Epistemologies in Practice: A Review of the Uses of Big Data in the Political and Social Sciences

Juho Pääkkönen*

University of Helsinki & Aalto University

1 Introduction

The role and value of big data in the social sciences have been a topic of heated discussion for almost a decade now (see e.g. Anderson 2008; Lazer et al. 2009; boyd and Crawford 2012; Floridi 2012; Mayer-Schönberger and Cukier 2013; Kitchin 2014; Burrows and Savage 2014; Frické 2015; Felt 2016; Symons and Alvarado 2016; Wheeler 2017).

A question central to the discussion is the issue of whether the availability of large-scale datasets and methods for analyzing them brings about a significant change in the *epistemology* of the social sciences. For instance, boyd and Crawford claim that

> Big Data creates a radical shift in how we think about research...it is a profound change at the levels of epistemology and ethics. Big Data reframes key questions about the constitution of knowledge, the processes of research, how we should engage with information, and the nature and the categorization of reality. (boyd and Crawford 2012, 665.)

Proposals in the literature concerning the nature and impact of this profound change have been varied, ranging from proclamations that big data brings about a Kuhnian paradigm shift in scientific methodology (Hey et al. 2009; Kitchin 2014), or that big data analytics leads to scientific theorizing becoming obsolete (Anderson 2008), to more modest arguments claiming that applications of computational methods should pay attention to the role of software in research (Symons and Alvarado 2016). As of today, however, no commonly accepted account of the change brought about by big data has emerged.

*juho.paakkonen@helsinki.fi

In this article, I propose that a fruitful way of approaching epistemological questions concerning big data in the social sciences is to look at *how social scientists use big data to empirically study social scientific phenomena*. The plurality of differing views about the change brought about by big data suggests that we should be cautious in making broad claims concerning big data epistemology. A more prudent approach grounds accounts of epistemology in the ways that big data is analyzed and thought about in social scientific research.

Towards this end, this article conducts a review of a selection of research articles which have appeared in recently published political and social science special issues about big data. Basing on the reviewed articles, a conceptual scheme is developed, which classifies them according to their epistemological perspective, type of data used, and methods of analysis. This classification is used to examine the different roles that methods and theory play in social scientific research with big data, and to explore the ways in which theory sanctions the choice of methodological strategy. In total, four different methodological strategies will be identified on the basis of the classification, which can be motivated by different uses of theory.

The article is organized as follows. In Section 2, I will elaborate on the aims of big data epistemology and formulate research questions, relying on perspectives from the Critical Data Studies literature. Section 3 presents the selection of articles to be examined, and Section 4 formulates the classification scheme that is used to answer the research ques-

tions. In Section 5, the articles are classified according to the proposed scheme. Basing on the classification, Section 6 identifies four different methodological strategies used in the articles to generate analysis results. In Section 7, the roles played by theory in each strategy are investigated, by looking closely at two examples from the reviewed articles. Finally, Section 8 concludes, discussing limitations and possibilities for further research.

2 Social scientific big data and the role of theory as an epistemological issue

Big data in the social sciences has been variously characterized. One way of defining the notion is to provide a list of the central features that make a certain dataset an instance of big data. For instance, Kitchin and McArdle (2016) argue that the central characteristics of big data are that the data are exhaustive, capturing an entire system, and have velocity, i.e. are created in near real-time.

Others have argued that social scientific big data should rather be thought about as the interplay of novel data processing and collection technologies, computational analysis methods, and social scientific theories (Olshannikova et al. 2017). To this list, boyd and Crawford (2012, 663) add hopes and beliefs to the effect that big data enables a "higher form of intelligence and knowledge that can generate insights that were previously impossible".

In the context of social science, the term "big data" most often arises in connection with various novel sources of digital data, such as digital trace data from social media platforms, or digitized online text corpora. The belief that big data fosters epistemic value for the social sciences is largely, too, linked to the availability of these new kinds of data sources.

However, as Floridi (2012) has argued, the *epistemological problems* pertaining to big data in research are not so much due to novel kinds of data being available, but rather have to do with the need to develop methods for their analysis. Also Gary King (2016) has argued that the most crucial issues with big data lie in data analytics, rather than data characteristics *per se*. The challenge here is to untap the potential of the rapidly accumulating masses of new data, and this requires the application of computational methods of analysis that are new to social scientific research.

