Big Data

Praktiken und Theorien der Datenverarbeitung im historischen Querschnitt

Cornelius Borck

Big Data ist ein schwacher Begriff. Vage umreißt er, was sich als neues Verhältnis zu großen Datenmengen erst abzuzeichnen scheint. Big Data meint einerseits das schiere Volumen von digitalen Informationen und verweist andererseits auf neue Formen des Umgangs mit ihnen. Lange Zeit galt das Telefonbuch als Inbegriff einer ungeheuren Datensammlung. Heutzutage wird es vielerorts gar nicht mehr gedruckt (geschweige denn benutzt), weil neue Methoden des Informationszugriffs sich als überlegen herausgestellt haben und elektronische Speicher weit mehr bereithalten, als sich im begrenzten Platz von Buchseiten unterbringen lässt. Wenn Big Data ein Phänomen der Gegenwart ist, dann weil heute täglich Datenmengen generiert und bewegt werden, die noch vor kurzem unvorstellbar waren, und weil diese Datenteppiche neue Methoden der Analyse und des Vergleichs hervortreiben, mit denen aus der Häufigkeit und Ähnlichkeit bestimmter Merkmale, Verteilungen und Muster auf Informationen geschlossen werden soll, für die es noch gar keine Anhaltspunkte, Vermutungen oder Hypothesen gab. Amazon empfiehlt dank der implementierten Algorithmen Bücher, nach denen wir gar nicht hätten suchen können, weil wir weder von ihnen wussten, noch davon, dass sie uns interessieren.

Die Phänomene sind weitgehend bekannt: Forschungszentren wie das CERN oder Konsortien wie das IPCC der Klimaforschung und das Human Brain Project generieren und bearbeiten riesige Datenmengen (zu deren Verarbeitung sie weltweit vernetzte Rechner benötigen), weil sich erst aus der Menge der Daten dem Forschungsgegenstand angemessene Simulationen und Modellierungen gewinnen lassen. Sogar wenn es nur um einen einzelnen menschlichen Körper geht, hantieren Genetiker und Systembiologen inzwischen mit gigantischen Datensets, denn auch hier geht es inzwischen um die Frage, wie sich das Zusammenspiel sehr vieler Faktoren in komplexen Systemen realitätsnah modellieren lässt, anstatt wie in der konventionellen Forschung abstrahierende Ursache-Wirkungsbeziehungen

herauszupräparieren. Während bei diesen Beispielen noch der fachspezifische Zugriff auf den Forschungsgegenstand die leitende Perspektive vorgibt (auch wenn dazu große interdisziplinäre Forschungsverbünde erforderlich sind), gehen andere Anwendungsfelder einen radikaleren Weg. Google baute sein Erkältungsfrühwarnsystem *Flu Trends* ohne besonderes medizinisches Wissen aus dem Suchverhalten seiner Nutzer auf und konnte damit sogar eine Veröffentlichung in *Nature* landen (Ginsberg et al. 2009).¹ Das Projekt exemplifizierte so den radikalen Anspruch einiger Big Data-Vordenker, allein aus digitalen Daten neues Wissen zu generieren: Bereits vor knapp zehn Jahren hatte Chris Anderson „das Ende der Theorie“ ausgerufen, weil inzwischen genügend Speicherplatz bereitstünde, um die ganze Welt unmittelbar abzubilden, und „with enough data, the numbers speak for themselves.“² In einer sich nicht zuletzt aufgrund der Informationstechnologien rasant verändernden Welt bleibt solche Vollständigkeit freilich ein Phantasma, das sich offenbar am ehesten noch als Überwachungsstaat oder als Monopol von Internetkonzernen konkretisiert. Seit Edward Snowden ahnt die Öffentlichkeit etwas von der Sammelwut der Geheimdienste und den Fangstricken, die sie mit ihren modernsten Methoden auslegen.³

