Accessibility statement

This is an accessibility statement for the journal: Encounters.

Conformance status

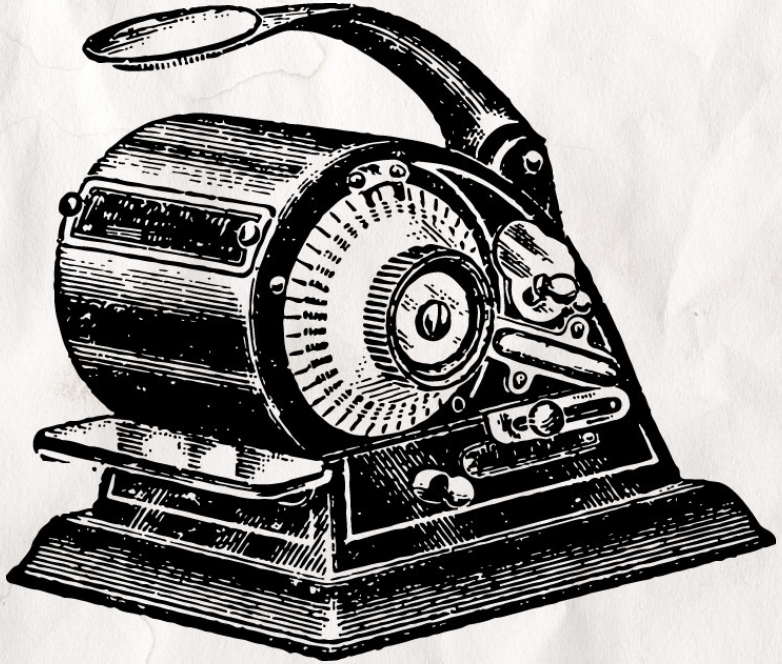
The Web Content Accessibility Guidelines (WCAG) defines requirements for designers and developers to improve accessibility for people with disabilities. It defines three levels of conformance: Level A, Level AA, and Level AAA. This statement is relevant for volume 10, number 5, 2018 through volume 12, number 1, 2021. Encounters is partially conformant with WCAG 2.1 level AA. Partially conformant means that some parts of the content do not fully conform to the accessibility standard. Despite our best efforts to ensure accessibility, footnotes and graphs may not be accessible for screen readers at this point in time.

Feedback

We welcome your feedback on the accessibility of the journal. Please let us know if you encounter accessibility barriers. You can reach us at:

E-mail: imvko@cc.au.dk
Address: STS Center, Helsingforsgade 14, 8200 Aarhus N

ENGAGING THE DATA MOMENT



Special issue
Volume 11• Number 1 • 2020

STS Encounters

SPECIAL ISSUE

Volume 11• Number 1• 2020

The rhetorical work of credibility-building for social scientific big data: Positioning arguments and legitimacy in empirical sociology

Juho Pääkkönen,
Department of Sociology, University of Helsinki, Finland
Computer Science Department, Aalto University, Finland

DASTS is the primary academic association for STS in Denmark. Its purpose is to develop the quality and breadth of STS research within Denmark, while generating and developing national and international collaboration.

Abstract

This article investigates the rhetorical work of building credibility for social scientific research designs with big data. Big data is discussed as a contested concept in the social sciences, one whose meaning and implications are under dispute. Proceeding from analysis of 29 sociology articles based on empirical research, the author argues that credibility is constructed in this context through the rhetorical positioning of disciplines as legitimate interpreters of big data. The article identifies three distinct positioning strategies: conservative, reformist, and supplementarist, each of which locates the legitimacy of interpretation in its own way. While conservative positioning fixes the locus of legitimate interpretation within the social sciences, those employing a reformist strategy seek to widen it to encompass methods from beyond established social scientific fields. Finally, supplementarist positioning portrays big data as inherently limited and ties the legitimacy of interpretation to alternative approaches. Through identifying and addressing these respective strategies, the article discusses rhetorical positioning as part of the work of enacting big data: a performative process that can foster several visions of the future methodology of the social sciences.

Keywords: Big data; Credibility; Rhetorical positioning; Locus of legitimate interpretation; Empirical sociology

Introduction

Over the past decade, the phenomenon known as 'big data' has received increasing attention in the social sciences (Manovich, 2012; Youtie et al., 2017), most often being characterized as involving high-volume, high-velocity data of varying structure (Kitchin and McArdle, 2016). However, in the social sciences, the notion commonly refers to *digital data* produced in the intertwined processes of digitalization and datafication (Van Dijck, 2014), particularly through human interaction in various digital environments (Lazer and Ratford, 2017). Examples include social media data, Web searches, blog posts, digital administrative records, and digitized texts. The proliferation of these data has inspired much enthusiasm in the social sciences, with big data being heralded as a revolution comparable to the invention of the telescope in astronomy (Watts, 2011: 266).

However, critics have recently argued that, as a phenomenon, big data not only consist of proliferating new data sources, but also involve a prevailing *rhetoric*, which works to rationalize computational methodology (e.g., boyd and Crawford, 2012; Kennedy 2016; Kennedy and Hill, 2018). For instance, Kitchin (2014: 113) proclaims the phenomenon to have given rise to a pervasive discourse that "provides the rationale for adopting new ideas and technologies, and legitimates their development and implementation". The worry is that the legitimating function of big data can privilege those with the resources needed for computational knowledge production while excluding others (Couldry, 2014). On the other hand, it has been argued that the rhetoric surrounding big data is business-driven in nature (Elish and boyd, 2018), while the extent of its influence in other contexts remains unclear. Ultimately, the legitimating function of big data rhetoric should not be taken for granted, particularly in academic research; rather, it constitutes an issue for investigation.

In this article, I examine the concept of big data in the context of the social sciences, where the notion has been caught up in debates pertaining to the future methodology of social research. I draw on a dataset of 29 empirical research articles in sociology to investigate the following research question: how do authors argue for the credibility of their research designs with big data? By 'research design' I mean the overall strategy through which the authors co-ordinate their data collection and methodology for the purposes of tackling their research problems. Focusing on cases wherein the use of big data is problematized, I analyse the set of articles to identify the conceptions of big data they display and how these are used to argue for certain notions related to credible social scientific research.

My theoretical foundation builds on recent work in science and technology studies (STS) by Bartlett et al. (2018), who suggest that key problems with exploiting big data in the social sciences are connected to the difficulty of establishing the legitimacy of a social scientific interpretation of data that were not originally generated for social scientific purposes. Indeed, the *locus of legitimate interpretation* (Collins and Evans, 2007) of big data often seems to reside outside the social sciences altogether. Working with these ideas, I analyse arguments for the credibility of research with big data as attempts at *rhetorically positioning* (Harré and Langenhove, 1998) particular disciplines – for instance, the social sciences or computer science – as legitimate interpreters of big data such that one may credibly draw on their methodological practices. I identify three distinct argumentation strategies, which I term the *conservative*, *reformist*, and *supplementarist* positionings, each of which locates the legitimacy of interpretation in its own specific manner. From this perspective, I argue that the concept of big data serves as an argumentation setting, within which the boundaries of credible social scientific knowledge are negotiated.

