# Social Research in Times of Big Data: The Challenges of New Data Worlds and the Need for a Sociology of Social Research

Diaz-Bone, Rainer; Horvath, Kenneth; Cappel, Valeska

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

**Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:**
GESIS - Leibniz-Institut für Sozialwissenschaften

**Empfohlene Zitierung / Suggested Citation:**
Diaz-Bone, R., Horvath, K., & Cappel, V. (2020). Social Research in Times of Big Data: The Challenges of New Data Worlds and the Need for a Sociology of Social Research. *Historical Social Research*, *45*(3), 314-341. https://doi.org/10.12759/hsr.45.2020.3.314-341

# Social Research in Times of Big Data. The Challenges of New Data Worlds and the Need for a Sociology of Social Research

*Rainer Diaz-Bone, Kenneth Horvath & Valeska Cappel* *

**Abstract**: *»Sozialforschung in den Zeiten von Big Data. Die Herausforderungen der neuen Datenwelten und die Notwendigkeit einer Soziologie der Sozialforschung«*. The phenomenon of big data does not only deeply affect current societies but also poses crucial challenges to social research. This article argues for moving towards a sociology of social research in order to characterize the new qualities of big data and its deficiencies. We draw on the neopragmatist approach of economics of convention (EC) as a conceptual basis for such a sociological perspective. This framework suggests investigating processes of quantification in their interplay with orders of justifications and logics of evaluation. Methodological issues such as the question of the "quality of big data" must accordingly be discussed in their deep entanglement with epistemic values, institutional forms, and historical contexts and as necessarily implying political issues such as who controls and has access to data infrastructures. On this conceptual basis, the article uses the example of health to discuss the challenges of big data analysis for social research. Phenomena such as the rise of new and massive privately owned data infrastructures, the economic valuation of huge amounts of connected data, or the movement of "quantified self" are presented as indications of a profound transformation compared to established forms of doing social research. Methodological and epistemological, but also institutional and political, strategies are presented to face the risk of being "outperformed" and "replaced" by big data analysis as they are already done in big US American and Chinese Internet enterprises. In conclusion, we argue that the sketched developments have important implications both for research practices and methods teaching in the era of big data.

**Keywords:** Big data, economics of convention, sociology of social research, sociology of quantification, sociology of health, data quality, political economy of quantification, data infrastructures, fact/value dichotomy.

* Rainer Diaz-Bone, Department of Sociology, University of Lucerne, Frohburgstrasse 3, 6002 Lucerne, Switzerland; rainer.diazbone@unilu.ch.
Kenneth Horvath, Department of Sociology, University of Lucerne, Frohburgstrasse 3, 6002 Lucerne, Switzerland; kenneth.horvath@unilu.ch.
Valeska Cappel, Department of Sociology, University of Lucerne, Frohburgstrasse 3, 6002 Lucerne, Switzerland; valeska.cappel@unilu.ch.

Historical Social Research 45 (2020) 3, 314-341 | published by GESIS
DOI: 10.12759/hsr.45.2020.3.314-341

## 1. Introduction

Big data and digitization have been met in the social sciences with mixed feelings, from almost utopian enthusiasm to more dystopian visions (Savage and Burrows 2007; Boyd and Crawford 2012; Pentland 2014; O'Neil 2016; Noble 2018; Zuboff 2019). From the standpoint of established quantitative social research methodologies, it is tempting to emphasize that many of the points raised in these debates do not sound that new at all. Huge datasets have been used by the social sciences for centuries, exploratory data analysis has been promoted from the 1950s onwards, and the current heralds of a coming age of "social physics" should probably be more aware of their 19th century predecessors (for which Auguste Comte and Adolphe Quetelet are just two prominent examples). Seen from this angle, we might give a double negative answer when asked about the "quality" of big data: they are neither really "new" nor particularly "good."

We argue that such a take on the problem risks missing the crucial point: The phenomenon of big data is indicative of an ongoing and fundamental shift in the social research landscape. This shift cannot be fully grasped from the standpoint of established methodological frameworks because these are the very frameworks that are being challenged in the first place. The key claim of this paper is that in order to grasp what is at stake, we need to discuss the methodological questions surrounding big data in relation to the precise historical and situational contexts in which they arise. In other words: We need to move towards a perspective of a sociology of social research that links methodological questions with analyses of both broad political and social configurations and the pragmatics and logics of doing social research. Seen from such an angle, we are not merely witnessing the emergence of new kinds of data and data analysis, but a whole set of transformations, involving, among others, epistemic values and orientations, institutions and forms of quantification, data and problem ownership, or the relations between research and the common good. We argue that it is crucial that the social sciences recognize and face the resulting challenges. The price of missing out may be high, given that social scientists have already lost part of their supremacy of analyzing and interpreting the social world (van Dijck 2014; Diaz-Bone 2019).

The field of health is among the social areas in which the disruptive transformation linked to datafication is most advanced and already observable in daily routines as well as in institutions and policies. We therefore use the field of health as the main example throughout this article to illustrate the changing arrangements of actors, data infrastructures, social and economic valorization

of data, and epistemic practices, values, and orientations.[1] Both the social sciences and the field of health face the challenge of negotiating what "good data" is about and on which common principles data are based.

In the following section, we briefly outline the conceptual foundation of our argument, which is mainly informed by the economics of convention (in short, EC) and related neopragmatist epistemologies (section 2). On this basis, we then deal with three questions. First, we ask what makes the new quality of big data (section 3). We provide a first account of what we believe are essential features of newly emerging data worlds, highlighting the close interplay between processes of quantification and datafication, political and economic contexts, and social processes that become organized in and around new data worlds. Secondly, we approach the question of what is at stake (section 4). We emphasize the implications of these new data worlds, pointing among others to the danger of algorithmic discrimination and knowledge production happening without any form of public control. Finally, we ask what kind of challenge exactly social research faces in dealing with these changed realities and why this challenge may be anything but trivial (section 5). We provide a sociological and methodological reading of the epistemic transformation we are witnessing, contrasting current data and research arrangements with the canonical forms of social research that dominated the 20th century. We claim that the challenges of big data are most pronounced vis-à-vis the canonical quantitative methodologies that have dominated social sciences for decades: they question established statistical techniques as well as key epistemic values and orientations underpinning these approaches. Overcoming this challenge will require taking alternative logics and techniques of doing social research seriously, as they have been developed in fields of exploratory statistical analysis (such as geometric data analysis; Le Roux and Rouanet 2010) but also in qualitative research contexts. The imminent crisis hence is also an opportunity for reconfiguring social research in ways that move beyond divides and boundaries that have dominated methodological debates for decades. Such a reconfiguration seems necessary lest the social sciences lose authority to analyze and interpret developments in the social world, with important implications regarding public accountability of societal knowledge production. We conclude that data quality is linked at once to methodological issues and to aspects of the political economy of quantification. Therefore the contribution calls for a new sociology of social research (section 6).

[1] The content of the article is also based on the current research project: "Digital health classifications in apps: Practices and problems of their development and of their situational application." It is a research project (2019-2022) applying convention theory, funded by the Swiss National Science Foundation (SNSF), and located at the University of Lucerne. For further information, see: <https://www.unilu.ch/fakultaeten/ksf/institute/soziologisches-seminar/forschung/digitale-gesundheitsklassifikationen-in-apps-praktiken-und-probleme-ihrer-entwicklung-und-situativen-anwendung/>.

## 2. Conventions, Coordination, and Contexts: A Conceptual Note on the Sociology of Social Research

This article argues for moving towards a sociology of social research in order to fully grasp the foundations and implications of the current critical moment in the development of empirical social research and its methodologies. We build on neopragmatist thinking and on the "economics of convention" (EC) as they have been developed especially in the context of the new French social sciences (Storper and Salais 1997; Boltanski and Chiapello 2005; Boltanski and Thévenot 2006; Eymard-Duvernay ed. 2006a, 2006b; Batifoulier et al. 2016; Diaz-Bone 2018). The objective is to systematically link the methodological question of "what makes (good) data" to a political economy of datafication as well as to an analysis of the pragmatics of social science knowledge production. Such a perspective allows bridging the discussion of methodological principles with analyses of dominant institutional and cognitive forms, epistemic values and orientations, forms of assessing and justifying data quality, and the political and economic links through which social research is related to public concerns and the common good. The rationale for such a perspective is that such a bridging is required for understanding the concrete challenge that social research faces.

From an EC-perspective, *conventions* are logics of coordination that actors rely on when they have to interpret, evaluate, and valuate others, actions, objects, or processes in situations (Storper and Salais 1997; Boltanski and Thévenot 2006). In situations of critique or in need of justification, actors can refer to conventions as deeper principles in order to relate to worth, "greatness" (in French "grandeur"), or common good. So far, EC scholars have identified and discussed a variety of such conventions, including the market convention, the industrial convention, the domestic convention, the civic convention, the convention of reputation, the convention of inspiration, the green convention, and the network convention.[2] Every empirical situation is influenced by a combination of these conventions, and further conventions are conceivable as possible resources for critique or different modes of coordination. Actors hence have to be practical metaphysicians and be capable of coping with a co-existing plurality of conventions as normative principles.