Das Erheben und Bearbeiten gigantischer Datenmengen, die noch vor kurzem unvorstellbar waren, ist in vielen Wissenschaftszweigen zur Selbstverständlichkeit geworden. Zu den Einsatzfeldern von Big Data gehören Teilchenphysik, genetische Sequenzen oder Klimadaten ebenso wie die Textcorpora der Literaturwissenschaften, historische Zeitschriften- und Buchbestände oder digitalisierte Archivalien. Längst sind ganze Bibliotheksverbünde digitalisiert worden. Zur Vision einer digitalen, dematerialisierten Welt gehört dabei der Traum einer Befreiung von Raum und Zeit.⁴ Denn für digitale Recherchen sollen weder die räumlichen Grenzen von Bibliothek, Stadt und Land, noch zeitliche – von so banalen wie Öffnungszeiten bis hin zu den tief in die Wissenschaftsinstitutionen eingeschriebenen – Grenzen von Epochen und Periodisierungen gelten. An die Stelle von Disziplinargrenzen, Tradition und Kanon treten damit Fragen danach, was überhaupt Wissen ausmacht und welche Daten gesammelt werden sollten.⁵ Selbst fachliche Trennungen wie die zwischen Disziplinen werden porös mit Datenbanken und Plattformen wie *Google Scholar*, Verlagsportalen oder JSTOR. Stattdessen treten umso deutlicher ökonomische und juristische Zugangsschranken hervor, weshalb Diskussionen über Zugangsrechte, Eigentumsfragen und Open Access umso dringlicher werden.

Die wachsenden digitalen Datenbanken und Archive eröffnen auch der Medizin-, Wissenschafts- und Technikgeschichte neue Perspektiven – und zwar nicht nur in Bezug auf eigene Forschungspraktiken, sondern auch auf ihren Gegenstand. In diesem Themenheft soll es um letzteres gehen. Big

Data eröffnet hier gleich mehrere Fragen: Sind große Datensammlungen tatsächlich so neu, was unterscheidet sie von vorausgegangenen Großprojekten und welche Parallelen lassen sich ziehen, mit denen das Neue genauere Konturen bekommt? Welchen Anteil hat die digitale Verschalt- und Verfügbarkeit der Daten an der behaupteten neuen Qualität von Big Data, und welche Effekte zeitigen diese Möglichkeiten in der Praxis? Diese Fragen sind der Ausgangspunkt des vorliegenden Themenheftes, das mit seinen fünf Fallstudien methodische und historische Reflexionspotenziale mobilisieren will, um sowohl die angebliche Neuartigkeit großer Datensammlungen, als auch das Versprechen einer damit möglich gewordenen unmittelbaren Wirklichkeitsabbildung kritisch zu hinterfragen.

Die Geschichte des Ausdrucks „Big Data“ verläuft sich im Dunkel der jüngsten Zeitgeschichte, denn einerseits ist die Kombination der beiden Wörter zu unspezifisch und andererseits ist noch nicht hinreichend klar, was genau mit Big Data gemeint sein soll und ob das mit diesem Namen jetzt Bezeichnete sich einmal als ein stabiles Phänomen herausstellen wird. Zugespitzt könnte man formulieren, dass der neue Ausdruck einer angemessenen Historisierung von Datenpraktiken geradezu im Wege steht, weil er mit seiner Neuheit Parallelen zu älteren Strategien und vorausgegangenen Datensammlungsprojekten verdeckt. Seit gut zehn Jahren ist der Ausdruck vor allem in Gebrauch, um eine neue Stufe, eine neue Qualität der informationellen Wirklichkeit anzuzeigen. „Big Data is a Big Deal“ verkündete die Obama-Regierung vor fünf Jahren zum Start ihrer *Big Data Development Initiative*:

By improving our ability to extract knowledge and insights from large and complex collections of digital data, the initiative promises to help accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning.⁶