My focus on empirical sociology as a case is motivated by recent calls for sociologists to rethink their methodology in the age of big data (e.g., Burrows and Savage, 2014). Without doubt, sociology is not alone as a field in dealing with the problem of incorporating novel data and computational methods (e.g., Grimmer, 2015; Wallach, 2018). However, empirical sociology represents a clear case wherein attempts at building credibility for big data can be expected to be visible. This article presents an analysis of such attempts, and how they are constituted rhetorically via positioning arguments.

I begin by introducing recent methodological debate about big data in the social sciences (Section 2), then move on to discussing my theoretical approach in more detail (Section 3). Against that backdrop, I present my empirical material (Section 4) and analysis (sections 5–8). I conclude the paper by discussing them in relation to critical accounts of big data (Section 9).

Big data as a contested concept in the social sciences

Previous research into big data as a conceptual phenomenon has emphasized that data always "need to be imagined as data to exist and to function as such" (Gitelman and Jackson, 2013: 3). Under this principle, using certain objects as data involves an act of interpreting them as useful for accomplishing certain analytical purposes (Bowker, 2013; Pentzold and Fischer, 2017: 2). For instance, Puschmann and Burgess (2014: 1691) argue in this vein that the various analytics technologies associated with big data are "still in a period of interpretative flexibility and ongoing contestation over their exact meanings and values" (see also Stevens et al., 2018). These studies demonstrate that, in a given context, big data can encompass a host of conceptions, which compete with each other to be the dominant interpretation of data and methods.

This contestation over the meaning of big data is apparent in the social sciences, where the notion has been associated with both high hopes of epistemic import and scepticism. It has been argued that technologies that generate digital data have a transformative effect on social research, due to both their flexibility and the wealth of data generated (Given, 2006). Digital traces accumulate in near-real time on platforms such as social media, and are thought to yield highly granular information about user activities without researcher intervention (Golder and Macy, 2014). While social scientists have been sceptical about these data supplanting traditional theory-driven methodology (Bowker, 2014; Rouvroy, 2013), they are regarded as important in *augmenting* or *reorienting* research by providing additional sources of information and by inspiring theories about action in novel settings (Edwards et al., 2013).

The enthusiasm notwithstanding, engaging with big data in the social sciences has proved challenging. More than a decade ago, Savage and Burrows (2007; 2009) famously argued that empirical sociology was facing a crisis arising from the field's slow reaction to rapidly proliferating commercial digital data sources and analytics.

Consequently, they implored sociologists to "intervene in the world of Big Data in order to ensure we command a voice in this new terrain" (Burrows and Savage 2014: 5). How exactly this intervention should be accomplished has become a matter of some debate (see, for example, Frade 2016 for a critique). Crucially, as Halavais (2015) argues, the difficulty of bringing big data to bear on social research does not lie in the scale involved, given sociologists' long history of expertise in analysing large datasets from sources such as administrative registers (Beer, 2016; Connelly et al., 2016; Hacking, 1991). Rather, the crux of the issue is that digital data and computational methods are *novel* for social scientists and lack clearly established use practices (Halavais, 2015: 586). For the proponents of big data, the central challenge lies in developing methodology and practices that credibly render them sources of social scientific evidence (Halavais, 2015: 591–592; Halford and Savage, 2017: 1138).

Accomplishing this involves a host of problems. As Halford and Savage (2017) note, sociologists have been profoundly sceptical about the value of big data, arguing that digital traces offer only part of the picture, without providing the contextual information that is vital for evaluating their validity (boyd and Crawford, 2012). Traditional methods such as surveys and interviews are still regarded as the gold standard of data generation (Crompton, 2008; Edwards et al., 2013), while computational methodology is criticized for relying on misguided conceptions of naturally occurring digital traces (Törnberg and Törnberg, 2018). Furthermore, digital data often exist in complex structures, necessitating methods such as machine learning, which lie outside the skill set of most social scientists (Goldberg, 2015; King, 2016; see Salganik, 2017 for work towards developing expertise in these areas). Social scientists are not typically trained in programming, which is an essential skill for critically engaging with the limitations of algorithmic data production and analysis (Gillespie, 2014; Halavais, 2015). One proposed solution is to encourage collaboration with computational scientists and data analysts (Halford and Savage, 2017). However, it remains unclear what form such collaboration should take, not least because differences in methodological paradigms complicate communication between social and computational scientists (McFarland et al., 2016).

These arguments from sceptics resonate with broader criticisms about the role of the digital in social research. As Ruppert and colleagues (2013) argue, digital platforms have the dual role of enabling social activities *while* generating data on them, and therefore their use necessitates a reflexive understanding of this simultaneous process of shaping and tracing action. Indeed, scholars in the digital methods literature (Rogers, 2013; Venturini et al., 2019) argue that researchers interested in the digital should learn to repurpose tools from these environments for social scientific purposes – a process that could involve, for instance, the development of collectively scrutinizable methods that facilitate transparent research processes on proprietary platforms (Venturini and Latour, 2010). However, as Marres (2017) has noted, such critical engagement also means that researchers must refine their methodological traditions so that they consistently mesh well with digital devices such as search engines, social media applications, and software for computational analysis. Ultimately, doing so could lead to reorienting the practices of social research towards increasing inclusion of actors and processes external to the context of academic social science (Marres, 2012).

As this critque indicates, the endeavour to take advantage of big data depends on more than merely building infrastructure for data access. If social scientists are to use big data credibly, they need to articulate how the methodological practices they adopt make sense in relation both to existing practices in their fields and to problems associated with novel data and computational methods. In the discussion that follows, I argue that how this is accomplished hinges on whether the social sciences are conceived of as *legitimate interpreters* of big data. In the next section, I present my theoretical approach for analysing attempts to establish the credibility of big data as positioning arguments.

Rhetorical positioning and the legitimacy of interpretation of big data

Bartlett et al. (2018) have recently suggested that the notion of the *locus of legitimate interpretation* from the STS literature (Collins and Evans, 2007) offers a way to understand problems with exploiting big data in the social sciences. In particular, they argue that, since most big data in the social sciences are *found data* – data produced "independently of the intent and design of the scientific community doing the analysis" (Bartlett et al., 2018: 4) – the social sciences face difficulties in claiming authority in interpreting them. This situation contrasts with contexts such as physics and biology, in which academic researchers generate their own data, consequently commanding exclusive authority over their interpretation. The locus of legitimate interpretation of big data, or the "location" across distinct expert communities "from which legitimate knowledge claims and judgements of those knowledge claims can be made" (Bartlett et al., 2018: 4), is more diffuse in the social sciences than in physics and biology. And, as the previous section elucidated, in many cases the locus is not only diffuse, but resides altogether outside the social sciences.

This account suggests that attempts to establish the credible use of big data in the social sciences are connected to ideas about *who can legitimately make knowledge claims from those data*. These are conceptions about the *status*, *authority*, and *expertise* of individual disciplines. The credibility of social scientific knowledge production involving big data depends in part on whether the social sciences can be portrayed as the legitimate interpreters of said data. Thus, to exploit big data, social scientists must be able to shift the locus of legitimate interpretation to include their respective expert communities.

Following Collins and Evans (2007: 123–125), such shifts can be understood as attempts to frame data use as legitimate via the allocation of *positions* for actors in a discussion. For example, as Bartlett et al. (2018: 5) document, although bioinformatics is central to data analysis in post-genomic biology, the field is often portrayed as merely performing service work that is subordinate to biology, and consequently regarded as outside of the locus of legitimate interpretation of biological data. Such positioning assigns to scientific disciplines the role of legitimate or illegitimate interpreters of big data, simultaneously shaping what can be considered credible knowledge production. For instance, credible interpretation in biology must involve more than mere computational work; it must also draw on the domain expertise of biologists.

