Analysetechniken der empirischen Sozialforschung gearbeitet werden.

Oft ist es im Big-Data-Kontext aber gerade interessant, viele einzelne Variablen, die aus sehr unterschiedlichen Quellen kommen und, wie beschrieben, einzeln oft nur einen geringen Informationsgehalt haben, zu informationsdichteren Variablen zu kombinieren (Cielebak/Rässler, Kapitel 31 in diesem Band). Dabei bietet es sich an, automatisiert mit Hilfe eines bestimmten Algorithmus nach Strukturen in den Daten zu suchen (Manderscheid, Kapitel 121 in diesem Band). Aufgrund der Vielfältigkeit der Daten bedarf es je nach Forschungsinteresse und Datenformat komplexer Analyseinstrumente, die sehr unterschiedlich sein können und teilweise sogar speziell für eine bestimmte Analyse programmiert sind. Für praktische Beispiele siehe Foster et al. (2017) und McLevey (2022).

Sei es, weil sehr viele Variablen vorliegen, die auf sehr unterschiedliche Art miteinander kombiniert werden können, sei es, weil es sich um sehr aufwändige Regeln

bei der Textanalyse handelt: Je komplexer der Datensatz wird, desto komplizierter kann es sein, einen Algorithmus zu programmieren, der eine automatisierte Analyse zulässt. Abhilfe können hier Verfahren des maschinellen Lernens schaffen, die ein computergestütztes Anlegen eben solcher Auswertungsalgorithmen erlauben, wobei zwischen *überwachtem* und *unüberwachtem Lernen* (supervised und unsupervised learning) (Ghani/Schierholz 2017) unterschieden wird.

- Beim überwachten Lernen werden z. B. Spam-Filter, Vorhersagen über Kreditwürdigkeit oder Krankheiten, Gesichtserkennung oder minutengenaue Stauvorhersagen auf Navigationsgeräten betrachtet. Algorithmen lernen Vorhersagen zu machen, indem ihnen anfangs gezeigt wird, was beispielsweise eine Spam-E-Mail ist und was nicht. Durch Lernen ist der Algorithmus letztendlich in der Lage, selbsttätig Vorhersagen zu machen. Welche Kriterien für die Vorhersage herangezogen werden, ist ab einem gewissen Punkt jedoch nicht mehr nachvollziehbar. Dabei stellt sich aber auch schon vorab die Frage, wer nach welchen Kriterien den kurierten Lerndatensatz designt hat, weil davon die Ergebnisse abhängen, die der angelernte Algorithmus produzieren wird. So ist es denkbar, dass hierbei – beabsichtigt oder unbeabsichtigt – bestimme Formen von Ergebnissen produziert werden, nicht weil sie wahr sind, sondern weil sie durch den zum Lernen verwendeten Datensatz als wahr erscheinen – was dann die Quelle eines massiven Bias sein kann.
- Im Gegensatz zum überwachten Lernen, hat das unüberwachte Lernen das Ziel Strukturen innerhalb der Daten zu entdecken. Hierzu wird nicht vorab eine Auswahl an Variablen aus theoriegeleiteten und sachlogischen Gründen festgelegt, auf welche das Verfahren angewendet wird, sondern mithilfe aller Variablen nach Mustern gesucht. Erst im Nachhinein erfolgt die Bewertung gefundener Zusammenhänge. Werbeempfehlungen beispielsweise basieren typischerweise auf der Gruppierung von Gewohnheiten und Vorlieben, um Personen mit ähnlichem Verhalten und Präferenzen personalisierte Werbung anzuzeigen. Für die Wissenschaft wäre die Analyse von Einstellungen und Verhalten bestimmter Milieus und Lebensstile interessant. Das Pedant zum unüberwachten Lernen sind in der konventionellen empirischen Sozialwissenschaft die Clusteranalyse, die Korrespondenzanalyse oder die Latente Klassenanalyse (Blasius/Baur, Kapitel 45 in diesem Band) als Verfahren zur Entdeckung von Ähnlichkeitsstrukturen.