Thus, epistemological problems pertaining to the use of big data in the social sciences are issues concerning the application of novel computational methods required for their analysis. What implications does the introduction of these novel methods have for how social scientific knowledge is produced? How should the benefits and limits of their use be conceived of? These are the central questions of big data epistemology in the social sciences.

As Symons and Alvarado (2016, 2-6) have noted, big data epistemology has mainly been discussed in the Critical Data Studies literature (see Iliadis and Russo 2016). Here, a prevalent view has been that the epistemological change brought about by big data in the social sciences amounts to a *reorientation of the role of theorizing in social scientific research* (Kitchin 2014, see also Anderson 2008; Hey et al. 2009; Frické 2015; Mayer-Schönberger and Cukier 2013). Following the Critical Data Studies literature, I will use this formulation of big data epistemology to guide the current investigation.

Kitchin (2014) has provided a clear account of the changing role of theories in social scientific big data research, by distinguishing between two epistemological positions he labels "empiricism" and "data-driven science". Empiricism is the view that the increasingly voluminous and comprehensive datasets and newly available methods for analyzing them allow researchers to draw inferences on the basis of data alone, without the need for theoretical knowledge or hypotheses. On this view, the scientific method is purely inductive in character, and big data allows for the automation of data-based inference to the effect that knowledge claims become objective, context-independent, and free from theory. (Kitchin 2014, 3-5)

The empiricist line of thought has often been argued against (e.g. boyd and Crawford 2012; Crawford 2013; Bowker 2014; Frické 2015), and Kitchin joins this criticism. As I think that these criticisms

are conclusive, I will not engage with their details here. Moreover, as Frické (2015, 652) has noted, the debate about theory free big data research was inaugurated by the *Wired* magazine editor Chris Anderson (2008), whose main motivation was to provoke debate by delineating an extreme epistemological position. This suggests that discussions of big data epistemology should move forward from criticisms of naïve empiricist epistemologies, to developing more plausible accounts.

One such account is Kitchin's data-driven science, which holds that the scientific method still depends on theory, but the function of scientific theorizing is changing due to big data. On this view, big data enables a new mode of research, in which theory-guided exploration of data with computational methods leads to the generation of novel hypotheses, which are then subjected to scrutiny using more traditional methods of hypothesis testing (Kitchin 2014, 5-6; see also Kelling et al. 2009, 613).

According to Kitchin (2014, 6), the data-driven approach is epistemologically novel in that it aims to extract most of the information contained in the data and use this information to formulate hypotheses, before commencing with testing them theoretically. This contrasts with more traditional, "knowledge-driven science", which is geared to developing hypotheses that would explain phenomena "under the conditions of scarce data and weak conditions" (Kitchin 2014, 5-6). Under such conditions, hypotheses are "born from the theory" rather than "born from the data" (Kitchin 2014, 5-6; Kelling et al. 2009, 613), that is, are deduced from pre-existing theoretical frameworks. Transcending the conditions of scarcity, data-driven science promises a new role for theories as guiding assumptions in data-intensive exploration (Kitchin 2014, 6).

Kitchin agrees with the view that computational methods are a key factor in the epistemological change brought about by big data. Although the fundamental epistemological shift in data-driven science is due to the changing role of theory, this change is enabled by the availability of new computational tools for analyzing big data:

> The challenge of analysing Big Data is coping with abundance, exhaustivity and variety, timeliness and dynamism, messiness and uncertainty, high relationality, and the fact that much of what is generated has no specific question in mind or is a by-rpoduct of another activity. Such a challenge was until recently too complex and difficult to implement, but has become possible due to high-powered computation and new analytical techniques...Big Data analytics enables an entirely new epistemological approach for making sense of the world; rather than testing a theory by analysing relevant data, new data analytics seek to gain insights 'born from the data'. (Kitchin 2014, 2.)

