Schwerpunkt Big Data

Herbert Hrachovec*

Schubladen und Wolkenfelder

Anmerkungen zur Bearbeitung großer Datenmengen

https://doi.org/10.1515/iwp-2018-0016

Zusammenfassung: „Big Data“ ist eine Kurzbezeichnung für algorithmisch gestützte Verfahren zur Verarbeitung großer Datenmengen. Der Erfolg dieses Vorgehens und – in höherem Maß – seine medienwirksame Präsentation hängt direkt mit dem von ihm als problematisch erkannten Paradigma der relationalen Datenbanktheorie zusammen. Darum ist einleitend ein Blick auf deren Grundzüge zu werfen. Nur so wird der Kontrast zum konkurrierenden Ansatz deutlich. Der entscheidende Punkt ist die überwältigende Masse an Information, die durch neue Technologien sowohl erzeugt als auch gesammelt wird. Ihrem Ansturm scheint die konventionelle Theorie nicht gewachsen zu sein. Diese Herausforderung wird in einem zweiten Abschnitt diskutiert. Es stellt sich heraus, dass die Datenflut, die eine mathematisch-statistische Behandlung nötig macht, eine geänderte Einstellung zu erkenntnistheoretischen Fragen provoziert. Ob man ihr folgen soll, ist abschließend zu fragen.

Deskriptoren: Algorithmus, Daten, Datenanalyse, Datenfeld, Datenstruktur, Informatik, Philosophie

Pigeonholes and banks of clouds. On processing large amounts of data

Abstract: „Big Data“ is a catchphrase referring to algorithmic procedures to process large amounts of data. It owes its success, and in particular its medial prominence, in part to its opposition to the established paradigm of relational database theory, whose main tenets, will be sketched in a first step. It is by contrasting „Big Data“ with this traditional procedure that the crucial contrast regarding the availability of an overwhelming mass of information can be drawn. Proponents of the former approach claim that a change of paradigm is called for because conventional theory is not equipped to handle these unprecedented challenges. The ensuing debate is presented in the second section of this paper. The opposing views can, as it turns out, be traced back to divergent epistemological attitudes. The contribution concludes with a discussion about the questionable move to establish the category „raw data“ as an underlying domain for an innovative theory of data processing.

Descriptors: Algorithm, Data, Data analysis, Data field, Data structure, Computer Science, Philosophy

Catégories et bancs de nuages. Commentaires sur le traitement de grandes quantités de données

Résumé: « Big Data » est le nom abrégé pour désigner les méthodes soutenues par des algorithmes pour le traitement de grandes quantités de données. Le succès de cette approche, et – dans une plus grande mesure encore – sa présence médiatique, est directement liée au paradigme de la théorie des bases de données relationnelles. Par conséquent, on doit d’abord jeter un coup d'œil à leurs caractéristiques de base. Ce n'est qu'alors que le contraste avec l'approche concurrente devient clair. Le point décisif est la masse écrasante des informations qui est à la fois générée par les nouvelles technologies et recueillie par elles. La théorie conventionnelle ne semble pas pouvoir faire face à son assaut. Ce défi sera discuté dans un deuxième chapitre. Il se trouve que le flot de données qui rend un traitement mathématique statistique nécessaire provoque une approche différente des questions épistémologiques. La question finale est : doit-on la suivre ?

Descripteurs: Algorithme, Données, Analyse de données, Champ de données, Structure de données, Informatique, Philosophie

Anmerkung: Dieser Artikel beruht auf einer früheren Fassung, vorgetragen auf dem Symposium „BIG DATA – Perspektiven kritischer Sozial- und Kulturwissenschaften“, veranstaltet im April 2017 an der Johannes Kepler Universität Linz (JKU) Linz/Österreich. Veranstalter: Institut für Philosophie & Wissenschaftstheorie und Kulturinstitut ebendort.