Einmal mehr wurde hier behauptet, der wissenschaftlich-technische Fortschritt läge fortan in einer Extraktion des in den Daten schon gespeicherten Wissens – woraus dann bezeichnender Weise im selben Atemzug ein Zugewinn an nationaler Sicherheit und eine bevorstehende Reform von Lehren und Lernen abgeleitet wurde. Vielleicht hat das seinem Selbstverständnis nach freieste Land der Erde tatsächlich weniger Anlass zur Sorge um ein Aushöhlen der Freiheitsräume durch Video- und Telefonüberwachung. Aber was soll es heißen, „teaching and learning“ an Big Data anzupassen, wenn dessen Versprechen doch gerade darin liegen soll, ohne große Vorkenntnisse Neues zu erkennen? Hatten die etablierten, traditionellen Techniken der Wissensverwaltung und der Lehre nicht bewusst auf Komplexitätsreduktion durch Methodentraining und Lehrbeispiele gesetzt,

Kanonbildung angestrebt und Theorie bis hin zum Ideal einer möglichst einfachen Formel zum Zentrum der Epistemologie erhoben?

Welche Formen des Wissens von Big Data stabilisiert werden und was das für das Verständnis von Wissenschaft und Unterricht heißen wird, lässt sich noch nicht abschätzen. Aber hier scheint eine irritierende Parallele auf, denn auch die Wissenschafts- und Technikgeschichte hat ja solche für die Wissenschaften bisher charakteristischen Verfahren der Komplexitätsreduktion in den vergangenen Jahren und Jahrzehnten kritisch hinterfragt. – Damit wird deutlich, welche Herausforderung Big Data für unser Fächerspektrum darstellt, denn man wird Big Data mit seinem weitgehend geschichtsblinden, aber gleichwohl epochemachenden Anspruch kaum einen natürlichen Partner nennen wollen. Ganz im Gegenteil: Wo Big Data mit immer genaueren Daten der Wirklichkeit habhaft zu werden vermeint, hat die Wissenschafts- und Technikgeschichte mit ihren diskursanalytischen und sozialkonstruktivistischen Ansätzen, mit dem *material turn* und der Fokussierung auf Praktiken immer wieder die komplexen materiellen wie sozialen Prozesse herausgearbeitet, die als methodische Zurichtung und disziplinäre Filterung wissenschaftliche Phänomene überhaupt erst stabilisieren. Daten sind buchstäblich *facta*, nämlich Gemachtes. Das umreißt das Potenzial der Wissenschafts- und Technikgeschichte auch für die Analyse von Big Data: Auf welche technischen Infrastrukturen und medialen Voraussetzungen bauen die aktuellen Datenpraktiken auf? Welche Wissensformen werden priorisiert, und welche Mechanismen der Selektion sind an welchen Stellen wirksam? Wie greifen technische Möglichkeiten (bzw. pragmatische Grenzen) und soziopolitische Kräfteverhältnisse ineinander? Welche Aushandlungsprozesse sind als Vorentscheidungen bereits in die Verfahren implementiert? Diese Leitfragen der Wissenschafts-, Medizin- und Technikgeschichte gelten auch für die Auseinandersetzung mit Big Data.

Stellen große Datenmengen und neue Techniken der Datenverarbeitung eine Zäsur dar? Vom Fachzuschnitt her wird die Wissenschafts-, Medizin- und Technikgeschichte dem vorschnellen Ausrufen einer Zeitenwende oder eines epistemischen Bruchs allein schon deswegen skeptisch gegenüberstehen, weil es in den verschiedensten Wissenschaften und Praxisfeldern schon seit langer Zeit sehr große Datensammlungen gab. Die Klage über die sogenannte Informationsexplosion ist vermutlich so alt, wie es überhaupt Speichermedien gibt. Die gebotene historische Skepsis bliebe allerdings solange wohlfeil, wie sie das spezifische Potenzial wissenschafts- und technikhistorischer Reflexion ungenutzt ließe. Deshalb versammelt das Themenheft historische und aktuelle Fallstudien, die am Einzelfall die Verfahren, Probleme und Strategien im Umgang mit großen Datenmengen beobachten und analysieren. Dazu werden in spezifischen Konstellationen

und lokalen Kontexten die konkreten Praktiken herausgearbeitet und die historischen und systematischen Tiefendimensionen freigelegt. Dabei sind die Fallstudien gezielt gestreut hinsichtlich des Untersuchungszeitraums und wissenschaftlichen Feldes. Sie reichen von der Entwicklung der preußischen Bevölkerungsstatistik und dem Beginn der systematischen Wetterbeobachtung in der Schweiz bis zur rezenten genetischen Forschung, den Simulationsalgorithmen am CERN und der Entwicklung relationaler Datenbanken als neuem Paradigma.