This discussion leaves us with the questions of *how* this new epistemological approach works in the context of social science, and whether it is a suitable account of social scientific big data in the first place. Kitchin (2014, 9) himself argues that data-driven science is not likely to become a dominant methodological paradigm in the social sciences, mainly due to the fact that they are largely dependent on qualitative methodology. Reconciling the computational methods required by data-driven science with qualitative research methods is an open problem, and one that will hinder the realization of a "paradigm shift" in the social scientific context (Hey et al. 2009). However, within computational social science, it might be that new modes of research are emerging that do not fit into the traditional quantitative framework of theory-driven hypothesis testing.

Basing on the discussion above, the central epistemological issues pertaining to the use of big data in the social sciences concern the introduction of new computational analysis methods, and the implications this has for the role played by theory in research. In line with the approach adopted in this article, an account of the epistemology of big data should pay close attention to how these novel methods are used in concert with more traditional methods in social scientific research employing big data. Doing so will enable the conceptualization of different roles given for theory, as it relates to the application of different kinds of analysis methods.

This objective can be expressed in the form of two research questions, which will be investigated in the remainder of this paper:

RQ1. *How are different analysis methods used in social scientific research with big data?*

RQ2. *What are the roles given for theory, and how do they relate to different analysis methods?*

These questions will be addressed by examining a selection of research articles from special issues about big data, published in social scientific journals. This selection and the classification used for its analysis will be presented in the next section.

3 The selection of big data special issues

The research articles used as data in this study were selected by searching for special issues, symposia and special sections in social scientific journals, published in 2014-2017, that explicitly mention big data as their topic.

Journal special issues were chosen as the focus in this study because searching for mentions of "big data" in the title, abstract, or content of social scientific research articles generates a very large volume of returned items. To give an example, a search between 2014-2017 in JSTOR for social scientific[1] research articles that contain the terms "big data" in their title, abstract or full text generated 6,430 items in total. This number rose if the disciplines of Business, Finance, Criminology, and Urban Studies were included, generating 7,342 articles. Further, if books were included in addition to articles, the search generated 29,879 items in total.

Analyzing the results of such a search would indeed be required in case one were to carry out a systematic literature review of the uses of big data in the social sciences. However, the aim in the present study is not to conduct a systematic review, but rather to construct a conceptualization of the ways in which big data are used to produce empirical social scientific knowledge, thus demonstrating how the examination of empirical research articles can inform an account of the epistemology of big data. The hope is that the conceptualization developed here can then in the future be subjected to scrutiny by using it as a framework in a systematic literature review.

Thus, the selection used here should not be considered to be representative of the whole field of social scientific big data research. First, the articles examined are published in journal special issues, and might not represent other empirical studies using big data that have been published in those same journals, or studies that have been published in other journals. Second, the findings from this selection cannot be generalized to other kinds of publications, such as books. Third, the findings cannot be generalized to hold in any particular social scientific field—that is, for instance for sociological big data research.

That being said, the special issues that were selected for this study aim at featuring cutting edge research done with big data in their respective fields, by providing a comprehensive view of the state-of-art in the field, or by identifying key issues confronting researchers using big data. Narrowing the analysis to a selection of articles representing social scientific big data at its best should provide a good basis for examining the potential epistemological novelty of big data research. With this purpose in mind, four special issues were included in the selection that explicitly claim to represent the cutting edge of big data research in their fields.[2] This selection included the issues *Journal of Communication* 64(2), 2014; *The ANNALS of the American Academy of Political and Social Science* 659(1), 2015; *Psychological Methods* 21(4), 2016; and *Social Science Computer Review* 35(1), 2017. In addition, in order to make the selection larger, and to include articles from the field of Sociology, the issue *Social Science Research* 59, 2016 was included as well. The resulting selection of five issues is depicted in Table 1.

The selection was further limited to articles that use big data to empirically study a phenomenon. Thus, articles which aim at developing or demonstrating a method for big data analysis, but do not actually analyze them in order to examine a phenomenon, were left out of the selection. So, too,

[1] Including Anthropology, Communication Studies, Development Studies, Economics, Political Science, Population Studies, Psychology, Public Policy & Administration, Social Work, and Sociology. The numbers reported here are from a search on 29 August 2017.