***Kontaktperson: ao. Univ.-Prof. i.R. Dr. Herbert Hrachovec,** Universität Wien, Institut für Philosophie, Universitätsstraße 7, 1010 Wien, Österreich, E-Mail: herbert.hrachovec@univie.ac.at

Schubladen

Rund um die ubiquitäre Notwendigkeit der Datenerfassung und -verwaltung hat sich eine komplexe Expertinnenkultur entwickelt, die in der Lage ist, den Ansprüchen des Verkehrs- und Finanzwesens, der Konsumgesellschaft, der Bürokratie und des Wissenschaftsbetriebs zu entsprechen. Diese Konstellation hat eine soziale und eine informationstechnische Seite. Der eine Teil besteht aus *grosso modo* standardisierten arbeitsteiligen Rollen, an deren beiden Außenpolen die Systemadministratorin und die Endbenutzerinnen stehen. Ohne ein entsprechend eingerichtetes und gewartetes Betriebssystem gibt es keine Datenbanken, ohne den gesellschaftlichen Bedarf keine Nachfrage nach deren Leistungen. Dazwischen liegt das Arbeitsfeld der Programmiererinnen, Designerinnen und Administratorinnen zweckorientierter Datenarchitekturen und Transaktionen. Das „Datenbank Management System“ (DMS) als zweiter Teil, besteht aus Vorkehrungen, die es erlauben, nach einem internen Schema gespeicherte Informationsbestände interaktiv, nach den Bedürfnissen der Benutzerinnen, in bereitgestellten Ansichten („views“) zu spezifizieren (Ein Klassiker zur Information über diese Zusammenhänge ist Date, C.J. (2003)).

Der theoretische Kern dieser Einrichtung ist eine Mathematisierung des intuitiven Begriffs einer Tabelle, die für bestimmte Attribute (z. B. Vornamen, Familiennamen und Geburtsdaten) einen Rahmen für mögliche Eintragungen bereitstellt. Eine „Relation“ im Sinn der Datenbanktheorie definiert die Anzahl und die zugehörigen Sachgebiete der Attribute und enthält in den einzelnen Zeilen des vorgegebenen Rasters Einzelwerte aus den betreffenden Bereichen, also etwa (Helga, Marek, 10.7.1990). Die Leistungsfähigkeit dieser Konstruktion beruht darauf, dass mittels einer „relationalen Algebra“ wirksame Such- und Transformationsoperationen auf den Tabelleninhalten durchgeführt werden können. Ohne deren Effektivität im Alltag ist unsere gegenwärtige Lebensform (zumindest in der nördlichen Hemisphäre) kaum vorstellbar.

Dieser Umstand ist darauf zurückzuführen, dass ein DMS einige Selbstverständlichkeiten unseres Umgangs mit der Welt mathematisch präzise nachbildet. Das geschieht in einer formalen Semantik, die auf der einen Seite mengentheoretisch konzipiert ist und auf der anderen Seite eine Brücke zum lebensweltlichen Gebrauch von Bedeutungen anbietet. So ist eine geordnete Aneinanderreihung wie das genannte Tupel (Helga, Marek, 10.7.1990) für sich genommen keine Information. Die Zeichen *per se* könnten sich auf Schildkröten und ihr Verkaufsdatum beziehen. Nur unter zusätzlichen Annahmen *bedeuten* sie etwas. Man spricht von intendierten Interpretationen im Hinblick auf lebenspraktisch erfasste und formal repräsentierbare Sachbereiche.

Formal werden in der Tabellendefinition die Inhaltsbezüge, auf welche die nominellen Einträge verweisen, modelltheoretisch festgelegt. Doch das Problem ist damit nicht gelöst. Dass „Helga“ ausgerechnet auf einen *Vornamen* verweisen soll, ist durch die formale Angabe einer Bezugsdomäne nicht garantiert. Mathematisch lässt sich aus abstrakt konzipierten Elementen ein semantisches Bezugsfeld definieren, das als Repräsentation des Begriffes „Vorname“ fungiert. Doch dabei handelt es sich um die symbolische Vertretung real existierender Vornamen, deren intendierte Interpretation nicht die formale Semantik, sondern der jeweilige Sprachgebrauch bestimmt. Andersherum gesagt: die Relationen einer Datenbank werden von Vorannahmen über die Welt geprägt, deren Strukturen sie präzisieren und informatisch handhabbar machen sollen. Die einzelnen Zeilen erhalten ihre intendierte Bedeutung aus dem Umgang mit realen Umständen. Der Zweck der Übung legt auch fest, an *welchen* Umgang dabei gedacht ist. Die Eintragungen sollen eine Konstellation der Welt wiedergeben; sie fungieren als Sätze, die das Bestehen von Sachverhalten behaupten. Helga Marek, so wird in unserem Beispiel festgehalten, wurde am 10.7.1990 geboren.