Hence, positioning can be viewed as part of the work of *enacting big data* in the social sciences, where "enactment" refers to the performative work done by scientific practices, research visions, and methodologies in the "making and re-making of scientific disciplines and their knowledge" (Bartlett et al., 2018: 4; see Law and Urry, 2004). Pickering (1995) labelled this performative process the *practice* of science, which consists of creatively building new methodologies, instruments, and theory on the basis of *models* provided by existing scientific culture. From this perspective, positioning can be viewed as a form of boundary work (Gieryn, 1983) or a screening procedure for ascertaining which disciplines should be considered to supply relevant models for establishing new methodological practices.

Although there are various audiences for legitimating work in the social sciences (funders, science journalists, etc.), one crucial audience consists of the social scientific community itself, especially the relevant publication venues. As Harré and Langenhove (1998: 105) have argued, scientific publications can be viewed as rhetorical descriptions of research processes; as such, they "always involve a positioning of the scientists towards a certain audience" for which the processes are made acceptable. I posit that examining *rhetorical positioning* in empirical research articles is important if one wishes to understand how the credibility of social scientific research with big data is argued. Next, I present the empirical material I used to investigate this question.

Material and method

My study employed a dataset of 29 peer-reviewed English-language articles (see Table 1 and the appendix), downloaded from the Clarivate Analytics Social Sciences Citation Index (SSCI) by means of the Web of Science (WoS) API. The sample was designed to include articles that present empirical analysis of data and explicitly argue for their research designs' credibility by drawing on conceptions of big data. Therefore, the sample is a subset of those empirical articles with a WoS classification as sociology that have big data as their topic.

Article	Sources of primary data	Argumentation strategy
Bulger et al. (2015)	Coursera events from meetup.com	B
Chen & Yan (2016a)	Digitised literature (Google N-gram)	B
Chen & Yan (2016b)	Digitised literature (Google N-gram)	B
Gunter & Önder (2016)	Google Analytics Web site traffic	B
Heerwig (2016)	Administrative records of financial support to candidates for federal office	B
Iannelli & Giglietto (2015)	Twitter, televised talk-show material	B
Kahn & Liu (2016)	Administrative data on hotels' energy consumption	B
Sachdeva et al. (2017)	Twitter	B
Su & Meng (2016)	Administrative records from a governmental discussion forum	B
Whang et al. (2017)	Digitised news articles	B
Xie et al. (2017)	TripAdvisor hotel reviews and manager responses	B
Burrows et al. (2017)	Commercial classification of residential addresses	C
Fitzhugh et al. (2016)	Twitter	C
McKelvey et al. (2014)	Twitter, census and administrative data on elections	C
Murthy (2017)	Twitter, digitised literature	C
O'Brien (2016)	Administrative records of citizen requests for city services	C
O'Brien et al. (2016)	Administrative records of citizen requests for city services	C
Bail (2017)	Twitter, survey data	R
Lycarião & dos Santos (2017)	Twitter	R
Nardulli et al. (2015)	Digitised news articles	R
Ogan & Varol (2017)	Twitter	R
Skeggs & Yuill (2016)	Facebook	R
Su et al. (2017)	Twitter	R
Tangherlini & Leonard (2013)	Digitised literature	R
Tinati et al. (2014)	Twitter	R
Barratt & Maddox (2016)	Digital ethnography	S
Cox (2017)	Semi-structured interviews	S
Mendenhall et al. (2017)	Digitised documents	S
Stephansen & Couldry (2014)	Participant observation, interviews, Twitter	S

Table 1: Articles, data sources, and argumentation strategies (B = big data as a change in the conditions of social research; C = conservative positioning; R = reformist positioning; S = supplementarist positioning).

I collected the sample by querying the WoS database for sociology articles that include the term 'big data' in their title, abstract, or keywords. In doing so, I followed the strategy proposed by Beer (2016: Note 1) and used the term 'big data' as an entry point to discussions about the concept. Applying this approach, I conducted an initial search for articles published prior to 2018, which yielded 117 results in total. From this initial set, I excluded non-English-language articles and classified each remaining result as empirical or non-empirical by inspecting article abstracts and, when necessary, the full text. This left me with 50 empirical articles. From these I excluded articles in which data served as the subject of the study. In these cases, big data was discussed as a set of practices to be investigated, and not as a concept guiding research design. That yielded the final sample, consisting of 29 articles. Intercoder testing of this classification procedure with a colleague for a random sample of 50 articles yielded a Cohen's kappa score of 0.72, indicating strong agreement.

Some limitations of this sampling approach should be acknowledged before I discuss the analysis. Firstly, delimiting the sample with the term 'big data' has the advantage of enabling one to explore the various meanings that the articles' authors attached to the notion, without having a fixed definition beforehand. However, this also caused articles that lack explicit use of the term to be left out of analysis (Taylor et al., 2014). Secondly, the SSCI focuses on academic journals and so excludes book-length discussions, conference proceedings, and other empirical work not published in journals; furthermore, it only indexes journals that meet its standards of quality and impact[1]. This narrows the sample to influential journals, and is likely to omit writings published in less institutionalized venues. Finally, at the time of download, the SSCI covered, in all, 129 English-language journals classified as sociology by the WoS[2]. While the list includes most major journals in sociology, it lacks exhaustive coverage of journals that might feature sociologically relevant work on the topic of big data[3]. Neither does it include various possibly relevant journals in fields removed from sociology. For these reasons, the sample should not be taken to be representative of all the various ways of thinking about novel data and methods in sociology or in the social sciences more generally. Rather, it was designed to provide focused evidence of the rhetorical work around big data in a contested context. Speaking to this aim, it provides a rich array of arguments that problematize and build credibility for big data.

To analyse the articles, I coded their full text contents with the Atlas.ti software. Firstly, I identified how the authors conceptualized big data, and how they described the benefits and shortcomings of using particular data and methods to address their research problems. Secondly, focusing on articles that problematize the use of big data, I coded their arguments in terms of credibility of research designs. Here, I focused especially on how particular research areas and relations between them were described and the characteristics that were deemed to constitute good research. The latter codes included desiderata such as comprehensiveness, systematicity, rigour, and sensitivity to context.

Guided by the theoretical framework discussed above, I analysed the coded excerpts qualitatively. Reading through the extracts under each code, I wrote a description of the argumentation strategy adopted in each article. In particular, I identified where the authors fixed the locus of legitimate interpretation of big data and which elements (e.g., theories, analytical tools, and methodological practices) they used to construct their arguments about credible data use.

On the basis of this analysis, I selected for discussion three contrasting argumentation strategies that serve as interesting cases. In the first, *conservative* positioning, credibility is constructed by fixing the locus of legitimate interpretation within the social sciences. In the second, *reformist* positioning, the locus is widened to encompass methods from outside the social sciences. Finally, in the *supplementarist* positioning, the locus of legitimate interpretation of big data is argued to

[1] See https://clarivate.com/webofsciencegroup/solutions/webofscience-ssci/

[2] The search function at https://mjl.clarivate.com/ can be used to inspect lists of journals by category

[3] For instance, the WoS sociology category does not include the journal Big Data & Society or Social Media + Society

be limited, and approaches alternative to big data analysis are portrayed as necessary. While various elements of these three strategies could be identified throughout the sample, they were most clearly distinguished in 18 of the 29 articles (see Table 1). These articles' authors engaged in extensive problematization of big data, arguing at length for credibility. My discussion of the three positioning strategies in sections 6–8 will focus on these articles, but let us begin with a look at the common context within which all the articles discussed big data.