Christine von Oertzen analysiert die materielle Kultur der Datenverarbeitung im Preußischen Zensus als einen Prozess der „Verdatung“ und legt dazu die verschiedenen Schichten, Massen, Methoden und Werkzeuge frei, die überhaupt erst die Kompilation der Aggregate ermöglichten – und dabei buchstäblich ganz Berlin durchwirkten. Franziska Hupfer zeichnet die nationalstaatliche Institutionalisierung der Wetterbeobachtung nach und rekonstruiert dabei, welchen vielfältigen lokalen wie epistemischen Transformationen dieses Projekt immer wieder unterworfen war. Gabriele Gramelsberger beschreibt das Zusammenspiel von DNA-Sequenziermaschinen und softwarebasierten Analyseprogrammen und analysiert dabei, wie die Validierungsstrategien für die ermittelten genetischen und epigenetischen Informationen inzwischen bereits in die Software für die Datenverarbeitung implementiert werden, so dass Aussagen über den epistemischen Status dieser Daten immer problematischer werden. Anne Dippel beobachtet, wie Spielen im Wortsinne zur Big-Data-basierten Wissensproduktion in der Hochenergiephysik beiträgt. In ihrer dichten Beschreibung der Arbeitsorganisation und Forschungspraktiken in Genf verschränkt sie die symbolische Dimension der Monte-Carlo-Simulationen mit den ontologischen Implikationen der kompetitiven Arbeitspraxis und den epistemischen Effekten der auf Online-Plattformen realisierten Algorithmen-Entwicklung. Ein eher biographischer Beitrag, in dem Nils Hanwahr Jim Grays These vom „Fourth Paradigm“ einer datenbankgestützten Wissensgenerierung kontextualisiert, rundet das Themenheft ab.

Die Beiträge des Themenheftes gehen auf die 99. Jahrestagung der Deutschen Gesellschaft für Geschichte der Medizin, Naturwissenschaften und Technik im Herbst 2016 in Lübeck unter dem Titel „Digitalisierung, Big Data und die Aufgabe der Theorie“ zurück, ohne dort bereits in einem besonderen Zusammenhang gestanden zu haben. Ich danke den Autor_innen für ihre Mitwirkung an diesem Themenheft und die sorgfältige Arbeit an den Beiträgen. Der NTM-Redaktion und den Herausgeber_innen danke ich für die Möglichkeit, das Thema der Tagung in diesem Themenheft aufzugreifen sowie für die tatkräftige Unterstützung dabei.

Mit diesen fünf Beiträgen lässt sich nicht bestimmen, was Big Data eigentlich ausmacht,⁷ und noch weniger liefern sie Hinweise für die

These, dass aus den großen Datenmengen dank innovativer Algorithmen und Modellierungen neue Wirklichkeitswissenschaften entstünden. Stattdessen zeigen sie auf, wie Verfahren der Datengewinnung, Datensammlung, Speicherung, Analyse und Verarbeitung immer schon und immer noch in ein vielschichtiges Netz epistemischer Vorannahmen, technischer Möglichkeitsbedingungen und materieller Limitierungen eingewoben waren bzw. sind. Diese Absage an eine Zeitenwende ist keine Bestätigung einer *déformation professionnelle* der Historiker, es gäbe nichts Neues unter der Sonne, sondern verweist darauf, dass in den aktuellen Debatten über Big Data die kritische Diskussion dieser so entscheidenden technischen Möglichkeitsbedingungen und praktischen Verfahrensweisen oft zu kurz kommt. Big Data mag das Versprechen auf eine Neue Zeit der Datenparadiese für neue Wirklichkeitswissenschaften sein, aber ohne solche Reflexionen werden daraus – das demonstrieren die historischen und aktuellen Fallstudien – allzu leicht neue Wirklichkeitswissenschaften, die erst schaffen, was dann als Wirklichkeit gilt.