Eine griffige Auffassung der Welt als Korrelat von Behauptungssätzen hat Wittgenstein in seinem „Tractatus logico-philosophicus“ vorgelegt. „Im Satz wird gleichsam eine Sachlage probeweise zusammengestellt.“ (Wittgenstein, L. (1984), 4.031) Es lässt sich gleichsam durchspielen, welche Eintragungen in einem vorgegebenen Formular die Umstände korrekt wiedergeben. Diese Charakteristik spannt einen Bogen zwischen den Konstrukten einer formalen Syntax (wie den angedeuteten Tabellen) und dem Arbeitsziel wissenschaftlicher Anstrengungen, nämlich der Entwicklung zertifizierter Erkenntnisse. Wittgenstein gibt ein suggestives Bild zum Verhältnis der beiden Pole. Die probeweise Darstellung einer Welt im Satz gleicht der Zusammenstellung eines Automobilunglücks durch Puppen im Gerichtssaal (Wittgenstein, L. (1980), 94f.) Einprägsam ist damit angezeigt, wie sich wissenschaftliche Entwürfe unter entsprechenden Bedingungen bewähren können. Sätze sind Hypothesen, deren Verifikation zur Erkenntnis führt.

Wolkenfelder

Das Bild der Darstellung durch Puppen hat seine Schwächen. Es kann auch als ein Hinweis auf die Unzulänglichkeit der klassischen relationalen Datenbanktheorien verstanden werden. Wo steht geschrieben, dass sich die Welt

derart distinkt darstellt, dass sie in Form konsolidierter Tabellen rubriziert werden kann? Eine Massenkarambolage mit Explosion und Fahrzeugbränden ist kein Puppenspiel. Ohne bildhafte Hilfsmittel ausgedrückt: zur Erfassung der Umwelt eignet sich die prästabilisierte Harmonie nur beschränkt. An dieser Stelle setzt der Einspruch gegen herkömmliche Datenbanken und das ihnen zugrunde liegende Erkenntnisverfahren an, Modelltheorie sei das falsche Werkzeug:

> This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves. [Anderson, Chris (2008)]

Die Ordnungsvorstellungen der prädigitalen Ära sind, kurz gesagt, durch die gegenwärtigen Datenströme überfordert (zum unterstellten Paradigmenwechsel s. Kitchin, R. (2014); Sathi, A. (2012)).

Um ein Datenmodell entwerfen zu können, benötigt man eine Vorstellung davon, was im gewählten Weltausschnitt zu erwarten sei. Das sind, metaphorisch ausgedrückt, die angesprochenen Schubladen. Ihre Verwendung hat die Epistemologie und Wissenschaftstheorie traditionell vorgegeben. Sie fasst den Erkenntnisprozess als Bestätigung, respektive Verwerfung, tentativer Behauptungen über den *Stand* der Dinge auf. Der Gegenentwurf operiert mit Masse statt Klasse:

> …taking vast quantities of data – usually on the scale of millions, if not billions, of individual data points – and running algorithms that look for the connections between them on supercomputers. This is the essence of big data, … (Steadman, I. (2013))

Angesichts der allerorts verbreiteten Erhebung und Akkumulation von Daten aus den verschiedensten Lebensbereichen, die in inhomogenen Clustern aggregiert werden, greifen die Formulare der relationalen Datenbanken nicht. Sie erlauben, das ist der Einspruch, keine Abweichungen von der vorgesehenen Beschaffenheit des Erkenntnisgegenstandes. Besser geeignete Instrumente zur Erschließung der Welt, die mit eventuell auftretenden Überraschungen besser zurecht kommen, sind Algorithmen, die in der Überfülle zusammengetragener Informationen nach Mustern suchen:

> The big data approach to intelligence gathering allows an analyst to get the full resolution on worldwide affairs. … The algorithms find the patterns and the hypothesis follows from the data. The analyst doesn't even have to bother proposing a hypothesis any more. (Steadman, I. (2013))