A change in the conditions of social scientific research

The common starting point in the materials was that recent technological developments, particularly in Internet-based data collection and computational analysis methods, have brought about a *change in the conditions of social scientific research* to which future research practices will have to adapt. The availability of increasingly large volumes of data of new kinds has created *normative pressure f*or utilizing these, which implies a need for methodological development and collaboration:

> The 'big data' revolution has enhanced the ability of scholars to create useful knowledge out of structured data such as ordered numbers and unstructured data such as text or images ... social researchers must find a way to leverage developments in data science if they are to advance social science knowledge and keep pace with other disciplines. (Nardulli et al., 2015: 149)

As this quote demonstrates, the pressure to utilize big data is often associated with the vast potential they offer as sources of information. The articles variously linked the informational potential of big data to large scale, which enables more *comprehensive* and *systematic* analyses, and to the data containing information that is at the same time *macroscopic* and *detailed* while also capturing *longitudinal* patterns. The authors claimed that, produced in digital settings, big data can provide evidence of *naturally occurring* behaviour, that is, 'information on what people do and say "in the wild", rather than what they say they do in interviews and surveys' (Tinati et al., 2014: 664). For the same reason, the data were characterized as affording entirely new information about processes that are themselves new, such as hybrid use of social media and traditional communication technologies (Iannelli and Giglietto, 2015), or processes that were difficult to observe previously, such as macro-scale word-use patterns in historical literature (Chen and Yan, 2016a).

These properties of big data were variously associated with digital administrative records, social media data, digitized news and literature, and Web-site traffic data. Importantly, the change brought about by big data was often explicated not in terms of just one attractive feature but, rather, as a combination of many factors – such as increased detail and large volume – which together enable granular comparisons between cases at a comprehensive scale. These novelties were typically described in terms of comparison with more traditional methods. Therefore, what was deemed to constitute big data was contingent on the methodological context of discussion in the given domain. This is in accordance with the working definition of big data proposed by Taylor et al. (2014: 1) for the social sciences, according to which "there is a step change in the scale and scope of the sources and materials" available with respect to certain objects of interest.

While authors anchored their adoption of big data through an appeal to the data sources' attractive properties, pressures to engage in methodological development and collaboration were identified in connection with several problems, such as the data's complexity or overwhelming volume. On a related note, traditional data-generation and analysis methods developed for small-scale settings were argued to be incapable of successfully harnessing the scale and other beneficial properties of large digital datasets.

As noted above, the articles varied in the extent to which they

problematized big data. Explicit arguments to support the credibility of research designs were found largely in connection with arguments for rethinking methodology or engaging in novel collaborative relationships. In the sections below, I focus on those articles featuring extensive problematization of big data, because this is where arguments for establishing credibility were most clearly visible. With the first argumentative strategy I discuss, scholars sought to establish credibility by fixing the locus of legitimate interpretation within the social sciences.

Conservative positioning: Giving meaning with established theory

Several of the articles portrayed the found nature of big data as presenting the social sciences with a dilemma. On the one hand, the data were argued to contain information about social processes that have proved difficult to study; on the other, the data have not been produced in line with rigorous protocols designed for research purposes. Hence, they frequently contain large volumes of irrelevant detail, lack clear structure, and display potential for unknown biases. This constitutes an impediment for exploiting big data in social research. For instance, addressing geodemographic data, Burrows and colleagues argued:

> The statistical procedures that each [commercial system] uses to cluster and then classify each address are proprietary and this is one of the main reasons why such systems have sometimes not proved popular with academics. Not only that but the veracity of the classifications are not primarily driven by social scientific sensibilities; they 'work' only in the sense that they ... have proven 'useful' to a wide range of commercial, public sector, and political bodies. (Burrows et al., 2017: 191)

Here, establishing a link to existing social scientific practice is emphasized as important for credible interpretation. In addition to geodemographic data, this idea was present in connection with, variously, digital administrative data (O'Brien, 2016; O'Brien et al., 2016), Twitter discussion data (Fitzhugh et al., 2014; McKelvey et al., 2014), and Twitter in combination with digitized texts (Murthy, 2017). While some authors described Twitter data as already well-established in the social sciences, it was argued that current uses lack theoretical underpinnings (McKelvey et al., 2014; Murthy, 2017). Well-developed theoretical understanding was considered crucial for the analysis of big data, and purportedly theory-free approaches to pattern discovery were criticized (e.g., Murthy, 2017: 18; O'Brien et al., 2017: 140). A lack of face value meaning of big data impelled researchers to tie their research designs to the 'fundamental understanding' provided by established theories. Failure to do so was argued to be dangerous:

> The challenges of detecting signals of social phenomena in the online environment implore us to develop a fundamental understanding of the social phenomena we intend to detect. Failure to understand the social processes underlying activity observed at large scale is dangerous and may lead to misleading or spurious results. (Fitzhugh et al., 2016: 138)

A strategy frequently employed in these articles to establish credibility consisted of *theoretically structuring the data* to make them interpretable in terms of already familiar methodology. In this context, 'theory' amounts to an organizing conceptual framework emerging from previous social scientific research. Theory in this sense was drawn upon for diverse objectives: to distinguish between relevant and irrelevant data, to identify some parts of the dataset as informing about important social scientific concepts and phenomena, and to validate new sorts of data against trusted sources.

For instance, McKelvey et al. (2014) argued that understanding how use practices on Twitter differ is necessary for exploiting the

associated data to study offline political phenomena, such as candidates' performance in elections. To develop such an understanding, they referred to political science's theory of issue publics, which implies that electoral performance should correlate positively with the attention a candidate receives from Twitter users who ordinarily do not discuss politics. They found support for this hypothesis by identifying multiple Twitter publics through content analysis and estimating correlations between discussion volumes and the candidates' performance. In another case, O'Brien et al. (2016) drew on the 'broken windows' theory in urban sociology to identify known types of civil disorder from digital administrative data about citizen requests for city services. The authors then used factor analysis techniques to identify the dimensions of these data and to construct metrics, which they validated statistically against audit-based measurements of disorder, alongside census and survey data. Finally, Fitzhugh et al. (2016) drew on the communication theory of 'rumouring' to identify disaster-related messages on Twitter. They argued that, while algorithmic methods for signal detection are not new to social research, their application to messy social media data is problematic. Rumouring theory gave the authors criteria for filtering the data to help them increase the signal strength of disaster-related messages and interpret the results as genuinely measuring disaster communication.

These examples show that big data research following this strategy can include traditional methods for drawing statistical inferences and describing the data (Burrows et al., 2017; McKelvey et al., 2014; O'Brien, 2016; O'Brien et al., 2016), but also methods such as algorithmic signal detection (Fitzhugh et al., 2017) or keyword searches of Twitter and literary material (Murthy, 2017). The key point here is that the methods should have tried-and-true uses in the fields where they are applied and that one can make them applicable by moulding unfamiliar data in line with established theory. Once this procedure of "translating" (O'Brien et al., 2016: 114) big data to familiar methodology is completed, the information contained within may be unlocked.