The analysis of processes of quantification was one of the starting points and building blocks for the economics of convention. EC was developed in the

[2] EC has introduced other models of conventions as the model of "worlds of productions" introduced by Storper and Salais (1997). But all these notions of convention share the general perspective that situations, burdened by uncertainty, are in need of being accomplished by conventions, which actors can apply to understand what is going on and what are ways of coordination.

analysis of social and methodological usages of classifications, quantifications and statistics (Diaz-Bone 2016, 2018). It can therefore be conceived as an approach in the field of sociology of social research (Diaz-Bone and Didier 2016).[3] Alain Desrosières' work on quantification serves as an emblematic point of reference (Desrosières 2000, 2008, 2002, 2011, 2015, 2016). Desrosières suggests distinguishing *issues of measurement* from *processes of quantification*. Quantification comes first and signifies the process of establishing definitions, standards, procedures for numerical classification and measurement. The process of quantification requires negotiation, coordination, and critique. It involves actors from different fields of practice and is always convention-based: Every quantification is anchored in a specific logic of justification that legitimates the counting or measuring of some phenomenon in a specific way. Such conventional logics are linked to orders of worth, related to the common good in specific ways, and imply certain ways of problematizing the social world. To see measurement and quantification this way means to insist on coherent and adequate conventions of measurement so that the result is data interpretable for coordinating actors and suitable for collective action, always with the overarching objective of pursuing a common good (Salais 2012, 2016). While the results of these conventionalizing processes may remain explicit or at least accessible (e.g., in codebooks or data manuals; Salais 2016, 119), most of the process is invisible in later stages in which instruments are treated as uncontested, uncomplicated, and, in this sense, necessary (rather than "conventional").[4]

Such a genuinely sociological view on social research and its methodologies does not imply a relativist standpoint (Desrosières 2002). An EC-inspired sociology of social research, to the contrary, aims to make social research more realistic by empirically investigating the actual forms of doing social research that have emerged and become accepted as feasible and legitimate ways of solving the uncertainties and complexities of knowledge production. Building on this understanding, we may define datafication as the institutionalized pro-

[3] The journal *Historical Social Research* has devoted a series of special issues to EC. See Diaz-Bone and Salais (eds.; 2011, 2012); Diaz-Bone et al. (eds.; 2015); Diaz-Bone and Didier (eds.; 2016); Diaz-Bone and Favereau (eds.; 2019).

[4] The German-speaking category of "migration background" provides a striking and topical example for the manifold social processes underlying statistical definitions and the effects that statistical categories can develop once they are established (Horvath 2019): The category was established as an official category ("defined in law") in the mid-2000s as a successor to the category of "foreigner" that had outlived its utility of capturing migrant populations. The aim was to capture a group of "othered" migrants without referring to explicitly ethnic markers. Educational professionals, migrant groups, and public administrations were among the players involved. Today, the category is used as if it denoted a self-evident feature of the social world. As such it has itself become a social reality that develops important consequences, for example by structuring narratives of educational inequality and resulting pedagogic strategies (Horvath 2018).

cesses by which social practices and processes are *transformed into and represented as data*, with the aim of using these data for specific purposes and in specific ways.

The key analytical benefit of this understanding is that we can develop the notion of datafication in two directions. First, we can ask how datafication relates to political and economic dynamics – we can investigate the political economy of data (Desrosières 2011, 2015). In this vein, we may ask what makes data valuable for different groups of actors. How is data made use of – how is their value realized and exploited? How are these forms of valuating data linked to societal dynamics? Second, we can focus our attention on the institutions, practices, and processes involved in datafication, on the inner workings and epistemology of data. One of the fundamental lessons of neopragmatist thinking in this context is that we conceive of these frameworks as involving epistemic values and orientations, which allow actors involved in knowledge production to coordinate, to reflect, and to evaluate (Putnam 2002). To give but one example that is relevant for the following discussion, quantitative research in the 20th century strongly favored parsimonious model building (rather than exploring complexity), motivated by value orientations such as transparency, neutrality, and communicability (Smith 1994).

There are a number of heuristic concepts that help applying a sociology-of-social-research perspective to concrete datafication processes. Alain Desrosières and Laurent Thévenot famously introduced the notion of the *statistical chain* that links different involved actors and situations (Desrosières and Thévenot 1979; Desrosières 2000). This allows for an understanding of how datafication and quantification can lead to invalid data when the involved actors switch conventions along the chain (Thévenot 1983; Desrosières 2008; Diaz-Bone 2016, 2017, 2019). Conventions hence are not only relevant as means of ensuring democratic control and public accountability, but also as basis for ensuring methodological coherence. In a similar vein, we can focus on *data infrastructures* as key interface between political and economic orders and everyday epistemic practices (van Dijck 2014; Gray et al. 2018; see also Bowker and Star 2000). Data infrastructures are conventionalized and institutionalized resources that guide and structure the datafication of social phenomena. One essential feature of these infrastructures is that they provide frameworks for assessing the quality of data. These frameworks have developed over time, resulting from form-giving activities by multiple authors; they provide schemes and criteria for evaluating data. Finally, we may conceive of concrete arrangements of these elements as *data worlds*: specific ensembles and interplays of players, institutions, and practices.

In the following, we use this conceptual framework and its heuristic resources in three steps. First, we look at knowledge production in broad societal, political, and economic contexts. What kind of data is produced? What is new about these current data worlds? What kinds of infrastructures and political

economies of data are we facing? Second, we ask what are the wider implications of these new data worlds? Third, we turn the perspective inside out and focus on knowledge orders and institutional dynamics in the field of the social sciences: What kind of challenge, exactly, results for the social sciences?

## 3. What Makes the "New Quality" of Big Data?

Big data is often introduced using a multiple-V definition. The most common one is the 3-Vs definition: volume, variety, and velocity. In this vein, Rob Kitchin argues that "big data is huge in volume, consisting of terabytes or petabytes of data; high in velocity, being created in or near real-time; diverse in variety, being structured and unstructured in nature" (Kitchin 2014, 1).[5] Big data, conceived of in this fashion, is at first glance characterized by an immensity of numerical information, too big as to be handled by ordinary data analysis technology. And big data is known to be "untidy" data, which differs from the more "tidy" and rectangular data file formats which have dominated the social sciences for more than half a century.

But to consider only aspects of mass and format would mean to overlook big data's social, scientific, economic, and political character and its already effective capacity to transform societies. Any serious attempt at characterizing ongoing transformations of data worlds must take into account that sensors, indicators, video surveillance, and manifold digital devices each storing masses of social data have become to pervade social life worlds. Smart phones, cars, entertainment technologies (such as Smart TVs), and digitized kitchen appliances in private homes are nowadays equipped with computer chips. The crucial development, which has persistently changed social life worlds as well as the public sphere, however, is the linkage of all these objects by the Internet ("Internet of Things," "Smart Cities"). Social life worlds today are digitized and datafied. Consequently, big data should not merely be conceived of as numerical *representation* of individual behavior (such as consumer behavior), but instead *social reality itself* needs to be seen as being transformed into digital processes.

The datafication of health provides one of the most impressive and obvious examples for this profound transformation (Ruckenstein and Schüll 2017). This transformation appears in health data infrastructures like data-driven medical research, governmental digital patient folders, individualized health care and in self-care practices. Biobanks and governmental databases run by artificial intelligence support health research and developing individualized methods of

[5] For more elaborated definitions, see Mayer-Schönberger and Cukier (2013); Wierse and Riedel (2017).

treatments (Abouelmehdi et al. 2018, 3). Governmental digital patient folders provide a data infrastructure to collect and share patient data trough hospitals, doctors, drugstores, health insurances, companies, and private persons. In everyday life, new technologies like health apps, wearables, implantable biosensors, drugs with sensors, patient monitoring devices, or digital consultation-hours are getting more and more part of the social reality (Ruckenstein and Schüll 2017, 262). Especially the intrusion of health apps generates a new understanding of individuals' health-status and behavior. There is a community called "Quantified Self" (QS), whose members follow the objective to gain "self-knowledge through numbers."[6] They try to get deep insights about their environment and physical conditions through quantification and data analysis of any aspects of social life (Nafus 2016). While the generated health data provides new insights, it is by no means a one-to-one account of one's life or identity (Sharon 2017). Rather, it equates a "situated objectivity," meaning that people engage with their personal data (Pantzar and Ruckenstein 2017). In this view the data reflects socially contextualized health assumptions, experiences, and aspects of the daily life and makes them visible, valuable, and socially negotiable through measurements (Ruckenstein and Schüll 2017, 266).