Den Unterschied der beiden Zugangsweisen kann man mit einem Beispiel aus der Urlaubsplanung illustrieren. (Eine allgemeine Übersicht über Einsatzbereiche von Big Data Algorithmen gibt Emrouznejad, A. (2016)). Reisebuchungen folgen einer übersichtlichen Logik: die Fluglinien, Flugdaten und Preise stehen fest. Es geht nur darum, aus den Optionen zu wählen. In diesem Zusammenhang gibt es jedoch auch andere Fragen. Zu welchem Zeitpunkt wird der günstigste Preis angeboten? Mit welcher Wahrscheinlichkeit treten Verspätungen auf? Antworten darauf findet man nicht in tabellarischen Aufstellungen. Die relevanten Faktoren sind zu unübersichtlich, in inhomogenen Formaten erfasst und dynamisch fluktuierend. Um sich ein Bild von solchen Wahrscheinlichkeiten zu machen, bieten sich explorative Algorithmen an. Die Ergebnisse dieser Operationen sind statistische Wahrscheinlichkeiten. In der elementaren klassischen Logik ist „zu einem gewissen Prozentsatz wahr" nicht vorgesehen. Das schließt allerdings die Entwicklung von Logiken für unscharfe Zustände („fuzzy logic") nicht aus. Dementsprechend kann man mit zwei unterschiedlichen formalen Methoden an einen Datenbestand herangehen. Suchausdrücke der „Structured Query Language" (SQL) liefern algebraisch eindeutige Resultate für Abfragen über geeignet konfigurierte Datenbanken. Dagegen ergeben Algorithmen, angewandt auf einen Informationspool, Befunde über statistisch signifikante, auffallende Korrelationen (vgl. etwa Bryant, A. and Raja, U. (2014); Mayer-Schönberger, V. and Cukier, K. (2013)).

Ein Einwand gegen die skizzierte Gegenüberstellung drängt sich auf. Die angesprochenen Algorithmen sind doch ihrerseits theoretische Konstrukte. Auch sie enthalten bestimmte Vorgriffe auf die Beschaffenheit der Welt. Um nach dem Einfluss diverser Parameter auf die Preisgestaltung von Fluglinien zu fragen, bedarf es einer Konzeption von *Preis*. Oder, um auf das Beispiel von „Helga Marek" zurückzukommen, auch ein Algorithmus wird unterscheiden müssen, ob die Inskription „Helga" einen Vornamen, eine Firmenbezeichnung, eine Abkürzung oder den Ausdruck einer Fremdsprache repräsentiert. Selbst wenn das zutrifft, sollte es dennoch den kritischen Punkt nicht verwischen. Relationale Datenbanken stehen in der Tradition einer Erkenntnisanalyse, die im Wesentlichen auf einer initialen Trennung von Vernunftkategorien und Erfahrungsdaten aufbaut. Im Kognitionsprozess verbinden sich die beiden Seiten zum Erfahrungswissen. Der Blick auf ein klassisches Problem dieses Ansatzes ist hilfreich, um die Funktionalitäten der Modellierung und der Anwendung von Algorithmen gegeneinander abzugrenzen.

Daten sprechen für sich selbst?

Der klassische Erkenntnisbegriff beruht auf einem Dualismus zwischen Verstand und Sinnlichkeit. Diese beiden menschlichen Fähigkeiten sind aufeinander bezogen, doch sie erfüllen, nach seiner Konstruktion, ganz unterschiedliche Aufgaben. Ein Stichwort ist die Debatte zwischen Rationalismus und Empirismus. Kategorien stellen einen Begriffsrahmen bereit, der, für sich genommen, leer bleibt. Nur auf konkrete Inhalte bezogen, verhilft er zu Erfahrungswissen. Diese Inhalte kommen ihrerseits aus der sinnlichen Wahrnehmung, deren Impulse kategorial verarbeitet und diskursiv aufbereitet werden. Im Hinblick auf dieses Arrangement ergibt sich eine methodische Schlüsselfrage: *Wie* finden solche Impulse Eingang in den gegebenen Erkenntnisrahmen? Spezifischer formuliert: Sinneseindrücke sind für sich genommen ungeordnet und unbeherrschbar; die Verstandeserkenntnis verleiht ihnen eine Form. Worauf kann sich die Anwendung der Ordnungsprinzipien berufen, um sicherzustellen, dass sie keine willkürlichen Operationen durchführt? Worin besteht der sachliche Grund, der das Ergebnis ihrer Anwendung legitimiert? Schließlich lassen sich Sinnesreizungen auf unterschiedliche Weise verarbeiten. Leistet die Sinnlichkeit einen eigenständigen Beitrag zur Validität von Erfahrungswissen?