Importantly, it should be noted that this emphasis on traditional methodology does not preclude joint efforts of the social sciences and other fields:

> [S]cholars should develop a systematic theory of how online discourse is related to offline discourse ... Such a theory, and the measurements it yields, would link informatics with allied social science fields such as sociology, political science, health, and economics. (McKelvey et al., 2014: 448)

While 'allied' fields such as informatics could provide the social sciences with an understanding of the techniques by which digital data are generated, interpreting what those data mean was presented as a matter to be articulated in terms of social scientific methodology: a credible interpretation of big data cannot be established without resorting to theory as a tool for organizing and giving meaning, because in isolation from social scientific methodology, the data *do not have a meaningful interpretation*. In this view, the locus of legitimate interpretation resides within the social sciences. Accordingly, the relevant articles positioned areas within the social scientific domain as possessing rigorous methodological protocols that can ensure the credible use of big data sources. This *conservative positioning* strategy is an effort to maintain the authority to legitimately make claims about big data within the social sciences. Simultaneously, a boundary was drawn between fields positioned as manifesting rigour, and alternative methodologies lying beyond the newly established locus of legitimacy:

> [Big data] must be demonstrated to be both reliable and valid in their measurement before modeling can begin, which unfortunately seems to be the default in many current approaches that emphasize 'econometrics' over 'ecometrics' or simply the power to predict. However powerful predictive analytics may be, it does not answer the substantive questions about social processes and

> mechanisms that motivate most social scientists. (O'Brien et al., 2016: 139)

In this extract, positioning is used to limit the locus of legitimate interpretation so that only certain methodological practices within the social sciences can be considered credible. This offers a contrast with the argumentation pattern discussed in the next section, wherein incorporating computational tools from outside is portrayed as necessary for credibly analysing big data in the social sciences.

Reformist positioning: Mediating with computational tools

A prevalent problem wrestled with in the materials involved the incapability of existing social scientific methodology to encompass digital data adequately. The shared feature behind these articles was that they were dealing with data that have a textual component, such as social media discussions (Bail, 2017; Ogan and Varol, 2017; Su et al., 2017; Tinati et al., 2014), news articles (Nardulli et al., 2015), or digitized literature (Tangherlini and Leonard, 2013). While standard methods of content analysis and close reading were regarded as the gold standard in terms of validity, applying them reliably to large volumes of text data was claimed to be impossible:

> Achieving high reliability in human-coded content analysis is often challenging, especially when analyzing large volumes of data, as it increases the likelihood that coders will make mistakes ... [W]hen relying on the subjective judgments of human coders, achieving perfect reliability is almost impossible. (Su et al., 2017: 408)

A related problem stems from social media data's lack of contextual and demographic details (Bail, 2017). When combined with the brevity of social media messages, the lack of contextual information was argued to make interpretation difficult even for methods with established validity (Ogan & Varol, 2017: 1224–1225).

Standard automated methods for analysing text content and network structure, while capable of reliably analysing data in large volumes, were argued to be incapable of grasping contextual nuances of meaning (Su et al., 2017: 409–411). One proposed solution for the problem of data volume was randomized sampling (Lycarião & Dos Santos, 2017: 378–379). However, others argued that sampling big data can, in extreme cases, distort the information held by the data:

> Big Data has commonly been approached with small-scale content analysis ... or larger scale random or purposive samples of tweets. Rendering Big Data manageable in this way overrides its nature as 'big' data, bypassing the scale of the data for its availability or imposing an external structure by sampling users or tweets according to a priori criteria, external to the data themselves. (Tinati et al., 2014: 665)

A more general problem raised has its roots in the proprietary nature of many digital datasets. In particular, the authors emphasized the artificiality of social media data, arguing that ready-made tools provided by platforms such as Facebook yield unreliable, black box representations of social media networks (Skeggs and Yuill, 2016). A large proportion of social media data were noted to be private and impossible to access via platform-provided tools (Bail, 2017). While Twitter was recognized as exceptional in its openness, even Twitter discussion data were argued to be artificial, being shaped by platform design (Tinati et al., 2014).

Thus, in this argumentation strategy, standard social scientific methods and the ready-made tools from digital platforms were portrayed as *incapable of accessing* the information contained in big data. The commonly adopted solution was to *extend the available methodology*

and tools with methods imported from other disciplines, most notably data science or computer science. In this vein, Bail (2017) introduced a Facebook application that facilitates obtaining users' consent to access private data, and supplements these with surveys to provide additional contextual information. Likewise, Skeggs and Yuill (2016) developed a browser plugin that tracks how Facebook monitors users elsewhere on the Internet. Another example is Nardulli et al.'s (2015) machine learning approach that combines context-sensitive human coding with scalable automated text classification to generate rich large-scale datasets from news articles. Finally, Tinati and colleagues (2014) introduced a software tool that draws together network metrics and visualizations into a dynamic workflow for alternating between large-scale representations and in-depth qualitative analyses of Twitter networks.

These examples highlight the difference between this argumentation strategy and the conservative positioning discussed in the previous section. Rather than rendering big data amenable to analysis via familiar methods, the authors in this strategy stressed that the information in digital data cannot be exploited adequately without importing or developing methods that are novel for the social sciences. The aim behind this *reformist positioning* is to extend and configure social scientific methodology to enable more flexible analysis of digital data, and to provide data access in cases of restrictions imposed by the material's proprietary nature.

In this regard, it is important to recognise this view's similarity to that expressed in the digital methods literature (Venturini et al., 2019). The twofold challenge of adapting digital tools to social scientific purposes and simultaneously retaining sufficient openness and control over the research processes also underlies the reformist positioning. Crucially, as digital methods scholars emphasize (Venturini et al., 2015), in many cases answering this challenge implies that social scientists should enter into collaboration with other disciplines. Likewise, in my sample, the success of reforms to social scientific methodology was deemed to depend on collaboration, because of the technical expertize and infrastructure required:

> [W]e believe that the most propitious path forward is to create collaborations between social scientists and data scientists. It is through such collaborations that social scientists will be able to capitalize on data science techniques while retaining the nuance needed for studying complex social phenomena. (Nardulli et al., 2015: 177)

This quote illustrates two points. It demonstrates that when credible models of methodological practice are found to be lacking in the social sciences, the locus of legitimate interpretation starts to become diffuse. However, it also makes it clear that the legitimacy of interpretation is extended beyond social science *only to the extent required to enable the application of nuanced social scientific methodology.* In the materials, computer science and data science were generally portrayed as emerging fields that, although developing rapidly, cannot independently solve the problems of interpreting textual meanings in large datasets. Social scientific theory was argued to be essential for interpreting the meaning of big data yet insufficient without methodological reform:

> [S]ociological concepts, theories and methods are critical to Big Data analysis ... the meaning of these data is not self-evident but requires robust methodologies, nuanced conceptual vocabularies and theoretical frameworks drawn inter alia from sociology. However, the existing sociological repertoire of methods ... will not be sufficient in this endeavour. (Tinati et al., 2014: 678)

In the reformist position, computational methods come to play a crucial *mediating* role between social scientific methodology and big data, enabling the application of sophisticated social scientific perspectives to big data, while retaining the information they contain. Accordingly, credible uses of big data demand hybrid methodology, which can scale

social scientific expertise to be responsive to the information inherent in big data.