Some theorists argue that big data will completely transform the organization of economies and of enterprises, among others by replacing money and making central banks obsolete (Mayer-Schönberger and Ramge 2018). With a view to the logics and practices of social research, big data, however, has other important aspects. In the following, we discuss various regards in which "big data" are qualitatively different from the kind of data the social sciences have become used to. These aspects are neither comprehensive nor mutually exclusive. In this section, they serve to illustrate the deep entanglement of epistemic, social, economic, and political dynamics involved. The discussion, at this point, aims to illustrate the complex and interrelated character of these developments.

First, the *type of data* is different in other regards as the ones underlying the 3Vs-definition. The data infrastructures, which generate and proceed big data are built out of technical platforms (Gillespie 2010) and devices implemented by private enterprises and used by these companies for their technological and commercial purposes. The big Internet Corporations Google, Amazon, Facebook, Apple, and Microsoft in the USA or Baidu, Alibaba, Tencent, and Xiaomi in China, and other big retailers such as Walmart but also smaller enterprises as well as "start-ups" which produce machines or services rely on sensors and apps that are connected with the Internet and deliver data continuously to these companies. Such developments have prominently taken place in the field of health where new health data infrastructures linked to a new economy of

[6] See <https://quantifiedself.com/>.

health have emerged. A development in the health system in Denmark has shown in recent years that not only individuals but also institutions can become part of an unconscious data economy (Wadman and Hoeyer 2018). These developments are partly further promoted by health professionals, government agencies, and health insurances, leading to invisible richness of data that these actors possess (Nissenbaum and Patterson 2016). For example, in Denmark, a group of doctors advocated an infrastructure for a seamless exchange of health data to assist them in treating patients superiorly and making their work easier. The infrastructure was perfectly connectable and in fact so useful for many purposes that the very richness of data finally led to the collapse of the infrastructure: Physicians realized that the data sourcing software was also used or was planned to be used by authorities and patients to control and monitor the doctors (Wadman and Hoeyer 2018, 9). Users of health and well-being apps mostly suspect nothing of the already established data infrastructure and data economy hidden in the background, while their data is getting attention and usage at several levels (Nissenbaum and Patterson 2016).

The data we see in these cases are much more fine-grained than for example survey data, which is still the most commonly used sort of data in academic social research. Also, these data are gathered mainly without individuals being aware of it. In contrast to survey data, big data mainly tracks individuals' behavior instead of opinions, values, or attitudes. While social research is in need of collecting data on actors' motives and aims as important explanatory variables, big data analysts instead focus on predictive accuracy of their data models, circumventing the inclusion of meaning and actors' interpretation.

These new types of data, *second, are utilized in new forms and linked to new sets of political and economic interests*. Data itself nowadays count as an economic resource. Most importantly, they are an important source of revenue for companies in the sector of new information technologies. Huge technology ventures like Google, Facebook, Apple, or Amazon push into the healthcare market because their business models are partly based on health data (Sharon 2018). For example, Apple is active with its apps on IOS devices, both in clinics and in private everyday lives of patients and healthy people. In everyday clinical practice, the apps should enable a more efficient data exchange, so that doctors have access to patient records all the time and nurses have access to all patient data to ensure safe medication. Apple is also developing new gadgets and tools, such as the Apple Watch or the Health App, that are connected to the iPhone and enable patients and healthy people to share health data with hospitals, doctors, and nursing teams. Apple is targeting healthcare providers such as hospitals and doctors and everyday life people with its health-products. The business model consists, on the one hand, of data aggregation and its distribu-

tion, but also of selling gadgets and tools such as Apple Watches or apps.[7] These enterprises have the necessary capacities to build new data infrastructures, not least because new consumer technologies are an essential ingredient of newly arisen health data infrastructures. Companies such as Google, Apple, or Facebook have a lot of experience in developing user-friendly and attractive applications to generate and collect data, including health data. Companies like Fit-Bit and Nike or online platforms like PatientsLikeMe share and sell their data with respect to affiliated companies, tech firms, insurers, marketers, pharmaceutical companies and medical device makers (van Dijck and Poell 2016; Ruckenstein and Schüll 2017). Other actors use new health data for other ends. Hospitals want to save costs and identify preventive health-scores (Hogle 2016; Ruckenstein and Schüll 2017). Advertisers, insurance companies and credit rating agencies are primarily interested in creating or completing data profiles in order to offer personalized products and services and to classify people. There are abstruse developments like growing "marketing-avatars," for example a "marketing-baby" made of connected health-data from pregnant woman. Companies and health services use new kinds of networked data to offer advertisement or products via e-mail, phone and post. Cynically, even in cases of miscarriages these avatars can live on, growing and developing for years because the "marketing-baby" is still in the databank growing (Ebeling 2016). These new health data infrastructures impressively show how different interests, (uninformed) actors, and non-chained situations are being linked along the statistical chain and lead to reorganized social realities. In particular, the digital connectivity of new health data on such infrastructures distinguishes them from previous health data. Context-independent reuse and the linking of new data-field situations open up completely new contexts of action and new forms of negotiation of health and health quality (Hogle 2016; Sharon 2016; van Dijck and Poell 2016).

A third key issue that is directly related to the uses and utilities of data is the *access to big data (generating) sources – including their costs and their ownership*. In contrast to engineering sciences and natural sciences with their manifold opportunities to collaborate with big private companies, there are only few possibilities for academic social research to have the research infrastructure financed by enterprises. There are likewise only few public initiatives and state organizations, which support the new needs of social research to build up data structures comparable to the technical Internet platforms of global players such as the big Internet companies mentioned above.[8] Hence, the access of academic social research to new types of data is limited to applying strategies such as

[7] See: <https://www.apple.com/healthcare/>.

[8] One could refer to the international project of Large Hadron Collider (LHC) or to national initiatives as social science research infrastructures (Gehring 2018).

web scraping and text mining or using APIs[9] to collect data (Wiedemann 2013; Munzert et al. 2015; Foster et al. 2017; Hampton 2017).[10] These strategies only partly solve the problem, since companies usually control and limit data access, as is the case, for example, with data obtained through the Twitter API (Boyd and Crawford 2012; Diaz-Bone 2016). Although, at first glance, the Internet suggests the possibility of free access to data, one must be aware that there is a second realm of data, hidden for ordinary users and social researchers: data proceeded and stored in the closed realms of Internet companies, in their source codes, and in their private algorithms.

The kind of data produced, their utility, and their ownership are inherently linked to methodological concerns and questions of the quality of data for social research. We now turn to these more methodological issues. A fourth feature of new data worlds follows directly from the problem of big data access: *social research lacks control over data formats and over biasing influences and other error sources in the data production*.[11] Big data is mostly generated automatically by technical and commercial processes in physical environments and in social worlds. Social researchers do not play the role of designers of instruments for data collection and they only partly have control over research designs (as would be the case when using big data for quasi-experimental research; see Salganik 2018). The process of operationalizing construct dimensions into variables is not in the hands of social researchers (as it is in classical surveys) but is operated by engineers, computer scientists, data scientists, and other professions involved in the everyday practices of implementing hardware (sensors) and programs (apps, algorithms) as the basis for gathering and proceeding big data, employed mainly by private enterprises. For social researchers, when relying on big data generated by these companies, the increasing division of labor, the different technical standards involved and the various platforms along the statistical chain undermine the transparency of possible influences on data formats and information "in the data" (including ad hoc decisions about categorization, unknown principles of data transformation, loss of metadata, missing knowledge about situational conditions of measurement, involved filtering processes, problems of storage and of adequate data flows, etc.). The first consequence is a loss of data quality because of this missing transparency about influences and possible error sources. Seen from the view of social researchers who do not control the statistical chain, the coherence of data and the meaning of data are questionable (Lagoze 2014, Diaz-Bone 2016). The second consequence is an erosion of the validity of conclu-

[9] API is the abbreviation for "application programming interface."
[10] For a German introduction into big data analytics, see Wierse and Riedel (2017).
[11] For a discussion of similar problems in the field of social bookkeeping, see Baur (2009).

sions based on such data (Jeese 2018).[12] In turn, big data producing companies which can control the data infrastructure and the statistical chain are gaining power over knowledge-production.[13] Big data is enforcing a change in the system of professions (Abbott 1988) and a change in the social distribution of power exerted in data based knowledge production.

Fifth, *some established approaches and concepts of social research are not applicable to the majority of big data analyses, including usual statistical parameters being not calculable*. An important example concerns the notions from sampling theory like representativeness, target populations, standard error, and others. In almost any study, big data is "found" or generated in everyday procedures. Therefore, empirical studies cannot enforce their own sampling design. It is obvious that ex post it will be difficult to define, delimitate, or identify the target population(s) about which generalized claims shall be made. There might be a commercial interest in detecting consumer behavior patterns in data and ignoring other populations. But for social research as social institution it can hardly be justified to exclude various societal groups just because they happen not to be covered by social science data. The emergence of new "data analysis ideologies" should be regarded skeptically, when big data analysts claim that representativeness is only a question of enough data or that big data can correct for measurement error and bias (Mayer-Schönberger and Cukier 2013). Rising nonresponse rates in surveys and different life style habits result in different social groups being represented to varying degrees in big data sources (Japec et al. 2015).