Wilfrid Sellars hat den Sachverhalt in der Mitte des 20. Jahrhunderts folgendermaßen auf den Punkt gebracht:

> The idea that observation ... is constituted by certain self-authenticating nonverbal episodes, the authority of which is transmitted to verbal and quasi-verbal performances when these performances are made „in conformity with the semantical rules of the language," is, of course, the heart of the Myth of the Given. [Sellars, W. (1997)]

Mit dieser Beschreibung der zentralen Annahme unmittelbarer Gegebenheiten hat Sellars den Finger auf einen wunden Punkt des Form-Inhalt-Schemas in dessen empiristischen Gebrauch gelegt. Die Verlässlichkeit der Wahrnehmung, die für eine sachliche Fundierung der Erkenntnis bürgt, beruht nach dem von ihm apostrophierten Mythos auf prä-verbalen Vorgängen („episodes") mit kognitiver Autorität. Hinter der von Sellars herangezogenen Fachterminologie steckt eine weit verbreitete Überzeugung. Nach ihr bilden Wahrnehmungen, die uns die Sinne bieten, den unbezweifelbaren Ausgangspunkt des Wissens. Über Schlussfolgerungen aus diesen Befunden lässt sich diskutieren, über ihr Vorliegen nicht. Die Gültigkeit der wissenschaftlich geprüften Erkenntnis verweist letztlich auf das unbezweifelbare Vorliegen der genannten Episoden. In einem plakativen Beispiel: „Hier befindet sich ein roter Fleck." Er ist „durch Sinneswahrnehmung gegeben". Sellars problematisiert den Übergang, der von dieser unterstellten „Gegebenheit" zu ihrer Weiterverwertung in höherstufigen kognitiven Kontexten führt.

Er charakterisiert den traditionellen Empirismus als eine Doktrin, die sich – als auf eine Letztinstanz – auf Entgegengenommenes („takings") stützt, dessen Bestehen nicht in Zweifel zieht.

> For the given, in epistemological tradition, is what is taken by these self-authenticating episodes.
>
> These ‚takings' are, so to speak, the unmoved movers of empirical knowledge, the ‚knowings in presence' which are presupposed by all other knowledge, both the knowledge of general truths and the knowledge ‚in absence' of other particular matters of fact. Such is the framework in which traditional empiricism makes its characteristic claim that the perceptually given is the foundation of empirical knowledge. [Sellars, W. (1997)]

Der Mythos des Gegebenen besagt demnach, dass Instanzen der Verlässlichkeit bestehen, die den diskursiven Versuchen, Verlässlichkeiten herzustellen und zu überprüfen, zugrunde liegen. Sie sind „self-authenticating", d. h., salopp gesagt, sie sprechen für sich. Darin liegt nun allerdings ein Problem, denn Sinneseindrücke besitzen keine Sprache. Sie können nur dadurch „überzeugen", dass sie auftreten und darin liegt die Crux der Sache. Man spricht zwar von „überzeugendem Auftreten", aber in dieser Redewendung ist impliziert, dass *jemand* überzeugt wird, also bestimmte Individuen unter bestimmten Umständen. Das Vorkommen bestimmter reizauslösender Impulse hat außerhalb der kognitiven Synthese, von der die klassische Erkenntnislehre ausgeht, keinen erkenntnisbegründenden Wert. Es kann nicht als unabhängig verlässlich in den kognitiven Prozess importiert werden. Die Überzeugungskraft, die ihm allenfalls zukommt, erhält es dadurch, dass sich jemand auf Episoden *beruft*, ihnen also eine *Rolle* in der Entstehung verlässlichen Wissens zuschreibt.

Ein Terminus, der in diesem Zusammenhang gerne verwendet wird, ist „Sinnesdaten". Er erlaubt den Querverweis auf die hier diskutierten Streitfragen der Informationstechnologie. „Bloße Daten", die keiner Aufbereitung bedürfen, um zu verwertbaren Resultaten zu führen, sind aus der Sicht herkömmlicher Datenbanktheorien ähnlich suspekt wie der Mythos des Gegebenen für die überlieferte Epistemologie. Um die Verbindung deutlicher zu machen, kann man die Bestandteile des Mythos weiter ausbuchstabieren. „Unbezweifelbare Sinnesdaten" sollen das Kunststück zuwege bringen, gleichzeitig einen garantierten Inhalt zu bieten und jeden Einspruch gegen diese Vorgabe auszuschließen. Sie treten als Begründungen auf, die keinen Widerspruch dulden – das ist ein Rezept für Fundamentalismus.