[2] See the special issue introductions (Parks 2014, 355; Shah et al. 2015a, 7; Harlow and Oswald 2016, 447-448; Gil de Zúñiga and Diehl 2017, 3).

Journal issue	Abbreviation	Articles in selection	Articles in total
Journal of Communication 64(2), 2014.	JC	8	8
The ANNALS of the American Academy of Political and Social Science 659(1), 2015.	AAPSS	7	20
Psychological Methods 21(4), 2016.	PM	2	10
Social Science Research 59, 2016.	SSR	6	12
Social Science Computer Review 35(1), 2017.	SSCR	6	7
		29	57

Table 1: Journal issues and research articles in the selection.

were articles that use an empirical example to illustrate the use of a method or a theoretical point, but do not primarily aim at establishing empirical results about the case examined.

This criteria led to the exclusion of three special issues or sections that were originally intended to be included in the selection, namely the special section on big data in *International Journal of Communication* 8, 2014; the symposium on big data in *Political Science & Politics* 48(1), 2015; and the big data special issue *International Journal of Sociology* 46(1), 2016. Although examining the arguments and conceptions present in non-empirical discussions of big data would be interesting, the present study is concerned to investigate what an examination of the actual uses of big data in social scientific knowledge production can provide for the development of big data epistemology. The numbers of articles per journal that are included in the selection are also summarized in Table 1.

A focus on special issues has the advantage that articles need not explicitly specify a notion of big data in order to be included in them. In other words, special issues might include articles that do not explicitly adopt a definition of "big data", but have been included in the issue nonetheless because they conduct research that could be described using this notion. Thus, focusing on big data special issues potentially leads to the selection including research, which operates with a notion of big data not captured using the commonly referred definitions.

4 Types of data, methods of analysis, and concepts of epistemic value

To examine the use of different analysis methods and the corresponding roles given for theory, the articles in the selection were classified according to three categories, namely *1)* Type of data used; *2)* Methods of analysis; and *3)* Conception of the epistemic value of big data.

The rationale for using these categories of classification is based on the assumption that the roles played by theory are connected to each of them. However, as I will be arguing below, how theory is used is not determined by any one category in isolation. Rather, theory can play different roles in studies that subscribe to the same conceptions of the epistemic value of big data, and analyze similar kinds of data with similar methods. Each of the classification categories are described in more detail below.

4.1 Types of data

Numerous authors have argued that the technological make-up and idiosyncrasies of the platforms used as sources of digital big data have implications for the use of these data in research (Rieder et al. 2015; boyd and Crawford 2012, 670-671). Further, it seems plausible to work under the assumption that the use of data from *natively digital* online sources novel to the social sciences might differ from the uses of data collected with more traditional methods like survey questionnaires (Rogers 2013).

Following this line of thinking, four different types of data used in the articles were distinguished from each other. I will label these types in the classification as follows.

Social media. Data from social networking platforms, such as Twitter and Facebook, and various discussion forum and wiki sites.

Other online. Data from internet sources other than social media, such as online news articles and scraped website data.

Mixed with social media. Data from traditional sources such as surveys and administrative data, mixed with online data sources that include social media data.

Mixed no social media. Data from traditional sources mixed with online data sources excluding social media data.

In the selection of 29 articles, each one used online data as one of their sources. This is reflected in the classification categories: while the first two categories include articles that used only online data, the latter two consist of articles that mixed online sources with data collected by other means.

4.2 Methods of analysis

Machine learning methods can be broadly classified into methods of *supervised* and *unsupervised* learning. In order to investigate the uses of different kinds of analysis methods, these categories will be used to classify the examined articles. Further, as the methodological pipeline of an article can simultaneously include both supervised and unsupervised methods, a mixed category will be included in addition, as described below.

Supervised methods. Methods that relate two or more data variables *with known values* to each other, in order to test their mutual dependency, or predict future values of the same variables. For instance, ordinary least squares regression, logistic regression, and machine learning classification using a training data set are all instances of supervised methods in this sense.

Unsupervised methods. Methods that use the values of one or more variables in the data to infer the values of a further, previously unknown variable. Principal Component Analysis and community detection in network analysis are examples of unsupervised methods in this sense.