As a sixth aspect one can point to the *development of data analytics towards real-time analysis run by algorithms which are based on machine learning and artificial intelligence*. This enables algorithms to evolve (to "learn"). Trading stocks in stock exchange is one example where algorithms have already replaced human decision making. The usage of pattern recognition to identify human beings (as it is already applied in some cities of China) is an example where machine learning meets big data analysis. There are countless further private, commercial, political, military or scientific applications in which artificial intelligence and machine learning is combined with big data. For social research this trend enforces the questioning of traditional methodological practices such as exploration, interpretation, and understanding. As long as social research regards data only as measurement of something else (as representations of attitudes, respondents' properties and so on), social researchers will not

[12] Commercial statistics portals (such as statista.com), data trading companies, and data trading platforms offer services in data analytics and compile different data sources. Thereby incoherent data sources and data formats can be linked to "big data." For an overview of big data markets, see Liang et al. (2018).

[13] Foucault developed the notion of power-knowledge to capture these relations between power and knowledge (Foucault 1995).

grasp this new phenomenon implied "big data" and their algorithmic analysis.[14] If, seemingly, the exploration of patterns and structures can be done by algorithms, if the development of hypotheses and assumptions in research can be done by computers, and if interpretation and understanding have become outdated, then social research is not only in danger of becoming privatized, it is in danger of becoming altogether obsolete.[15]

## 4. What is a Stake? Why We Should Care about New Data Worlds

Taken together, these aspects (the new types of data we witness, the utilities and forms of valuation of these data, control over data chains and infrastructures and their possible influences on data, the incoherence between these new data sources and established social research methods, and finally the implications of algorithmic processing of data) indicate a profound transformation. This transformation raises important methodological issues, even if it is still to be seen if big data and big data analysis will actually prove to be the kind of disruptive innovation for social research and research methodologies that it has been for other fields (Kitchin 2014), such as economics (Einav and Levin 2014)[16] or the social sciences more generally (Lazer et al. 2009; Pentland 2014; Alvarez [ed.] 2016). So far, there have been only few methodological reviews, sketching methodological potentials and deficiencies of big data rather programmatically (Boyd and Crawford 2012; Japec et al. 2015; Kleiner et al. 2015; Lazer and Radford 2017). Nonetheless, the emergence of big data has been repeatedly linked to diagnoses of an upcoming crisis of empirical social research and its established methodologies (Savage and Burrows 2007; Burrows and Savage 2014).

These methodological concerns are inherently linked to questions regarding the authority of interpreting social phenomena. In a nutshell: What is at stake is the ability and mandate of the social sciences to analyze and interpret newly emerging social formations (Bartlett et al. 2018). Increasingly, other professional groups (such as computer scientists, often employed by or linked to private businesses) produce knowledge of the social world (see, e.g., Pentland

[14] To be more precise: if big data is (still) conceived of as a representation of individual behavior, it mostly captures social behavior in more indirect ways as digital traces (as in the case of geo-data or Internet use tracked by cookies). Visual and audio data is overall closer to data representing actual behavior.

[15] In mathematics, algorithms are developed and engaged to find new mathematical equations and proof them automatically (Novotny 2015). This way the main business of mathematicians is taken over by algorithms and computers.

[16] For an evaluation for economic sociology, see Diaz-Bone (2017).

2014). Their great resource is that they can claim to be up-to-date regarding analytical techniques and to have access to a so far unseen richness of data and information. These authors, however, seldom have a background in the social sciences and are hence unaware of conceptual and methodological foundations that have been built over the past two centuries. The risk hence is one of a complete backlash into forms of knowledge production that are more reminiscent of Lombrosian craniology than of modern relational social sciences – as the example of algorithms that are presumably capable of classifying criminals versus non-criminals discussed by Bartlett et al. (2018) shows.[17]

The methodological question of data quality cannot be artificially decoupled from the social and political contexts in which it arises. In other words, any methodological critique also requires discussing political and social foundations and implications of new data worlds. This is particularly obvious considering the key role that conventions play for datafication. Algorithms need criteria for optimization and fitting, and these have to conform to publicly debated standards and objectives of collective action. In this sense, conventions are foundational for the programming of algorithms, if the results are meant to correspond to a common good and to deliver information relevant for collective action. Otherwise, algorithms rest on nothing else than ad hoc decisions or some programmers' opinions about adequate criteria, and will generate performative effects (Callon 1998), which realize their own (negative) realities.[18] As Paul Dourish (2016) has shown, algorithms are, as procedures of calculation, distributed over different steps and may be distributed over different devices (computers, programs, etc.). Dourish further argues that developers of algorithms can have a stronger interest in the opacity of algorithms than in their transparency, partly because they have an interest in shielding their programming work (Dourish 2016, 4). The opacity of algorithms impedes their analysis and their social or methodological critique. This is especially an issue for research ethics, when opacity restricts the analysis of normative principles of algorithms and their consequences. The "algorithmic" claim that research can

[17] Another prominent example is "Compass," a "risk assessment algorithm," which is a product of a private enterprise and offers to judges in court "quantifications" of the risk of recidivism (Israni 2017). This privately owned algorithm is not transparent to judges or the public. Also, the juridical adequacy of sentences based on the algorithm cannot be validated, when persons are sent to prison for many years. A second example is the commercial software Compstat, which was implemented in police departments of cities as New York and Paris and invents indicators to quantify criminal activities in urban areas. It has been shown that some commanders of these police departments "[…] asked for wrongful arrests and police work dedicated primarily to reaching better numbers" (Didier 2018, 525). And it has to be added that actors resist effects of quantification and also quantification itself. Alain Desrosières has labelled this resistance as "retroaction" (Desrosières 2015).

[18] Cathy O'Neil has stated that an algorithm is an "opinion formalized in code" (O'Neil 2016, 53). For an introduction into the (critical) sociological research about algorithms, see Kitchin (2017).

dismiss the search for causal explanations and instead focus on correlation patterns alone (Anderson 2008) is not only hard to sustain when social research aims to understand social reality, social structures, and social processes, but also highly consequential. A-theoretic predictions become all the more troubling when they turn out to work as self-fulfilling or self-enforcing decisions. Algorithms, which are unable to understand social categories, their genealogy and interpretation, implicit meanings, and social inequality as context, have correspondingly been shown to re-enforce racism (Noble 2018) and to increase social inequality (Eubanks 2017). From the perspective of EC, measurement and datafication therefore need to be based on conventions, on which social groups, social movements, and representatives of social institutions have agreed. On this basis, these conventions enable the generation of data that are adequate as informational basis to decide and frame actions, which pursue a common good. Data analysis ideologies, which ignore this need as well as the demand for coherent statistical chains, in the end, neglect the legitimate public expectation that social research should be oriented towards the needs of the public and should address the common goods. The notion of data quality in EC therefore is linked to both the idea of adequacy and the idea of justice.

Again, the field of health provides striking examples for the necessity of relating to a common good as well as for the deep entanglement of politics, economics, and data methodologies. Even before the emergence of the new data economy and digital capitalism, shifts from welfare state's organizational principles to market-based organizational principles were observed, particularly in European health care systems (Batifoulier et al. 2018; Da Silva 2018). Evidence-based medicine was introduced in the 1980s with a focus on practitioners' clinical experience, state-of-the-art clinical research, and data collection on current treatments as evidence (Staii 2018, 199). This form of data collection for medical data has been steadily expanded and has been reflected at the policy level through standardization efforts (Da Silva 2018; Staii 2018, 199). These developments at the political level have been informed in particular by the problematic assumption that health institutions are primarily to be understood as incentive systems. In this way, actors of mainstream economic theory have been politically conceived as rational, calculating, and self-interested actors who pose a potential risk to health care costs and who do not follow a common good orientation (Batifoulier et al. 2018). For health insurances and physicians, the risk of illnesses could only be identified by statistical distributions, and evidence-based medicine interpreted the individual case against the background of collective data (Staii 2018, 200). With the datafication of health, a renewed epistemic change seems to be emerging. The emergence of new data infrastructures and networking of new (individualized) data with old health data promotes in particular an individualized, predictive, and preventive medicine. Actors are thus increasingly made responsible and mobilizable as empowered and preventive patients (Staii 2018, 199), but at the same time their orientation