Zwar lassen sich Phänomene selbstverständlich jenseits unserer Praxis des Begründens denken, z. B. Erdbeben. Naturereignisse kommen ohne Argumentation aus. Aber sie versehen uns nicht von sich aus mit Sichtweisen und Erklärungen. An Aristoteles ist die Frage zu richten, wie etwas Unbewegtes Bewegung hervorbringen könne. An unerschütterliche Sinnesdaten als Fundamente, ist die Frage zu richten, wieso sie sich je nach den Rezeptionsbedingungen ständig ändern. Theoriefreie Gegebenheiten als Ausgangspunkt für Erkenntnis sind dementsprechend eine seltsame Konstruktion. Es soll sich um gänzlich unbestimmte Auffälligkeiten handeln. Doch eine Gegebenheit muss sich von anderen Gegebenheiten abheben, damit sie auffällt. Sie ist diskret und das heißt, dass sie nicht für sich genommen *etwas* sein kann.

Ein Merksatz aus der Wissenschaftsforschung stammt von Geoffrey Bowker. „Raw Data' is both an oxymoron and a bad idea; to the contrary, data should be cooked with care" (Bowker, G. (2008), S. 184) „raw" impliziert Unbearbeitetes und „data", dessen Gegebenheit. Dazu muss es *als etwas* aufgefasst, d. h. in irgendeiner Weise registriert werden. Ungefilterte visuelle Reize entstehen vor der Einordnung als Farben, aber sie sind *Einwirkungen* auf das *Sehvermögen*. Jemand, der hinter die Mustertafeln für Farben zurückgehen will, kann nicht darauf verzichten, sinnliche Affektionen von Phantasien und Sehnerven vom Gehörorgan zu unterscheiden. Geoffrey Bowker hat das Thema im Rahmen des Disputs zwischen Datenmodellen und Big Data ausgeführt. Er bestreitet nicht, dass man jenseits vorgeformter Kategorien statistische Auffälligkeiten z. B. im globalen Proteintransfer oder im Lohnniveau finden kann. (Es handelt sich auch dabei um theoretische Abstraktionen, aber davon einmal abgesehen.) Die Frage sei jedoch, wie diese Gegebenheiten weiter verwendet werden. Die Häufung New Yorker Schadensmeldungen (auf Twitter, nach einem Hurrikan) besagt nicht, dass die Verwüstung dort am stärksten war, sondern dass sich in New York überdurchschnittlich viele Smartphones finden (Crawford, K. (2013); vgl. auch Boellstorff, T. (2013); Vis, F. (2013)). Die registrierten Werte zu Protein bzw. Gehaltszahlungen lassen offen, warum festgestellte Datenverteilungen zu beachten wären. Bowker führt aus:

> It does not just happen that there is a net protein, natural resource drain from the Third World to the First, nor that women in the United States are consistently paid less for the same quality of work as men. These categories represent a reality. ([Bowker, 2003], S. 1995)

Um etwas mit Proteinniveaus anfangen zu können, benötigt man in diesem Beispiel die Kategorien „Erste Welt" und „Dritte Welt". Sie können nicht aus „Rohdaten" gewonnen werden. Damit ist nicht gesagt, dass auf solche Daten bei der *Konzeptualisierung* verzichtet werden könnte. Wohl aber wird darauf bestanden, dass sie – gerade um ihre Überzeugungskraft entfalten zu können – im Rahmen eines kategorialen Gefüges die Realität repräsentieren (können).