Mixed. The simultaneous use of supervised and unsupervised methods in the analysis of data.

As was discussed in Section 2, the epistemological novelty of big data has been claimed to be in enabling a data-driven approach, where hypotheses no longer are based on theory, but instead are ”born from the data” (Kitchin 2014, 5-6; Kelling et al. 2009, 613). However, even a cursory glance at the articles included in the selection reveals that in their vast majority, existing social scientific theory is used to formulate explicit research questions and hypotheses that guide data collection sanction method selection. Thus, a more nuanced view is required in order to identify the ways in which big data analytics might be used to generate hypotheses born from the data. As machine learning methodology is often argued to be central to the data-driven approach (Wheeler 2017, Kitchin 2014, 2), regimenting analysis methods according to the supervised-unsupervised distinction provides a conceptual tool with which the uses of different methods can be fruitfully investigated.

4.3 Conceptions of the epistemic value of big data

Finally, the articles were classified according to their stated conception of the epistemic value of big data. It should be noted at the outset that not all articles explicitly specified how they understand the notion of big data. For this reason, in order to identify the conception adopted, the articles were examined to determine the methodological or epistemic novelty that the authors saw in their approach. This novel feature or characteristic was then used as a proxy for the epistemic value of big data. This approach is based on the assumptions that empirical research articles included in special issues on big data indeed use big data in one way or another, and moreover that their

stated methodological novelties are linked to this use.

On the basis of this approach, four different categories of epistemic value were distinguished. In the analysis below, these will be referred to as

Natural behavior. Big data is advantageous because it allows researchers to study human behavior as it occurs in a natural environment or setting.

Access to non-reachable phenomena. Big data provides researchers with access to data about phenomena that would otherwise be hard or even impossible to study.

Flexible research settings. Big data enables novel kinds of research settings that enable researchers to ask social scientific questions in new ways.

Scale. Big data enables easier and cheaper data collection than earlier methods, and enables research with greater scope, breadth, timespan, accuracy, or generalizability.

These conceptions of epistemic value have also been discussed in the Critical Data Studies and computational social science literature. For instance, boyd and Crawford (2012, 663) argue that the belief that large datasets offer previously inaccessible insights and increases in the objectivity or accuracy of knowledge claims is a central constituent of the big data phenomenon. Lazer et al. (2009, 722) argue that the new digital sources enable the continuous analysis of of human interaction in areas that have so far relied mainly on "one-time, self-reported data on relationships." Further, exhaustive datasets from virtual worlds enable "experimentation that would otherwise be impossible or unacceptable" (Lazer et al. 2009, 722). In sum, Lazer et al. (2009, 722) conclude that the introduction of computational methods to social science enable the collection and analysis of data "with an unprecedented breadth and depth and scale."

It is important to note that these conceptions of epistemic value are not mutually exclusive. Thus, classifying the articles on basis of their conceptions of epistemic value helps examine, whether articles expressing similar combinations of epistemic values applied analysis methods in a similar manner, and used social scientific theory for the same purposes.

5 Classification of the research articles

Table 2 presents the classification of the 29 research articles included in the selection according to the categories described above. This preliminary analysis already suggests several interesting insights.

First, as was mentioned above, all of the articles use data gathered online as one of their data sources. 25 out of the 29 articles (86.2%) use social media data as one of their data sources, 17 of which use Twitter data (58.6%). 12 out of the 29 papers use social media as their only data source (41.4%), eight of which use only Twitter data (27.6%).

Albeit the sample used here might not representative of social scientific big data research in general, these numbers agree with previous research, suggesting that Twitter is over-used as a data source (Rains and Brunner 2015). Moreover, although 12 out of 29 articles in total (41.4%) use online data mixed with other kinds of data, this classification suggests that data collected from social media strongly represent what counts as big data in the social sciences.

However, when publication years of the journals are taken into account, a different trend is revealed. 14 of the 15 articles published during 2014-2015 use social media data as one of their sources (93.3%), 12 of which use Twitter (80%). Moreover, only four out of 15 of these papers use online data mixed with other data sources (26.6%). By contrast, while 11 out of the 14 articles published during 2016-2017 use social media as one of their data sources (78.6%), only five of these 14 articles use Twitter data (35.7%). Furthermore, eight out of these 14 articles use online data mixed with other kinds of data sources (57.1%).