10.8 Forschungsethik

Wie wir gesehen haben, hat Big Data eine Reihe begehrenswerter Eigenschaften, die für sozialwissenschaftliche Analysen sehr vielversprechend erscheinen. Bei genauerer Betrachtung ergeben sich hier aber mehrere große Probleme: So lässt die bisherige Darstellung Zweifel daran aufkommen, wie viele dieser Versprechen bei genauerem Hinsehen tatsächlich eingehalten werden können, während gleichzeitig genau die Eigenschaften, die Big Data so begehrenswert machen, erhebliche ethische Probleme implizieren (Zwitter 2014, siehe auch Friedrichs, Kapitel 21 in diesem Band).

Den meisten Personen dürfte zwar klar sein, dass personenbezogene Daten anfallen, aber nicht unbedingt, welche es tatsächlich sind und was sich aus diesen gerechtfertigt oder ungerechtfertigt ableiten lässt. Letzteres gilt dann insbesondere auch für nutzergenerierte Daten, bei denen oft vergessen wird, dass selbst ohne weitere Freigabe zumindest der Betreiber eines Datendienstes diese einsehen und analysieren kann. Die Existenz von vermeintlich objektiven Daten über sehr viele Individuen wird also letztlich über den Verlust informationeller Kontrolle erkauft (Mühlichen, Kapitel 23 in diesem Band). Aus Sicht einer datenverarbeitenden Instanz mag sich dieses Problem oberflächlich durch eine Form von Zustimmung, die dem Nutzer abverlangt wird, lösen lassen. Ob eine solche Konstruktionen aber zuverlässig funktionieren oder eine realistische Möglichkeit zur Kontrolle in der massiv asymmetrischen Beziehung zwischen einzelnem Nutzer und Datensammler bedeutet, darf bezweifelt werden – zumal dem Nutzer oft gar keine Wahl zwischen Verzicht auf Nutzung oft notwendiger Dienste oder der vollumfänglichen Zustimmungen aller vorgegebenen Nutzungsbedingungen eines Dienstes bleibt. Dies mag problematischer sein, wenn die auswertende Institution eine der Überwachung oder eine wirtschaftlichen Prinzipien verpflichtete ist und nicht so sehr, wenn es um wissenschaftliche Forschung geht. Es stellt sich aber selbst dann die Frage, ob sich sozialwissenschaftliche Forschung bei Big Data bedienen kann, ohne gleichzeitig diejenigen (und damit auch deren fragwürdigen Intentionen) zu legitimieren, die solche Daten ursprünglich erzeugt haben (vergleiche zum Beispiel die Diskussion bei Zimmer 2010).

Auch stellt sich die Frage, welche Folgen eine probabilistisch gewonnene Erkenntnis aus einer Big-Data-Analyse haben darf. Wenn eine staatliche Überwachungsbehörde unerwünschtes Verhalten oder eine Kreditauskunftei die Kreditwürdigkeit mit Hilfe von Big Data prognostizieren möchte, dann sind deren Modelle vielleicht gut genug, um eine probabilistische Tendenz für bestimmte Gruppen erkennbar zu machen, sie sind aber sehr wahrscheinlich nicht dazu geeignet, um (dann deterministische) Rückschlüsse über spezifische Individuen zuzulassen – und so beispielsweise jemanden als nicht kreditwürdig zu führen.

Spätestens bei automatisierten Echtzeituntersuchungen riesiger Datenmengen dürften Methoden verwendet werden, die nicht den Ansprüchen genügen, wie sie beispielsweise medizinische, normierte psychometrische Testverfahren einsetzen, die – immer noch mit einer Fehlerwahrscheinlichkeit – Rückschlüsse über Indivi-