„Big Data" bezeichnet einerseits die Herausforderung, dem Datenüberfluss mit neuen informatischen Strategien bislang unerreichbare Ergebnisse abzugewinnen. Die Attraktivität des Unternehmens ist offensichtlich und seine Berechtigung steht hier nicht zur Debatte. Andererseits handelt es sich um einen publizistisch gut verwertbaren Begriff, mit dem versucht wird, mit Rückenwind von den großen Datensammlern (Google, Facebook, Twitter) die Standards aufzuweichen, die bisher verbreitet für die Entwicklung und Bestätigung wissenschaftlicher Theorien galten. Dieser Tendenz wurde hier widersprochen. Ein Beispiel für die Aktualität der Debatte zum Abschluss. Severin Schwan, der Vorstandsvorsitzende des Pharmakonzerns Roche, gab der Frankfurter Allgemeinen Sonntagszeitung im September 2017 ein Interview. Er wurde gefragt, ob seiner Firma die algorithmische Aufbereitung von Daten nicht bald den Rang ablaufen würde. Seine Antwort stimmt mit der hier vertretenen Auffassung überein. „Sie müssen die Abläufe im Körper verstehen, sonst wissen sie nicht, wonach sie in ihren Daten suchen sollen." (Schwan, S. (2017), S. 25) Es trifft zu, dass dabei Vorgaben im Spiel sind, welche die Entdeckung unvorhergesehener Zusammenhänge erschweren. Doch für Überraschungen gilt umgekehrt, dass sie sich nur vor einem etablierten Hintergrund abzeichnen und dass sie, wenn es sich nicht bloß um Irritationen eines Musters handelt, nach einer Erklärung verlangen. Da müssen wieder Theorien helfen. Reichtum an Daten alleine macht nicht klug.

Literatur

Anderson, Chris (2008). The end of theory: The data deluge makes the scientific method obsolete. https://www.wired.com/2008/06/pb-theory/ [28.1.2018].

Boellstorff, T. (2013). Making big data, in theory. http://journals.uic.edu/ojs/index.php/fm/article/view/4869 [28.1.2018].

[Bowker, 2003] Bowker, G. (2003). The theory/data thing. *International Journal of Communication*, 8:1795–1799.

Bowker, G. (2008). *Memory Practices in the Sciences*. The MIT Press.

Bruns, A. (2013). Faster than the speed of print: Reconciling 'big data', social media analysis and academic scholarship. *First Monday*, 18(10).

Bryant, A. and Raja, U. (2014). In the realm of big data … 19(2).

Crawford, K. (2013). The hidden biases in big data. https://hbr.org/2013/04/the-hidden-biases-in-big-data [28.1.2018].

Date, C.J. (2003). *An Introduction to Database Systems*. Pearson.

Emrouznejad, A., editor (2016). *Big Data Optimization: Recent Developments and Challenges*. Springer International Publishing.
Floridi, L. (2012). Big data and their epistemological challenge. http://link.springer.com/10.1007/s13347-012-0093-4 [28.1.2018].
Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society,* pages 1–12.
Mayer-Schönberger, V. and Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work and Think*. John Murray.
Sathi, A. (2012). *Big Data Analytics: disruptive technologies for changing the game*. MC Press.
Schwan, S. (2017). Krebs ist heilbar. Das ist revolutionär. *Frankfurter Allgemeine Sonntagszeitung*, 37.
Sellars, W. (1997). *Empiricism and the philosophy of mind*. Harvard University Press.
Steadman, I. (2013). Big data and the death of the theorist. http://www.wired.co.uk/article/big-data-end-of-theory [28.1.2018].
Vis, F. (2013). A critical reflection on big data: Considering APIs, researchers and tools as data makers. http://journals.uic.edu/ojs/index.php/fm/article/view/4878 [28.1.2018].
Wittgenstein, L. (1980). *Tractatus logico-philosophicus. Tagebücher 1914 – 1916. Philosophische Untersuchungen.* Suhrkamp.
Wittgenstein, L. (1984). *Tractatus logico-philosophicus*. Suhrkamp.

ao. Univ.-Prof. i.R. Dr. Herbert Hrachovec
Universität Wien
Institut für Philosophie
Universitätsstraße 7
1010 Wien
Österreich
herbert.hrachovec@univie.ac.at
http://hrachovec.philo.at

Herbert Hrachovec ist Hochschullehrer im Ruhestand am Institut für Philosophie an der Universität Wien. Er arbeitete über analytische Philosophie, Metaphysik und Ästhetik. Gegenwärtiger Arbeitsschwerpunkt sind Neue Medien. Die Unterlagen seiner Vorlesung zu „Big Data“ 2015 sind abrufbar unter http://wiki.philo.at/index.php?title=Big_Data_(Vorlesung_Hrachovec,_WS_2015). 2001 bis 2010 war Herbert Hrachovec stellvertretender, zuletzt Institutsvorstand des Instituts für Philosophie, 2006 bis 2010 und 2012 bis 2013 Mitglied des Senates, 2005 bis 2010 Vorsitzender der Curricularkommission der Universität Wien und 2010/11 Koordinator des EU-Projektes „Agora. Open Access in the Humanities“.