towards the common good is ignored (Batifoulier et al. 2011, 153; Ruckenstein and Schüll 2017, 272). The epistemic rupture is particularly evident in the evolving redistribution of disease risk. In an individualized medicine, the risk of illness is no longer an expression of a random distribution with arbitrary fate for a person. On the contrary, predictive medicine mobilizes general and individualized knowledge for a personalized diagnosis and a direct allocation of risk. In the process, the data approaches do not dissolve but complement each other with the aim of eliminating the disease risk via measurements and knowledge, or are at least redistributing it to an individual level (Staii 2018, 200). In addition, economic policies promote an active, well-informed patient in order to give them market power, making the field of health more accessible to the market and generating demand (Batifoulier et al. 2011). The problem of all those developments is the ignorance of the plurality of political, ethical, economic, and professional values, especially in the health system (Batifoulier et al. 2018). Reforms in the healthcare system based on the same arguments of legitimacy for improving quality of care and reducing costs have not worked out properly. Rather, they have led to the increase of certain qualities, such as an industrial quality, and a decline in domestic and civic quality (Batifoulier et al. 2018; Da Silva 2018). This non-negotiated shift of public interest orientations in the health system is already leading to worrying developments and crises: competitive dynamics in hospital health fees that create social inequalities in access to health, and standardized inhumane medical interventions that are not necessarily for the benefit of patients. Finally, this raises the question of the legitimacy of health decisions and the urgency to negotiate them together (Batifoulier et al. 2018). With the emerging data economy in the field of health, the problem of missing negotiation processes is further compounded because common goals, values, and qualities can be negotiated even more difficultly. This is, on the one hand, through the new players and infrastructures in the health field (Staii 2018, 202) and, on the other hand, because their concerns are even more obscured by the data-processing processes and technologies. These problems are exacerbated by the lack of public accountability – private enterprises define the problems for research and own the data to provide answers. In other words, neither are the algorithmic procedures transparent nor is there any form of public negotiation of the conventions that structure the framing of the categories, classifications, etc. involved. This is a methodological problem, but at the same time a decisively political issue. Seen in their social contexts, it seems clear that social researchers need to take big data-induced transformations seriously.

## 5. What Kind of Challenge Does Social Research Actually Face?

The developments described above leave little room for doubt: Big data challenges established forms of doing social research. In the following, we argue that this challenge is especially pronounced in relation to the canonic form of predominantly quantitative social research that dominated most of the 20th century. In line with our conceptual framework we maintain that this (still prevalent) post-WWII methodological canon is marked by a specific set of preferred methods, of epistemic values and orientations, and of concomitant quality criteria as well as by typical institutional forms and political-economic contexts. New data worlds put the still dominant canon in question in many of these regards. A comprehensive depiction and appraisal of all the issues involved goes beyond the scope of this article. In the following, we briefly discuss some selected issues that seem particularly relevant for the discussion of the implications of big data and digitization.

The one idea that probably captures the still canonic model of social research most accurately is the notion of the hypothetic-deductive method, also referred to as "the scientific method" from the 1920s onwards (mostly with reference to Popper 2002). Although of course contested and questioned from the outset, this idea served as an organizing principle for methods teaching as well as for designing research projects and for developing/presenting empirical arguments for decades. As an underlying principle, it linked ontological understandings to methodological perspectives and concrete research techniques. For example, the rise and relevance of regression as the main form of statistical analysis is deeply entangled with specific understandings of causality (Vandenberghe 1999). The style of social research that emerged was variable-oriented and heavily focused on measurement as a guiding idea for defining and solving methodological problems (i.e., the valid numeric representation of phenomena that are conceptualized in substantialist terms as always already there, as pre-existing entities or traits). This general understanding yielded concrete forms and procedures for doing social research. To mention only two prominent examples of these preconfigured forms, after WWII null-hypothesis-significance-testing (in short, NHST; Nickerson 2000) and the framework of Total Survey Error (in short, TSE; Weisberg 2005) provided researchers with important guidance for dealing with the ever over-demanding tasks of knowledge production.

Both NHST and TSE mirror specific research logics and favor specific empirical procedures. Under the surface of concrete guidelines, they bridge ontological presuppositions and methodological strategies. But apart from these "classical" epistemological issues, they also reflect particular epistemic values (Putnam 2002) and political-economic contexts (Desrosières 2011). The still

widely used blueprint of null-hypothesis-significance-testing (NHST) is an obvious case in point. To this day, NHST has remained dominant as a mode of doing social research although it was heavily contested within statistics from the beginning and actually never managed to develop into a consensual or coherent framework for designing research or doing statistical analysis (Cohen 1994; Gigerenzer 2004). To understand NHST's stellar career, we need to take the contexts of the time into account. Arguably, what made NHST so attractive was its adequacy in relation to a set of epistemic values that prevailed in quantitative empirical research (Smith 1994): be it a preference for neutrality, transparency, and communicability of procedures and findings in times of political polarization or a predilection for parsimony and elegance paralleling the supposed ideal of the natural sciences (Hossenfelder 2018). NHST also corresponded well to basic epistemic orientations, such as understandings of generalization that focus on statistical representativity, which in turn was used almost synonymously with the idea of a random sample from a target population defined in national terms. NHST and the whole idea of significance and representativeness thus became deeply intertwined with the methodological nationalism that pervaded 20th century social sciences and continue to do so in many fields of research (Wimmer and Glick-Schiller 2002, 2003; Horvath 2012). In other words, NHST might have become and remained influential not in spite of but rather because of its very incoherence: it was and still is far more than just a guide for pure and tidy statistical analysis; NHST has become a ritual (Harlow et al. 1997; Gigerenzer 2004) serving a particular demand that organizes a specific style of doing social research.

Dominant methodological forms such as NHST and their corresponding epistemic values are deeply related to political and economic contexts and, more specifically, to the governmental roles of social research. Desrosières (2008, 2011a, 2011b, 2016) offers a systematic overview of different models of how quantification and statistics are linked to different forms of governing. For example, he contrasts the uses that early mercantilist politics made of statistical endeavors of surveys of populations and territories with the very different 19th century laissez-faire governmental strategy of providing all market players with the information needed for fair and equal terms of competition. Each of these forms of government corresponds to specific forms of economic relations and practices, and they imply certain functions and methods of quantification (of counting and measuring the social). During the 20th century, the social sciences gained massive relevance because they offered means for combining two other forms of making political and economic use of statistics: on the one hand, Keynesian macroeconomics required reliable numbers for indicator-based policy-making (e.g., inflation rates, GDP); on the other hand, the rise of national welfare states led to a demand for social surveys that would allow for monitoring the distribution of incomes, wealth, living conditions, educational opportunities, etc. After WWII, both these areas saw rapid and impressive

developments. The social sciences played a crucial role in providing the statistical toolkit (Savage and Burrows 2007). Methodological innovation and competence were linked to academic social sciences. This constellation of social sciences, economic models, and forms of government was feasible and meaningful, not least because it allowed public accountability and a certain degree of democratic control over social knowledge production. Against this background, the representative population survey and randomized controlled experiments evolved into the two main and emblematic forms of doing social research in academic contexts. Both allowed sustaining the image of social research as value-free, detached, and neutral profession that provides the evidence needed in other, value-based fields of practice, most importantly national politics.

It is important to note that the post-WWII canonic form of social research was tension-ridden and contested long before big data and digitization entered the stage. A few disparate examples illustrate the diversity of issues that triggered debate and critique from the 1970s onwards: the whole blueprint framework of NHST was debated intensely from within the statistical community (e.g., Cohen 1994; Harlow et al. 1997); alternative and more exploratory (and hence inductive) statistical frameworks were put forward (Tukey 1977; Benzècri 1992); the overall logic of social research was disputed early on from rivaling methodological perspectives (one prominent among many examples being Glaser and Strauss 1967); the dichotomization of facts and values has over and over been criticized as unrealistic and unsustainable (Putnam 2002; Putnam and Walsh 2012); the shortcomings of methodological nationalism and its corresponding conceptions of generalization and research to deal with transnational phenomena such as migration became obvious (Wimmer and Glick-Schiller 2002, 2003); and finally relational perspectives have criticized the substantialist thinking underlying mainstream statistical approaches for decades (Emirbayer 1997; Vandenberghe 1999).

Nonetheless, the configuration of challenges that big data poses to this canonic form is striking. Methodologically, inductive strategies, data-driven analysis, and prediction are set against deductive thinking, model-driven approaches, and explanation (Jones 2018). Epistemic values such as transparency and parsimony are replaced by an orientation towards complexity and a tolerance for or even appreciation of opacity – as long as the algorithm works (Dourish 2016). The main institutional context of methodological innovation and competence has moved from academic social sciences to the private business world (van Dijck 2014; Diaz-Bone 2019; Zuboff 2019). The role of data and quantitative information has shifted from informing (national) policy-making to using numbers for benchmarking, individualized surveillance, and pattern prediction (Zuboff 2019). Ownership of data and accountability for defining and doing research have moved from the public sector, and so on. Any effective critique of current data worlds must take this all-encompassing rupture seriously. It does not suffice to critically assess the quality of big data

because the very basis for this assessment is put into question. What is needed is an epistemological reconfiguration that lives up to transformed social realities and corresponds to new data worlds. The task is to challenge current big-data realities by urging a move from a-theoretical prediction and affirmative algorithms to exploratory research and reflexive methodologies.