Thus, while it seems that—in this selection of articles—social media maintains its positions as the prevalent source of big data, during the timespan of 2014-2017 the sources of social media data have diversified, and are being mixed more with data from offline sources.

Second, a majority of the articles relied only on supervised methods in their analysis. While nine out of the 29 articles (31%) used unsupervised methods as part of their approach, only one article used unsupervised methods as their only methods (3.4%).

Table 2: Analysis of the use of big data in 29 social scientific research articles.

				Conception of the epistemic value of big data			
Article	Journal	Types of data (T = Twitter)	Methods of analysis	Natural behavior	Access to non-reachable phenomena	Flexible research settings	Scale
Colleoni et al. (2014)	JC	Social media (T), other online	Supervised	No	No	No	Yes
Emery et al. (2014)	JC	Social media (T)	Supervised	Yes	No	No	Yes
Giglietto and Selva (2014)	JC	Social media (T)	Supervised	No	No	No	Yes
Jungherr (2014)	JC	Mixed with social media (T)	Mixed	No	No	Yes	Yes
Park et al. (2014)	JC	Mixed with social media (T)	Supervised	Yes	No	No	Yes
Russell Neuman et al. (2014)	JC	Social media (T), other online	Supervised	Yes	No	Yes	No
Shaw and Hill (2014)	JC	Social media	Supervised	No	No	No	Yes
Vargo et al. (2014)	JC	Social media (T)	Supervised	No	No	No	Yes
Bode et al. (2015)	AAPSS	Social media (T)	Unsupervised	Yes	No	No	Yes
Freelon et al. (2015)	AAPSS	Social media (T)	Mixed	Yes	No	No	No
Guggenheim et al. (2015)	AAPSS	Social media (T), other online	Supervised	No	No	Yes	Yes
Park et al. (2015)	AAPSS	Mixed with social media (T)	Mixed	No	No	Yes	No
Shah et al. (2015)	AAPSS	Mixed with social media (T)	Supervised	Yes	No	Yes	Yes
Soroka et al. (2015)	AAPSS	Other online	Mixed	No	No	No	Yes
Welles and Contractor (2015)	AAPSS	Social media	Supervised	No	No	Yes	No
Jones et al. (2016)	PM	Social media (T)	Supervised	Yes	Yes	Yes	Yes
Stanley and Byrne (2016)	PM	Social media (T)	Supervised	No	No	Yes	Yes
Chen and Yan (2016)	SSR	Mixed no social media	Mixed	No	No	No	Yes
Deutschmann (2016)	SSR	Mixed with social media	Supervised	No	No	No	Yes
Keuschnigg et al. (2016)	SSR	Mixed no social media	Supervised	Yes	No	Yes	Yes
Reed et al. (2016)	SSR	Mixed with social media	Supervised	Yes	No	Yes	Yes
Su and Meng (2016)	SSR	Social media	Supervised	No	No	Yes	Yes
Westlake and Bouchard (2016)	SSR	Other online	Mixed	No	Yes	No	No
Brandtzaeg (2017)	SSCR	Mixed with social media	Supervised	Yes	No	No	Yes
Maireder et al. (2017)	SSCR	Social media (T)	Mixed	No	No	No	No
Kwon and Cho (2017)	SSCR	Social media	Supervised	No	No	No	Yes
Vargo and Hopp (2017)	SSCR	Mixed with social media (T)	Supervised	No	No	No	Yes
Wells and Thorson (2017)	SSCR	Mixed with social media	Mixed	No	No	Yes	Yes
Zhu (2017)	SSCR	Mixed with social media (T)	Mixed	No	No	No	No

This is somewhat surprising, given that the use of unsupervised methods for data exploration is often thought to be the novel epistemological feature that distinguishes big data research from earlier "paradigms" of social science (Kitchin 2014; Wheeler 2017; Hey et al. 2009). Basing on this sample, then, it seems that either the majority of big data research in the social sciences is not yet geared to untap the true potential of the data-driven approach, or else that the epistemological novelty in social scientific big data might not amount to a shift from theory-driven to data-driven research.