Supplementarist positioning: Counterbalancing big data

The previous two sections focused on attempts at establishing the credibility of taking advantage of big data in the social sciences. This section demonstrates that the notion can also be used to argue for alternative research designs.

The starting point in this strategy was a characterization of big data as an established research agenda in the social sciences, yet one unable to answer important social scientific questions. Big data approaches were portrayed as large-scale quantitative analyses of online communication, such as network analyses or quantitative measurements of macro-scale discussion dynamics (Barratt and Maddox, 2016; Cox, 2017; Stephansen and Couldry, 2014). These approaches were presented as holding appeal in that they "map out large-scale communication patterns and network structures" (Stephansen and Couldry, 2014: 1215), and enable *unobtrusive observation* of behaviour in settings that are otherwise difficult to access, such as stigmatized online populations (Barratt and Maddox, 2016).

The main criticism of big data was that large-scale analyses of digital traces lose nuances of the context of production. Big data approaches were argued to be based on problematic assumptions, and the artificial nature of digital data was emphasized:

> Claims about large-scale quantitative analyses of digital traces … being more 'complete' or less 'biased' than surveys or interviews are premised on assumptions that native digital data objects are produced, stored and analysed 'objectively', yet researchers must choose what to select and what to store and often must rely on 'black box' media analysis tools, built by and for corporate interests … Furthermore, the meaning of the data may be lost or misinterpreted when taken out of the social and cultural context within which it was produced … This critique … suggests that there are limits to what researchers can expect from these new digital artefacts of social behaviour, both in terms of interpretation and representativeness. (Barratt and Maddox, 2016: 702)

Similar criticisms are present in the other two positionings; however, in those strategies the stated aim is to overcome these problems, whether by introducing computational tools, or by establishing methodological protocols that anchor interpretations of big data to theory. In contrast, in the articles at hand, the argumentation strategy was to align oneself with alternative approaches seen as a *counterbalance* to big data.

Along these lines, Stephansen and Couldry (2014: 1224) argued that an ethnographically and hermeneutically oriented 'small data' approach is necessary for understanding the 'micro-processes' of community-formation on Twitter. Barratt and Maddox (2016: 715), on the other hand, argued that the interaction with research subjects in digital ethnography is uniquely able to provide researchers with the contextual information needed for understanding "key community issues, like the tensions between publicity and secrecy".

The argument in this *supplementarist positioning* is that there *are bounds to the legitimate interpretation of big data in the social sciences.* With respect to certain knowledge claims, it does not matter whether computational tools are imported or the methodology is modelled on established protocols. Certain information simply cannot be accessed within a big data approach:

> While quantitative metrics can provide important insights into the form that online communities might take and the extent of their interactions, an ethnographic and hermeneutic approach is needed to understand how

> Twitter and other digital platforms become embedded within particular contexts and used by social agents for their own purposes. (Stephansen and Couldry, 2014: 1224)

This argument is premised on positioning big data research as an established branch of the social sciences, one that focuses on large-scale quantitative analysis of macro patterns in digital trace data. Big data cannot be legitimately used to address certain epistemic interests because the approach consists of large-scale unobtrusive analyses of macro structures, which by definition cannot access context-sensitive information. Here we see an instance of boundary work between big data approaches and alternative perspectives, wherein engagement with the alternatives is motivated by a portrayal of big data as an established yet epistemically limited agenda.

Importantly, the authors did not advocate rejecting big data approaches outright, but rather described them as "undoubtedly useful" (Stephansen and Couldry, 2014: 1215). The upshot is that big data analysis should be *supplemented* with context-sensitive information produced by in-depth studies, "with which we can better interpret the findings of studies based solely on the analyses of their digital traces" (Barratt and Maddox, 2016: 715). Thus, alternative approaches are able to carve out a position for themselves next to the established big data agenda, gaining support by appealing to the epistemic promise of large-scale digital data.

Concluding discussion: Enacting big data via positioning rhetoric

I have argued above that in empirical social scientific research, arguments for the credibility of research designs involving big data are shaped by the rhetorical positioning of research areas. Given conceptions of the problematic yet promising properties of big data, positioning rhetoric works to establish their locus of legitimate interpretation. How this is accomplished depends on whether the social scientific methodological practice – that is, the theories, methods, and data that are already familiar to social scientists – can be portrayed as providing readily applicable models for utilizing big data. When successful, as in the conservative positioning, the legitimacy of interpretation can be located within the social sciences, and credibility argued for by drawing on traditional methodological protocols that tie interpretation of the data to theory. Otherwise, the locus must either be widened, as the reformist position argues (to enable methodological imports from other fields), or limited, as those taking the supplemetarist position maintain (to argue for alternative approaches).

The account fleshed out above speaks interestingly to themes discussed in previous literature. Rhetorical positioning in empirical publications can be understood as part of the performative process of *enacting* big data in the social sciences (Bartlett et al., 2018). As scholars of the rhetoric of science have argued, the procedure of review and revision of scientific articles can be seen as negotiation of the status assigned to their claims by the relevant scientific community (Myers, 1985). Hence, the positioning of disciplines as legitimate interpreters of big data in empirical articles can be taken to reflect the process of constructing the boundaries within which credible social scientific knowledge claims can be made. In this light, conceptions of big data in the social sciences constitute an argumentation setting for enacting particular kinds of knowledge production. This is consistent with the idea put forward by McFarland et al. (2016) that big data represents an opportunity for establishing novel collaborative relations between the social sciences and computational disciplines. Positioning is a process that contributes to determining whether or not this negotiation can lead to the creation of a productive 'trading zone' (Collins et al., 2007; Galison, 1997), where "researchers from entirely different paradigms, despite differences in language and culture, collaborate with each other to exchange tools, information, and knowledge" (McFarland et al., 2016: 13).

In the critical literature, such hopes have been dampened by arguments that the hype-inflated rhetoric surrounding big data can create unbalanced power structures by rationalizing computational forms of knowledge production (Couldry, 2014; Elish and boyd, 2018; Kitchin, 2014). Worries about such a *digital divide* (boyd and Crawford, 2012) emerging between the social sciences and computational approaches have spurred methodological debate among social scientists, lending weight to attempts to incorporate novel data into social scientific methodology. This leads us to the question of how this incorporation is negotiated, and what kinds of social scientific knowledge production are simultaneously enacted.

Taken in its entirety, my analysis provides a balanced view of the rhetorical work around big data in the social sciences. While big data approaches were met with enthusiasm overall, critical conceptions were frequently articulated to counter their attractive properties. Moreover, this interplay between problematic and promising facets was what ultimately constituted the thrust for both methodological reform and adherence to established social scientific practice. Importantly, conceptions of big data were used to bolster arguments both in favour of the use of said data and those favouring alternative approaches, depending on how the associated disciplines were positioned. That said, given that the sample examined in my study was not representative, one should not consider this analysis to provide evidence of the prevalence of each argumentation pattern discussed. Research seeking such evidence would be worthwhile, however, and similar work is already being carried out in other contexts (Stevens et al., 2018).