## 6. Conclusion: Contours of and Starting Points for a Sociology of Social Research and Data Quality in Times of Big Data

New data worlds change social relations, the problems of data quality, and transform the power relations, which are in control of data production and data analysis – social research therefore needs to engage with them. The link between data quality and the control of data infrastructures (statistical chains) cannot be dissolved and needs to be addressed beside narrow methodological definitions of "quality." The epistemological foundations of social research (when relying, e.g., on Twitter data or Facebook data), is neither controlled by researchers nor are they fully transparent to them. Methods teaching in times of big data has to teach about the potentials and deficiencies of big data but also about Internet platforms and the politics of data infrastructures and address the need for publicly controlled data infrastructures that implement coherent statistical chains, which are based on such conventions for measurement, for analysis and interpretation of data, for data distribution and public representation, in a way that legitimate and publicly deliberated common goods can be pursued and achieved. Here the perspective of EC contributes to the collapse of the fact/value dichotomy because EC emphasizes the normative basis of data production and data quality assessment, which is reduced in mainstream methodology to technical aspects only. The argument itself is an empirical one: actors in society such as citizens, NGOs, governments, academics, and enterprises are regarded as competent in applying convention as normative order of justification and situations require conventions "as normative equipment" to offer meaning, orientation and benchmarks for evaluation. Research and teaching (not only) in the field of big data can also profit from acknowledging the factual plurality of normative orders in social research as a more realist and more radical empirical approach to social research. In this sense, one can speak of a neopragmatist sociology of social research. This is a new approach (but for reinventing classical pragmatist positions, see Dewey 1938) and a contribution of convention theory (EC) to the reflection on data quality, which can be regarded as part of the ongoing "evolution of data quality" (see Keller et al. 2017). Research of and teaching about *big data analytics* (doing and training the analysis of big data sets) cannot be separated from research of and teaching

about the analysis of big data conceived as the political economy of quantification. To conceive the plurality of conventions (as normative orders) as the starting point and as the inevitable foundation of coordination in the linked situations of big data analytics also opens new perspectives to address (1) questions of methodological strategies such as exploratory data analysis (including visualizing), linking big data (analysis) to established social research data formats (such as survey data), combining qualitative digital research methods with statistical tools (Lupton 2015; Marres 2017), and (2) questions of grounding research questions in the pursuit of collective goals and the enabling of achieving a common good with data bases and forms of data analyses that are coherent with this collective intentionality. Also (3) the need for theory in research and teaching is a core issue. The forms of knowledge that emerge from current big data perspectives are marked by a-theoretic pattern recognition and automated decision-making tools – affirmative algorithms, artificial intelligence, and data-driven predictions that threaten to reinforce existing patterns of discrimination and orders of inequality. These forms of knowledge production by and large happen without any form of public ownership or democratic control. The overarching objective hence must be to regain authority and to promote forms of theory building and exploratory research that are capable of triggering true social innovation, not least because they can build on decades of existing social science discourse. (4) Related to these points is the importance of elaborating specific forms of designs and methodological strategies as is argued by Halford and Savage (2017), who invented the notion of "symphonic social sciences" to characterize some new modes for working with heterogeneous data. They refer to some exemplary studies:

> These are all, fundamentally, "data-books". Each deploys large-scale heterogeneous data assemblages, re-purposing findings from numerous and often asymmetrical data sources – rather than a dedicated source, such as a national representative sample or an ethnographic case study. These works build on earlier traditions of comparative analysis, using strictly comparable forms of data […] but are innovative in the use of far more diverse data sources to make their comparative points. (Halford and Savage 2017, 1135)

(5) A last point is to be made about the consequences for a sociology of social research.[19] To understand and to evaluate big data, its quality, and its production, one has to replace the established epistemological models. Traditionally, epistemology has focused on single persons or groups of academics. Nowadays, data production is a distributed process, including different kinds of intermediaries (objects, technologies, or persons), integrating different time frames (real-time analysis, but also possibilities for historical analysis) and a

[19] For a first sketch, see Leahey (2008), but his contribution does not include digitalization or big data.

high degree of division of labor.[20] Big data is linked to data infrastructures, which are (in most cases) not built up and are not controlled by social researchers; the quality assessment of big data calls for a new sociology (of this new form) of social research. This sociology of social research would not be restricted to academic researchers, research institutions, or the "science system" but instead include all the statistical chains, situations, intermediaries, and conventions that are connected in the production of big data – for its formatting, filtering, distribution, storing, aggregating, analyzing, etc. It would not only focus on the situational practices and its coherences but also tensions and contradictions all along the chain, on the entangled values and common goods, to which collective intentionality is oriented. This sociology of social research would track the practices, which bring in meaning, interpretation, and value to big data as a new kind of social resource and would study the effect on big data production but also its effects on society. Big data is *proceeding in society* but step by step pervading, representing, analyzing, structuring, and *proceeding society*; sociology of social research would be a critical, pragmatist, and reflexive methodology of distributed knowledge production in a digital society. This way the wrong dichotomy of utopian or dystopian perspectives on big data should be avoided and social research should improve its societal agency.

## References


Abbott, Andrew. 1988. *The system of professions. An essay on the division of expert labor*. Chicago: Chicago University Press.

Abouelmehdi, Karim, Abderrahim Beni-Hessane, and Hayat Khaloufi. 2018. Big healthcare data. Preserving security and privacy. *Journal of Big Data* 5 (1): 1-18. doi: 10.1186/s40537-017-0110-7.

Alvarez, R. Michael, ed. 2016. *Computational social science*. Cambridge: Cambridge University Press.

Anderson, Chris. 2008. The end of theory. The data deluge makes the scientific method obsolete. *Wired*. <https://www.wired.com/2008/06/pb-theory/>

Bartlett, Andrew, Jamie Lewis, Luis Reyes-Galindo, and Neil Stephens. 2018. The Locus of Legitimate Interpretation in Big Data Sciences: Lessons for Computational Social Science from -Omic Biology and High-Energy Physics. *Big Data & Society* 5 (1). doi: 10.1177/2053951718768831.

Batifoulier, Philippe, Jean-Paul Domin, and Maryse Gadreau. 2011. Market empowerment of the patient. The French experience. *Review of Social Economy* 69 (2): 143-162.


[20] Edwin Hutchins (1995) invented the notion of distributed cognition to grasp the complex procedure of navigation done in a US marine ship. In a similar way, big data analysis can be portrayed as a complex and distributed analysis, which has itself to be the object of analysis to assess its epistemological properties, the ontologies and politics of data, and its resulting qualities.

Batifoulier, Philippe, Franck Bessis, Ariane Ghirardello, Guillemette de Larquier, and Delphine Remillon, eds. 2016. *Dictionnaire des conventions*. Villeneuve-d'Ascq: Presses Universitaires du Septentrion.
Batifoulier, Philippe, Nicolas Da Silva, and Jean-Paul Domin. 2018. *Economie de la santé*. Paris: Armand Colin.
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. *Historical Social Research* 34 (3): 9-50. doi: 10.12759/hsr.34.2009.3.9-50.
Benzècri, Jean-Paul. 1992. *Correspondence analysis handbook*. New York: Paul Dekker.
Boltanski, Luc, and Eve Chiapello. 2005. *The new spirit of capitalism*. New York: Verso Books.
Boltanski, Luc, and Laurent Thévenot. 2006. *On justification. Economies of worth*. Princeton: Princeton University Press.
Bowker, Geoffrey C., and Susan L. Star. 2000. *Sorting things out. Classification and its consequences*. Cambridge: MIT Press.
Boyd, Danah, and Kate Crawford. 2012. Critical questions for big data. *Information, Communication and Society* 15 (5): 662-679.
Burrows, Roger, and Mike Savage. 2014. After the crisis? Big data and the methodological challenges of empirical sociology. *Big Data & Society* 1 (1): 1-6.
Callon, Michel. 1998. Introduction: the embeddedness of economic markets in economics. In *The laws of the markets*, ed. Michel Callon, 1-57. Oxford: Blackwell Publishers.
Cohen, Jacob. 1994. The earth is round ($p < .05$). *American Psychologist* 49 (12): 997-1003.
Da Silva, Nicolas. 2018. L'industrialisation de la médicine libérale: une approche par L'Économie des conventions. *Management & Avenir Santé* (1) 3: 13-30. doi: 10.3917/mavs.003.0013.
Desrosières, Alain. 2000. Measurement and its uses. Harmonization and quality in social statistics. *International Statistical Review* 68 (2): 173-187.
Desrosières, Alain. 2008. *Pour une sociologie historique de la quantification: L'Argument statistique I*. Presses des Mines. doi: 10.4000/books.pressesmines.901.
Desrosières, Alain. 2002. *The politics of large numbers. A history of statistical reasoning*. Cambridge: Harvard University Press.
Desrosières, Alain. 2011. Worlds and numbers. For a sociology of the statistical argument. In *The mutual construction of statistics and society*, ed. Ann Rudinow Saetnan, Heidi Mork Lomell, and Svein Hammer, 41-63. New York: Routledge.
Desrosières, Alain. 2015. Retroaction. How indicators feed back onto quantified actors. In *The world of indicators*, ed. Richard Rottenburg, Sally Engle Merry, Sung-Joon Park, and Johanna Mugler, 329-353. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781316091265.013.
Desrosières, Alain. 2016. The quantification of the social sciences. An historical comparison. In *The social sciences of quantification*, ed. Isabelle Bruno, Florence Jany-Catrice, and Béatrice Touchelay, 183-204. Cham: Springer. doi: 10.1007/978-3-319-44000-2_15.