duen zulassen, aber andererseits genau für eine solche Zuordnung verwendet werden. Eine solche Big-Data-Echtzeitanalyse käme beispielsweise zum Einsatz, um beim Besuch einer Webseite zu versuchen, den Nutzer samt dessen Präferenzen, Einstellungen etc. zu identifizieren, um so die Erfahrung zu individualisieren, indem u. a. Werbung und Informationsblasen – also eine Auswahl der für dieses Individuum als interessant vermutete Informationen – auf diesen angepasst werden. Das ethische Dilemma mag bei schlecht personalisierter Werbung geringfügig sein, nimmt aber spätestens dann deutlich zu, wenn z. B. Suchergebnisse nach nicht nachvollziehbaren und gegebenenfalls auch noch falschen Kriterien auf Individuen angepasst werden und so eine verzerrte Vermittlung von Realität droht. Noch größer wird das Dilemma, wenn Big Data in der Forschung eingesetzt wird, um hier auch komplexere Konstrukte messbar zu machen, wie die Neigung zu psychischen Erkrankungen. Allein das Verschweigen des Ergebnisses gegenüber dem gemessenen Individuum, das vielleicht nicht einmal von seinem Status als Proband weiß, hat dabei bereits eine ethisch problematische Dimension (vgl. z. B. De Choudhury et al. 2014).

10.9 Schlussbemerkung

Die Aufbruchsstimmung, die in Bezug auf Big Data auch die Sozialwissenschaften erreicht hat, erscheint angesichts der vielen methodischen Probleme, die bei der tatsächlichen Durchführung zu erwarten sind, auf den ersten Blick vielleicht als zu optimistisch. Tatsächlich liegt es aber nahe, die Situation mit der Zeit der Anfänge von Online-Umfragen zu vergleichen. Auch dort gab es auf der einen Seite einen initialen Hype, der gepaart mit unrealistischen Erwartungen und methodischer Unwissenheit zu zweifelhaften Ergebnissen geführt hat. Erst eine sorgfältige Aufarbeitung der Grundlagen, also insbesondere auch der methodischen Probleme und Implikationen der neuen Verfahren, hat dann zur Etablierung eines heute selbstverständlich erscheinenden Erhebungsmodus geführt. Die Sozialwissenschaften stehen dabei weniger vor dem Problem, durch Big Data ersetzt zu werden, als methodologische und ethische Fragen angehen zu müssen. Aus methodischer Perspektive erscheint es besonders wichtig, eine Systematisierung wissenschaftlicher Standards im Bereich Nonresponse, Selektivität, Fallzuordnung, Messfehler, Replikation und Transparenz anzustreben.

Neben dem Methodischen erscheint es wichtig, dass sich die Sozialwissenschaft beim Umgang mit Big Data nicht von der Masse der Daten blenden lässt. Big-Data-gestützte Analysen verbleiben bis dato meist auf der Ebene deskriptiver Beschreibungen von Zusammenhängen – einer Feststellung dessen was passiert –, das Warum bleibt dagegen meist offen. Diese eher explorative Vorgehensweise erscheint auf sich beschränkt eher als Vorstufe zur Hypothesen- und Theoriebildung geeignet, eine stärkere Einbindung soziologischer Theorien verspricht aber Big Data zunehmend mit Inhalten zu füllen und auch kausale Zusammenhänge zu erklären.

Literatur

Baur, Nina (2009): Measurement and Selection Bias in Longitudinal Data. A Framework for Re-Opening the Discussion on Data Quality and Generalizability of Social Bookkeeping Data. In: Historical Social Research 34 (3): 9–50

Bick, Wolfgang/Müller, Paul J. (1984): Sozialwissenschaftliche Datenkunde für prozessproduzierte Daten: Entstehungsbedingungen und Indikatorenqualität. In: Bick, Wolfgang/Mann, Reinhard/Müller, Paul J. (Hg.): Sozialforschung und Verwaltungsdaten. Stuttgart: Klett-Cotta, 123–159

Bryson, Maurice C. (1976): The Literary Digest Poll: Making of a Statistical Myth. In: The American Statistician 30: 184–185

De Choudhury, Munmun/Counts, Scott/Horvitz, Eric/Hoff, Aaron (2014): Characterizing and Predicting Postpartum Depression from Shared Facebook Data. In: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing. Baltimore, MD: ACM Press: 626–638. URL: https://doi.org/10.1145/2531602.2531675

Foster, Ian/Ghani, Rayid/Jarmin, Ron S./Kreuter, Frauke/Lane, Julia (Hg.) (2017): Big Data and Social Science: A Practical Guide to Methods and Tools. Boca Raton, FL: CRC Press