That the positionings discussed above work to enact different kinds of knowledge production is evident when, for instance, one considers their diverging takes on the proprietary nature of digital data – in particular, with regard to the recent data-access limitations imposed on social media platforms (see Schroepfer, 2018). Whereas the conservative response to access restrictions would be to draw on those sources of big data still accessible via traditional methods, the reformist would respond by re-configuring social scientific methodology to improve access possibilities. In contrast, the supplementarist strategy would be to mount a critique of such data by pointing to their proprietary nature and emphasizing the need for in-depth studies. In each case, data-access limitations are cited in support of different visions of what the future methodology of social research might look like.

However, it is also important to note that these positionings might not conflict with each other in any strong sense. Instead, different rhetorical strategies are likely to be suitable for different purposes. For instance, it may be that reformist positioning is effective for credibility-building with large unstructured sets of textual data while the conservative strategy works for data more familiar to social scientists, such as digital administrative records. Indeed, the account proposed here points to an unanswered question that calls for future empirical work: what determines which disciplines and methodological practices are positioned as legitimate in enacting big data? Pickering (1995) has argued that the elements employed in creatively constructing novel scientific practices are selected as part of a somewhat indeterminate real-time process of discovery. Comprehensive enquiry examining the conceptions that guide positioning in different contexts wherein big data are enacted could provide insights into how this creative process of repurposing and discovery works.

Acknowledgements

I would like to thank the anonymous reviewers for their instructive comments. I also wish to thank the Digital Content Communities research group at Aalto University for helpful discussions, and Matti Nelimarkka for assisting with the categorization of the empirical materials used in this study. This work was supported by the Finnish Foundation for Economic Education and the KONE Foundation (project: 'Algorithmic Systems, Power and Interaction').

References

Bartlett A, Reyes-Galindo L and Stephens N (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.

Beer D (2016) How should we do the history of Big Data? Big Data & Society 3(1). DOI: 10.1177/2053951716646135.

boyd d and Crawford K (2012) Critical questions for big data. Information, Communication & Society 15(5): 662–679.

Bowker G (2013) Data flakes. In: Gitelman L (ed) "Raw Data" is an Oxymoron. Cambridge: The MIT Press, pp.167–171.

Bowker G (2014) The theory/data thing. International Journal of Communication 8: 1795–1799.

Burrows R and Savage M (2014) After the crisis? Big Data and the methodological challenges of empirical sociology. Big Data & Society 1(1). DOI: 10.1177/2053951714540280.

Collins H and Evans R (2007) Rethinking expertise. Chicago: University of Chicago Press.

Collins H, Evans R and Gorman M (2007) Trading zones and interactional expertise. Studies in History and Philosophy of Science Part A 38(4): 657–666.

Connelly R, Playford C, Gayle V et al. (2016) The role of administrative data in the big data revolution in social science research. Social Science Research 59: 1–12.

Couldry N (2014) Inaugural: A necessary disenchantment: Myth, agency and injustice in a digital world. The Sociological Review 62(4): 880–897.
Crompton R (2008) Forty years of sociology: Some comments. Sociology 42(6): 1218–1227.

Edwards A, Housley W, Williams M et al. (2013) Digital social research, social media and the sociological imagination: Surrogacy, augmentation and re-orientation. International Journal of Social Research Methodology 16(3): 245–260.

Elish MC and boyd d (2018) Situating methods in the magic of Big Data and AI. Communication Monographs 85(1): 57–80.

Frade C (2016) Social theory and the politics of Big Data and method. Sociology 50(5): 863–877.

Galison P (1997) Image and logic: A material culture of microphysics. Chicago: University of Chicago Press.

Gieryn T (1983) Boundary-work and the demarcation of science from non-science: Strains and interests in professional ideologies of scientists. American Sociological Review 48(6): 781–795.

Gillespie T (2014) The relevance of algorithms. In: Boczkowski P, Foot K, and Gillespie T (eds) Media technologies: Essays on communication, materiality, and society. Cambridge: The MIT Press, pp.167–193.

Gitelman L and Jackson V (2013) Introduction. In: Gitelman L (ed) "Raw Data" is an Oxymoron. Cambridge: The MIT Press, p.1–14.

Given J (2006) Narrating the digital turn: Data deluge, technomethodology, and other likely tales. Qualitative Sociology Review 2(1): 54–65.

Goldberg A (2015) In defense of forensic social science. Big Data & Society 2(2). DOI: 10.1177/2053951715601145.

Golder S and Macy M (2015) Digital footprints: Opportunities and challenges for online social research. Annual Review of Sociology 40(1): 129–152.

Grimmer J (2015) We are all social scientists now: How big data, machine learning, and causal inference work together. Political Science & Politics 48(1): 80–83.

Hacking I (1991) How should we do the history of statistics? In: Burchill G, Gordon C and Miller P (eds) The Foucault effect. Chicago: University of Chicago Press, p.181–195.

Halavais A (2015) Bigger sociological imaginations: Framing big social data theory and methods. Information, Communication & Society 18(5): 583–594.

Halford S and Savage M (2017) Speaking sociologically with Big Data: Symphonic social science and the future for Big Data research. Sociology 51(6): 1132–1148.

Harré R and Langenhove L (1998) Positioning theory: Moral contexts of intentional action. Oxford: Blackwell.

Kennedy H (2016) Post, mine, repeat: Social media data mining becomes ordinary. London: Palgrave Macmillan.

Kennedy H and Hill L (2018) The feeling of numbers: Emotions in everyday engagements with data and their visualisation. Sociology 52(4): 830–848.

King G (2016) Preface: Big Data is not about the data! In: Alvarez RM (ed) Computational Social Science: Discovery and Prediction. New York: Cambridge University Press, pp.vii–x.

Kitchin R (2014) The data revolution. London: Sage.

Kitchin R and McArdle G (2016) What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. Big Data & Society 3(1). DOI: 10.1177/2053951716631130.

Law J and Urry J (2004) Enacting the social. Economy & Society 33(3): 390–410.

Lazer D and Ratford J (2017) Data ex machina: Introduction to Big Data. Annual Review of Sociology 43: 19–39.

Manovich L (2012) Trending: The promises and the challenges of big social data. In: Gold MK (ed) Debates in the Digital Humanities. Minneapolis: University of Minnesota Press, pp.460–475.

Marres N (2012) The redistribution of methods: On intervention in digital social research, broadly conceived. The Sociological Review 60(1): 139–165.

Marres N (2017) Do we need new methods? In: Marres N, Digital sociology: The reinvention of social research. Cambridge: Polity, pp.78–115.

McFarland D, Lewis K and Goldberg A (2016) Sociology in the era of Big Data: The ascent of forensic social science. The American Sociologist 47(1): 12–35.

Myers G (1985) Texts as knowledge claims: The social construction of two biology articles. Social Studies of Science 15(4): 593–630.

Pentzold C and Fischer C (2017) Framing Big Data: The discursive construction of a radio cell query in Germany. Big Data & Society 4(2). DOI: 10.1177/2053951717745897.

Pickering A (1995) The mangle of practice: Time, agency and science. Chicago: University of Chicago Press.

Puschmann C and Burgess J (2014) Metaphors of Big Data. International Journal of Communication 8: 1690–1709.

Rogers R (2013) Digital methods. Cambridge: The MIT Press.