Desrosières, Alain, and Laurent Thévenot. 1979. Les mots et les chiffres: les nomenclatures socioprofessionnelles. *Economie et statistique* 110: 49-65.
Dewey, John. 1938. *Logic. The theory of inquiry*. New York: Henry Holt.
Diaz-Bone, Rainer. 2016. Convention theory, classification and quantification. *Historical Social Research* 41 (2): 48-71. doi: 10.12759/hsr.41.2016.2.48-71.
Diaz-Bone, Rainer. 2017. Classifications, quantifications and quality conventions in markets – Perspectives of the economics of convention. *Historical Social Research* 42 (1): 238-262. doi: 10.12759/hsr.42.2017.1.238-262.
Diaz-Bone, Rainer. 2018. *Die Economie des conventions. Grundlagen und Entwicklungen der neuen französischen Wirtschaftssoziologie*. 2nd ed. Wiesbaden: Springer VS.
Diaz-Bone, Rainer. 2019. Convention theory, surveys and moral collectives. In: *Moralische Kollektive. Theoretische Grundlagen und empirische Einsichten*, ed. Stefan Joller and Marija Stanisavljevic, 115-135. Wiesbaden: Springer VS.
Diaz-Bone, Rainer, and Emmanuel Didier, eds. 2016. Conventions and quantification – Transdisciplinary perspectives on statistics and classifications (special issue). *Historical Social Research* 41 (2). <https://www.gesis.org/hsr/archiv/2016/412-conventions-and-quantification>.
Diaz-Bone, Rainer, and Emmanuel Didier. 2016. The sociology of quantification – Perspectives on an emerging field in the social sciences. *Historical Social Research* 41 (2): 7-26. doi: 10.12759/hsr.41.2016.2.7-26.
Diaz-Bone, Rainer, and Olivier Favereau, eds. 2019. Markets, organizations, and law – Perspectives of convention theory on economic practices and structures. *Historical Social Research* 44 (1). <https://www.gesis.org/hsr/aktuelle-hefte/2019/441-markets-organizations-and-law> (Accessed 9 May 2020).
Diaz-Bone, Rainer, and Robert Salais, eds. 2011. Conventions and institutions from a historical perspective (Special Issue). *Historical Social Research* 36 (4). <http://www.gesis.org/hsr/archiv/2011/364-conventions-institutions>
Diaz-Bone, Rainer, and Robert Salais, eds. 2012. The Économie des Conventions – Transdisciplinary discussions and perspectives (HSR-focus). *Historical Social Research* 37 (4). doi: 10.12759/hsr.37.2012.4.9-14.
Diaz-Bone, Rainer, Claude Didry, and Robert Salais, eds. 2015. Law and conventions from a historical perspective (special issue). *Historical Social Research* 40 (1) <https://www.gesis.org/hsr/archiv/2015/401-law-and-conventions>.
Didier, Emmanuel. 2018. Globalization of quantitative policing. Between management and statactivism. *Annual Review of Sociology* 44: 515-534.
Dourish, Paul. 2016. Algorithms and their others. Algorithmic culture in context. *Big Data & Society* 3 (2): 1-11.
Ebeling, Mary F. E. 2016. *Healthcare and big data. Digital specters and phantom objects*. New York: Palgrave Macmillan.
Einav, Liran, and Jonathan Levin. 2014. Economics on the age of big data. *Science* 346: 715-721.
Emirbayer, Mustafa. 1997. Manifesto for a relational sociology. *American Journal of Sociology* 103 (2): 281-317. doi: 10.1086/231209.
Eubanks, Virginia. 2017. *Automating inequality*. New York: St. Martin's Press.
Eymard-Duvernay, François, ed. 2006a. *L'économie des conventions. Méthodes et résultats, vol. 1: Débats*. Paris: La Découverte.

Eymard-Duvernay, François, ed. 2006b. *L'économie des conventions. Méthodes et résultats, vol. 2: Développements*. Paris: La Découverte.
Foster, Ian, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter, and Julia Lane, eds. 2017. *Big data and social science. A practical guide to methods and tools*. Boca Raton: CRC Press.
Foucault, Michel. 1995. *Discipline and punish. The birth of the prison*. New York: Vintage Books.
Gehring, Petra. 2018. Forschungsinfrastrukturen. Ein Flussbett für Datenströme. *Frankfurter Allgemeine Zeitung* 19 December. <https://www.faz.net/aktuell/feuilleton/hoch-schule/forschungsinfrastrukturen-ein-flussbett-fuer-die-datenstroeme-15946883.html> (Accessed 9 May 2020).
Gigerenzer, Gerd. 2004. Mindless statistics. *Journal of Socio-Economics* 33 (5): 587-606. doi: 10.1016/j.socec.2004.09.033.
Gillespie, Tarleton. 2010. The politics of "platforms". *New Media & Society* 12 (3): 347-364.
Glaser, Barney G., and Anselm L. Strauss. 1967. *The discovery of Grounded Theory. Strategies for qualitative research.* Chicago: Aldine Publishing Company
Gray, Jonathan, Carolin Gerlitz, and Liliana Bounegru. 2018. Data infrastructure literacy. *Big Data & Society* 5 (2): 1-13.
Halford, Susan, and Mike Savage. 2017. Speaking sociologically with big data. Symphonic social science and the future for big data research. *Sociology* 51 (6): 1132-1148.
Hampton, Keith N. 2017. Studying the digital. Directions and challenges for digital methods. *Annual Review of Sociology* 43: 167-188.
Harlow, Lisa L., Stanley A. Mulaik, and James H. Steiger, eds. 1997. *What if there were no significance tests?* Mahwah: Lawrence Erlbaum.
Hogle, Linda F. 2016. Data-intensive resourcing in healthcare. *BioSocieties* 11(3): 72-93.
Horvath, Kenneth. 2012. National numbers for transnational relations? Challenges of integrating quantitative methods into research on transnational labour market relations. *Ethnic and Racial Studies* 35 (10): 1741-1757. doi: 10.1080/01419870.2012.659270.
Horvath, Kenneth. 2018. Fixed Narratives and Entangled Categorizations. Educational Problematizations in Times of Stratified and Politicized Migration. *Social Inclusion* 6(3): 237-247.
Horvath, Kenneth. 2019. Migration background – Statistical classification and the problem of implicitly ethnicising categorisation in educational contexts. *Ethnicities* 19(3). doi: 10.1177/1468796819833432.
Hossenfelder, Sabine. 2018. *Lost in math. How beauty leads physics astray*. New York: Basic Books.
Hutchins, Edwin. 1995. *Cognition in the wild*. Cambridge: MIT Press.
Israni, Ellora Thadaney. 2017. When an algorithm helps send you to prison. *New York Times* 26 October. <https://www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html> (Accessed 9 May 2020).
Japec, Lilli, Frauke Kreuter, Marcus Berg, Paul Biemer, Paul Decker, Cliff Lampe, Julia Lane, Cathy O'Neil, and Abe Usher. 2015. Big data in survey research. AAPOR task force report. *Public Opinion Quarterly* 79 (4): 839-880.