Anmerkungen

- 1 Die Vorhersagekraft von *Google Flu Trends* erwies sich später als so mangelhaft, dass der Dienst inzwischen eingestellt wurde. Zwar lässt sich nicht bestreiten, dass unerwartete Informationen in großen und komplexen Datensammlungen enthalten sind, aber oftmals liefern die vermeintlich vollständigen Daten keineswegs eine vollständige Abbildung, sondern eine zufällige und oft sogar noch durch die Sammlung selbst systematisch verzerrte Auswahl, vgl. Harford (2014).
- 2 Anderson (2008).
- 3 Das geht schnell über dystopische Diagnosen, wurde doch Big Data auch für den Wahlsieg Trumps verantwortlich gemacht, vgl. Grassegger und Krogerus (2016).
- 4 Wobei jedoch zumeist vergessen wird, welche Ressourcen zur Aufrechterhaltung der digitalen Räume erforderlich sind (vgl. Ensmeyer 2013).
- 5 Beispielsweise hat die Library of Congress inzwischen einen Vertrag mit Twitter geschlossen, um auch Tweets zu archivieren, weil hier an der Schwelle von Mündlich- zur Schriftlichkeit der Zeitgeist genauer verfolgt werden kann.
- 6 Kalil (2012).
- 7 Entsprechend kann auch diese Einführung keinen Überblick über den Forschungsstand geben. Für einen ersten Einstieg, vgl. Hacking (1982), Porter (1995), Leonelli (2014), Hounshell & Midena (2014), Reichert (2014).

Literatur

Chris Anderson 2008: The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, *Wired Magazine* 23.06.2008. URL: <https://www.wired.com/2008/06/pb-theory/> (09.10.2017).

- Ensmenger, Nathan 2013. Computation, Materiality, and the Global Environment. *IEEE Annals of the History of Computing* (35): 80.
- Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski und Larry Brilliant 2009. Detecting Influenza Epidemics Using Search Engine Query Data. *Nature* (457): 1012–1014.
- Grassegger, Hannes und Mikael Krogerus 2016. Ich habe nur gezeigt, dass es die Bombe gibt. *Das Magazin* 03.12.2016. URL: <https://www.dasmagazin.ch/2016/12/03/ich-habe-nur-gezeigt-dass-es-die-bombe-gibt/> (08.10.2017).
- Hacking, Ian 1982. Biopower and the Avalanche of Printed Numbers. *Humanities in Society* (5): 279–295.
- Harford, Tim 2014: Big Data. Are We Making a Big Mistake? *Financial Times* 28.03.2014. URL: <https://www.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0> (09.10.2017).
- Hounshell, Eric und Daniel Midena 2014. Historicizing Big Data. *MPIWG Preprint* 462.
- Kalil, Tom 2012. Big Data is a Big Deal. *White House* 29.03.2012. URL: <https://obamawhitehouse.archives.gov/blog/2012/03/29/big-data-big-deal> (08.10.2016).
- Leonelli, Sabina 2014. What Difference Does Quantity Make? On the Epistemology of Big Data in Biology. *Big Data & Society* (1): 1–11.
- Reichert, Ramón (Hg.) 2014. *Big Data. Analysen zum digitalen Wandel von Wissen, Macht und Ökonomie*. Bielefeld: transcript.
- Porter, Theodore M. 1995. *Trust in Numbers. The Pursuit of Objectivity in Science and Public Life*. Princeton: Princeton University Press.

Cornelius Borck
Institut für Medizingeschichte und Wissenschaftsforschung
Universität zu Lübeck
Königstraße 42
23552 Lübeck
Deutschland
borck@imgwf.uni-luebeck.de