Ghani, Rayid/Schierholz, Malte (2017): Machine Learning. In: Foster, Ian/Ghani, Rayid/Jarmin, Ron S./Kreuter, Frauke/Lane, Julia (Hg.) (2017): Big Data and Social Science: A Practical Guide to Methods and Tools. Boca Raton, FL: CRC Press, 147–186

Heiberger, Raphael/Riebling, Jan (2016): Installing Computational Social Science: Facing the Challenges of new Information and Communication Technologies in Social Science. In: Methodological Innovations 9: 1–11

Klochikhin, Evgeny/Boyd-Graber, Jordan (2017): Text Analysis. In: Foster, Ian/Ghani, Rayid/Jarmin, Ron S./Kreuter, Frauke/Lane, Julia (Hg.) (2017): Big Data and Social Science: A Practical Guide to Methods and Tools. Boca Raton, FL: CRC Press, 187–214

Laney, Douglas (2001): 3-D Data Management: Controlling Data Volume, Velocity, and Variety. In: META Group Research Note. URL: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lazer, David/Kennedy, Ryan/King, Gary/Vespignani, Alessandro (2014): The Parable of Google Flu: Traps in Big Data Analysis. In: Science 343: 1203–1205

Mayerl, Jochen (2015): Bedeutet Big Data das Ende der sozialwissenschaftlichen Methodenforschung? Soziopolis. URL: https://soziopolis.de/beobachten/wissenschaft/artikel/bedeutet-big-data-das-ende-der-sozialwissenschaftlichen-methodenforschung/; zuletzt zugegriffen am 21. 9. 2017

McLevey, John (2022): Doing Computational Social Science. A Practical Introduction. Sage Publications Ltd.

Morstatter, Fred/Pfeffer, Jürgen/Liu, Huan (2014): When Is It Biased? Assessing the Representativeness of Twitter's Streaming API. In: Proceedings of the 23rd International Conference on World Wide Web. New York, NY, USA: ACM: 555–556. URL: https://doi.org/10.1145/2567948.2576952

Neylon, Cameron (2017): Working with Web Data and APIs. In: Foster, Ian/Ghani, Rayid/Jarmin, Ron S./Kreuter, Frauke/Lane, Julia (Hg.) (2017): Big Data and Social Science: A Practical Guide to Methods and Tools. Boca Raton, FL: CRC Press, 23–70

Schmitz, Andreas/Skopek, Jan/Schulz, Florian/Blossfeld, Hans-Peter (2009): Indicating Mate Preferences by Mixing Survey and Process-generated Data. The Case of Attitudes and Behaviour in Online Mate Search. In: Historical Social Research 34 (1): 77–93

Zimmer, Michael (2010): „But the Data is already Public": On the Ethics of Research in Facebook. In: Ethics and Information Technology 12 (4): 313–325

Zwitter, Andrej (2014): Big Data ethics. In: Big Data & Society 1 (2): 1–6

Miriam Trübner (geb. Schütte) ist wissenschaftliche Mitarbeiterin in der Abteilung für Soziologie und quantitative Methoden an der Johannes Gutenberg-Universität Mainz. *Ausgewählte Publikationen:* The Dynamics of „neither agree nor disagree" in Attitudinal Questions, in: Journal of Survey Statistics and Methodology (2021); Effect of Header Images on Different Devices in Web Surveys, in: Survey Research Methods (2020); Arbeitsteilung in Paarhaushalten. Eine dyadische Untersuchung partnerschaftlicher Aufgabenverteilung. Opladen: Barbara Budrich (2020); *Webseite:* https://quantitative-methoden.soziologie.uni-mainz.de/dr-miriam-truebner/. *Kontaktadresse:* truebner@uni-mainz.de.

Andreas Mühlichen arbeitet am C³RDM als wissenschaftlicher Mitarbeiter der Universitäts- und Stadtbibliothek der Universität zu Köln. *Ausgewählte Publikationen:* Privatheit im Zeitalter vernetzter Systeme. Eine empirische Studie. Opladen: Barbara Budrich (2018). *Webseite:* https://fdm.uni-koeln.de/c3rdm-team. *Kontaktadresse:* muehlichen@ub.uni-koeln.de.