Rouvroy A (2013) The end(s) of critique: Data behaviourism versus due process. In: De Vries K (ed) Privacy, Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology. Abingdon: Routledge, pp.143–167.

Ruppert E, Law J and Savage M (2013) Reassembling social science methods: The challenge of digital devices. Theory, Culture & Society 30(4): 22–46.

Salganik M (2017) Bit by bit: Social research in the digital age. Princeton: Princeton University Press.

Savage M and Burrows R (2007) The coming crisis of empirical sociology. Sociology 41(5): 885–899.

Savage M and Burrows R (2009) Some further reflections on the coming crisis of empirical sociology. Sociology 43(4): 762–772.

Schroepfer M (2018). An update on our plans to restrict data access on Facebook. In: Facebook Newsroom. Available at: https://newsroom.fb.com/news/2018/04/restricting-data-access/ (accessed 20 October 2019).

Stevens M, Wehrens R and De Bont A (2018) Conceptualizations of Big Data and their epistemological claims in healthcare: A discourse analysis. Big Data & Society 5(2). DOI: 10.1177/2053951718816727.

Taylor L, Schroeder R and Meyer E (2014) Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same? Big Data & Society 1(2). DOI: 10.1177/2053951714536877.

Törnberg P and Törnberg A (2018) The limits of computation: A philosophical critique of contemporary Big Data research. Big Data & Society 5(2). DOI: 10.1177/2053951718811843.

Van Dijck J (2014) Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. Surveillance & Society 12(2): 197–208.

Venturini T and Latour B (2010) The social fabric: Digital traces and quali-quantitative methods. In: Proceedings of Futur en Seine 2009, Paris, France, 29 May ¬– 7 June 2009, pp.87–101. Paris: Editions Futur en Seine.

Venturini T, Jensen P and Latour B (2015) Fill in the gap: A new alliance for social and natural sciences. Journal of Artificial Societies and Social Simulation 18(2). DOI: 10.18564/jasss.2729.

Venturini T, Bounegru L, Gray J et al. (2019) A reality check(list) for digital methods. New Media & Society 20(11): 4195–4217.

Wallach H (2018) Computational social science != computer science + social data. Communications of the ACM 61(3): 42–44.

Watts D (2011) Everything is obvious: Once you know the answer. New York: Crown Business.

Youtie J, Porter A and Huang Y (2017) Early social science research about Big Data. Science and Public Policy 44(1): 65–74.

Appendix: Articles included in the analysis

Bail C (2017) Taming big data: Using app technology to study organizational behavior on social media. Sociological Methods & Research 46(2): 189–217.

Barratt M and Maddox A (2016) Active engagement with stigmatised communities through digital ethnography. Qualitative Research 16(6): 701–719.

Bulger M, Bright J and Cobo C (2015) The real component of virtual learning: Motivations for face-to-face MOOC meetings in developing and industrialised countries. Information Communication & Society 18(10): 1200–1216.

Burrows R, Webber R and Atkinson R (2017) Welcome to 'Pikettyville'? Mapping London's alpha territories. Sociological Review 65(2): 184–201.

Chen Y and Yan F (2016a) Economic performance and public concerns about social class in twentieth-century books. Social Science Research 59: 37–51.

Chen Y and Yan F (2016b) Centuries of sociology in millions of books. Sociological Review 64: 872–893.

Cox J (2017) The source of a movement: making the case for social media as an informational source using Black Lives Matter. Ethnic and Racial Studies 40(11): 1847–1854.

Fitzhugh S, Gibson B, Spiro E et al. (2016) Spatio-temporal filtering techniques for the detection of disaster-related communication. Social Science Research 59: 137–154.

Gunter U and Önder I (2016) Forecasting city arrivals with Google Analytics. Annals of Tourism Research 61: 199–212.

Heerwig J (2016) Donations and dependence: Individual contributor strategies in house elections. Social Science Research 60: 181–198.

Iannelli L and Giglietto F (2015) Hybrid spaces of politics: The 2013 general elections in Italy, between talk shows and Twitter. Information Communication & Society 18(9): 1006–1021.

Kahn M and Liu P (2016) Utilizing "Big Data" to improve the hotel sector's energy efficiency: Lessons from recent economics research. Cornell Hospitality Quarterly 57(2): 202–210.

Lycarião D and Dos Santos MA (2017) Bridging semantic and social network analyses: The case of the hashtag #precisamosfalarsobre-aborto (we need to talk about abortion) on Twitter. Information Communication & Society 20(3): 368–385.

McKelvey K, DiGrazia J and Rojas F (2014) Twitter publics: How online political communities signaled electoral outcomes in the 2010 US house election. Information Communication & Society 17(4): 436–450.

Mendenhall R, Brown N and Black M (2017) The potential of big data in rescuing and recovering black women's contributions to the Du Bois-Atlanta School and to American sociology. Ethnic and Racial Studies 40(8): 1231–1233.

Murthy D (2017) Comparative process-oriented research using social media and historical text. Sociological Research Online 22(4): 3–26.

Nardulli P, Althaus S and Hayes M (2015) A progressive super-vised-learning approach to generating rich civil strife data. Sociological Methodology 45(1): 148–183.

O'Brien T (2016) Using small data to interpret big data: 311 reports as individual contributions to informal social control in urban neighborhoods. Social Science Research 59: 83–96.

O'Brien T, Sampson R and Winship C (2016) Ecometrics in the age of big data: Measuring and assessing "broken windows" using large-scale administrative records. Sociological Methodology 45(1): 101–147.

Ogan C and Varol O (2017) What is gained and what is left to be done when content analysis is added to network analysis in the study of a social movement: Twitter use during Gezi Park. Information Communication & Society 20(8): 1220–1238.

Sachdeva S, McCaffrey S and Locke D (2017) Social media approaches to modeling wildfire smoke dispersion: Spatiotemporal and social scientific investigations. Information Communication & Society 20(8): 1146–1161.

Skeggs B and Yuill S (2016) The methodology of a multi-model project examining how Facebook infrastructures social relations. Information Communication & Society 19(10): 1356–1372.

Stephansen H and Couldry N (2014). Understanding micro-processes of community building and mutual learning on Twitter: A 'small data' approach. Information Communication & Society 17(10): 1212–1227.

Su Z and Meng T (2016) Selective responsiveness: Online public de-mands and government responsiveness in authoritarian China. Social Science Research 59: 52–67.

Su L, Cacciatore M, Liang X et al. (2017) Analyzing public sentiments online: Combining human- and computer-based content analysis. Information Communication & Society 20(3): 406–427.

Tangherlini T and Leonard P (2013) Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and humanities research. Poetics 41: 725–749.

Tinati R, Halford S, Carr L et al. (2014) Big data: Methodological challenges and approaches for sociological analysis. Sociology 48(4): 663–681.

Whang T, Lammbrau M and Joo H (2017) Talking to whom? The chang-ing audience of North Korean nuclear tests. Social Science Quarterly 98(3): 976–992.

Xie K, Kwok L and Wang W (2017) Monetizing managerial responses on TripAdvisor: Performance implications across hotel classes. Cornell Hospitality Quarterly 58(3): 240–252.

Author Bio

Juho Pääkkönen is a doctoral candidate in sociology at the University of Helsinki. His thesis work investigates the use of digital big data sources and computational methods in the social sciences and data analytics. He is also affiliated with the Aalto University department of computer science.