Jeese, Norbert. 2018. Internet of things and big data. The disruption of the value chain and the rise of new software ecosystems. *AI and Society* 33: 229-239.
Jones, Matthew L. 2018. How we became instrumentalists (again). Data positivism since World War II. *Historical Studies in the Natural Sciences* 48 (5): 673-684. doi: 10.1525/hsns.2018.48.5.673.
Keller, Sallie, Gizem Korkmaz, Mark Orr, Aaron Schroeder, and Stephanie Shipp. 2017. The evolution of data quality. Understanding the transdisciplinary origins of data quality concepts and approaches. *Annual Review of Statistics and Its Application* 4: 85-108.
Kitchin, Rob. 2014. Big data, new epistemologies and paradigm shifts. *Big Data & Society* 1 (1): 1-12.
Kitchin, Rob. 2017. Thinking critically about and researching algorithms. *Information, Communication and Society* 20 (1): 14-29.
Kleiner, Brian, Alexandra Stam, and Nicolas Pekari. 2015. Big data for the social sciences. *FORS Working Papers* 2015-2. Lausanne: FORS.
Lagoze, Carl. 2014. Big data, data integrity, and the fracturing of the control zone. *Big Data & Society* 1 (2): 1-11.
Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. Computational social science. *Science* 323: 721-723.
Lazer, David, and Jason Radford. 2017. Data ex machine. Introduction to big data. *Annual Review of Sociology* 43: 19-39.
Leahey, Erin. 2008. Methodological memes and mores. Toward a sociology of social research. *Annual Review of Sociology* 34: 33-53.
Le Roux, Brigitte, and Henri Rouanet. 2010. Multiple correspondence analysis. Los Angeles: Sage.
Liang, Fan, Wie Yu, Dou An, Qingyu Yang, Xinwen Fu, and Wie Zhao. 2018. Survey on big data market. Pricing, trading and protection. *IEEE Access* 6: 15132-15154. doi: 10.1109/ACCESS.2018.2806881.
Lupton, Deborah. 2015. *Digital sociology*. London: Routledge.
Marres, Noortje. 2017. *Digital sociology*. London: Polity Press.
Mayer-Schönberger, and Thomas Ramge. 2018. *Reinventing capitalism in the age of big data*. New York: Basic Books.
Mayer-Schönberger, Viktor, and Kenneth Cukier. 2013. *Big data. A revolution that will transform how we live, work, and think*. Boston: Houghton Mifflin Harcourt.
Munzert, Simon, Christian Rubba, Peter Meißner, and Dominic Nyhuis, eds. 2015. *Automated data collection with R. A practical guide to web scraping and text mining.* New York: Wiley.
Nickerson, Raymond S. 2000. Null hypothesis significance testing. A review of an old and continuing controversy. *Psychological Methods* 5 (2): 241-301. doi: 10.1037//1082-989X.5.2.241.
Nafus, Dawn, ed. 2016. *Quantified. Biosensing technologies in everyday life*. Cambridge: MIT Press.
Nissenbaum, Helen, and Heather Patterson. 2016. Biosensing in context. Health privacy in a connected world. In *Quantified. Biosensing technologies in everyday life*, ed. Dawn Nafus, 78-100. Cambridge: MIT Press.

Noble, Safiya Umoja. 2018. *Algorithms of oppression. How search engines reinforce racism*. New York: New York University Press.
Novotny, Rudi. 2015. Das Orangen-Projekt. Der Amerikaner Thomas Hales löste eines der größten mathematischen Rätsel. Das veränderte sein Leben – und die Mathematik. *Die ZEIT* 27. <https://www.zeit.de/2015/27/mathematik-thomas-hales-computer-algorithmus-johannes-kepler-kuenstliche-intelligenz> (Accessed 9 May 2020).
O'Neil, Cathy. 2016. *Weapons of math destruction. How big data increases inequality and threatens democracy*. New York: Penguin Books.
Pantzar, Mika, and Minna Ruckenstein. 2017. Living the metrics: Self-tracking and situated objectivity. *Digital Health*. doi: 10.1177/2055207617712590.
Pentland, Alex. 2014. *Social physics. How good ideas spread. The lessons from a new science*. New York: Penguin Books.
Popper, Karl R. 2002. *The logic of scientific discovery*. London: Routledge.
Putnam, Hilary. 2002. *The collapse of the fact/value dichotomy and other essays*. Cambridge: Harvard University Press.
Putnam, Hilary, and Vivian Charles Walsh, eds. 2012. *The end of value-free economics*. London: Routledge.
Ruckenstein, Minna, and Natasha Dow Schüll. 2017. The datafication of health. *Annual Review of Anthropology* 46: 261-278.
Salais, Robert. 2012. Quantification and the economics of convention. *Historical Social Research* 37 (4): 55-63. doi: 10.12759/ hsr.37.2012.4.55-63.
Salais, Robert. 2016. Quantification and objectivity. From statistical conventions to social conventions. *Historical Social Research* 41 (2): 118-134. doi: 10.12759/ hsr.41.2017.1.118-134.
Salganik, Matthew J. 2018. *Bit by bit. Social research in the digital age*. Princeton: Princeton University Press.
Savage, Mike, and Roger Burrows. 2007. The coming crisis of empirical sociology. *Sociology* 41 (5): 885-899.
Sharon, Tamar. 2016. The Googlization of health research. From disruptive innovation to disruptive ethics. *Personalized Medicine* 13(6): 563-574.
Sharon, Tamar. 2017. Self-Tracking for health and the Quantified Self. Re-articulating autonomy, solidarity, and authenticity in an age of personalized healthcare. *Philosophy & Technology* 30(1): 93-121.
Sharon, Tamar. 2018. When digital health meets digital capitalism, how many common goods are at stake? *Big Data & Society* 5 (2): 1–12.
Smith, Mark C. 1994. Social sciences in the crucible. The American debate over objectivity and purpose, 1918-1941. Durham: Duke University Press.
Staii, Adrian. 2018. Connected health. Between common aspirations and specific interests. In *Confidence and legitimacy in health information and communication vol. 1*, ed. Paganelli, Céline, 195-221. New York: Wiley.
Storper, Michael, and Robert Salais. 1997. *Worlds of production. The action frameworks of the economy*. Cambridge: Harvard University Press.
Thévenot, Laurent. 1983. L'économie du codage social. *Critique de l'économie politique* 23/24: 188-222.
Tukey, John. 1997. *Exploratory data analysis*. Reading: Addison-Wesley.

Vandenberghe, Frederic. 1999. The real is relational. An epistemological analysis of Pierre Bourdieu's generative structuralism. *Sociological Theory* 17 (1): 32-67. doi: 10.1111/0735-2751.00064.
van Dijck, José. 2014. Datafication, dataism and dataveillance. Big data between scientific paradigm an ideology. *Surveillance & Society* 12 (2): 197-208.
van Dijck, José, and Thomas Poell. 2016. Understanding the promises and premises of online health platforms. *Big Data & Society* 3(1): 1-3. doi: 10.1177/20539517 16654173.
Wadman, Sarah, and Klaus Hoeyer. 2018. Dangers of the digital fit. Rethinking seamlessness and social sustainability in data-intensive healthcare. *Big Data & Society* 5 (1): 1-13.
Weisberg, Herbert F. 2005. *The total survey error approach. A guide to the new science of survey research*. Chicago: University of Chicago Press.
Wiedemann, Gregor. 2013. Opening up to big data. Computer-assisted analysis of textual data in social sciences. *Historical Social Research* 38 (4): 332-358. doi: 10.12759/hsr.38.2013.4.332-358.
Wierse, Andreas, and Till Riedel. 2017. *Smart data analytics. Zusammenhänge erkennen, Potentiale nutzen und Big Data verstehen*. Berlin: De Gruyter.
Wimmer, Andreas, and Nina Glick-Schiller. 2002. Methodological nationalism and the study of migration. *European Journal of Sociology* 43 (2): 217-240. doi: 10.1017/S000397560200108X.
Wimmer, Andreas, and Nina Glick-Schiller. 2003. Methodological nationalism, the social sciences, and the study of migration. An essay in historical epistemology. *International Migration Review* 37 (3): 576-610.
Zuboff, Shoshana. 2019. *The age of surveillance capitalism. The fight for a human future at the new frontier of power*. New York: Public Affairs.

Historical Social Research
Historische Sozialforschung

All articles published in this Forum:

Nina Baur, Peter Graeff, Lilli Braunisch & Malte Schweia
The Quality of Big Data. Development, Problems, and Possibilities of Use of Process-Generated Data in the Digital Age.
doi: 10.12759/hsr.45.2020.3.209-243

Peter Graeff & Nina Baur
Digital Data, Administrative Data, and Survey Compared: Updating the Classical Toolbox for Assessing Data Quality of Big Data, Exemplified by the Generation of Corruption Data.
doi: 10.12759/hsr.45.2020.3.244-269

Gertraud Koch & Katharina Kinder-Kurlanda
Source Criticism of Data Platform Logics on the Internet.
doi: 10.12759/hsr.45.2020.3.270-287

Martin Weichbold, Alexander Seymer, Wolfgang Aschauer & Thomas Herdin
Potential and Limits of Automated Classification of Big Data – A Case Study.
doi: 10.12759/hsr.45.2020.3.288-313

Rainer Diaz-Bone, Kenneth Horvath & Valeska Cappel
Social Research in Times of Big Data. The Challenges of New Data Worlds and the Need for a Sociology of Social Research.
doi: 10.12759/hsr.45.2020.3.314-341

Michael Weinhardt
Ethical Issues in the Use of Big Data for Social Research.
doi: 10.12759/hsr.45.2020.3.342-368