Third, perhaps unsurprisingly, the prevalent conception of the epistemic value of big data was scale. In 22 of the 29 articles (75.9%) big data was mentioned as being epistemically valuable due to the increased scale or breadth of research they enable. 12 articles (41.4%) indicated that big data enables new kinds of research settings that lead to valuable insights, while 10 articles (34.5%) held that big data is advantageous because it allows for observing behaviors in a natural setting. Finally, two articles (6.9%) stated that big data provides researchers with access to data about phenomena that would otherwise be hard or even impossible to study. Two of the 29 articles made no explicit statements concerning the epistemic value of big data.

It is interesting to note that in 12 out of the 21 articles (57.1%) which mentioned scale as an epistemic benefit of big data, some one of the other three categories was mentioned too. Thus it seems that although scale is the most commonly expressed conception of the epistemic value of big data, other conceptions often coexist with it. This suggests that an account of the epistemology of big data built only on considerations of data volume is more often than not incomplete if not altogether misguided.

6 Methodological strategies

Let us now take a closer look at how the different kinds of methods are used in the articles. I will here propose that using the supervised-unsupervised distinction, four different *methodological strategies* can be distinguished in how the articles generate their analysis results.

Unsupervised methods were used in the articles for two distinct purposes. In seven of the nine articles that used unsupervised methods, the purpose was to cluster a dataset for further analysis with supervised methods (Freelon et al. 2015; Park et al. 2015; Chen and Yan 2016; Westlake and Bouchard 2016; Maireder et al. 2017; Wells and Thorson 2017; Zhu 2017). In two articles, unsupervised methods were used also to generate the final results of the analysis (Jungherr 2014; Bode et al. 2015). In one article, factor analysis was used to "make clear(er)" the results from correlations yielded by a supervised dictionary-based sentiment analysis (Soroka et al. 2015).

In the articles that used only supervised methods, in addition to testing for dependencies between variables in the data, these were used for classifying or measuring values of variables in data for further analysis (Vargo et al. 2014; Jones et al. 2016; Welles and Contractor 2015; Shaw and Hill 2014; Kwon and Cho 2017; Brandtzaeg 2017; Deutschmann 2016; Reed et al. 2016; Keuschnigg et al. 2016; Stanley and Byrne 2016; Russell Neuman et al. 2014; Colleoni et al. 2014; Guggenheim et al. 2015; Park et al. 2014; Vargo and Hopp 2017; Shah et al. 2015b). In one article, the classification results with supervised machine learning were directly used to test the research hypotheses investigated (Emery et al. 2014). Similarly, in (Giglietto and Selva 2014), the proportions of categories resulting from an automated content analysis were examined to assess the research question and hypothesis.

Thus, basing on the examined sample, we can distinguish between at least three purposes for using unsupervised methods and four purposes for using supervised methods, presented in Table 3. These purposes are not mutually exclusive. Rather, I propose that these different purposes of use can now be represented in the form of four broad *methodological strategies* for the use of supervised and unsupervised methods in concert. Table 4 presents these strategies, along with the purposes of use that can figure in them.

It is important to note here that the methodological pipeline need not proceed in the order indicated in the table cells. However, the defining feature of each

Unsupervised methods	Supervised methods
Clustering data for analysis	Classifying or measuring data for analysis
Analyzing data by clustering	Analyzing data by classification
Clarifying results of supervised analysis	Testing dependencies between variables in data
	Testing the reliability of methods

Table 3: The purposes of using unsupervised and supervised methods.

strategy are the methods used to generate the analysis results. For instance, in the strategy "Unsupervised with supervised", clustering the data with unsupervised methods are expected to yield the results of the analysis, even though supervised methods might be used to asses the reliability of these results.

By contrast, in the "Supervised with unsupervised" strategy, data can be clustered for analysis using unsupervised methods, and then analyzed with supervised methods to examine dependencies between data variables in different clusters. However, this strategy might also involve the analysis of data by classifying it, and then using unsupervised methods to further explore latent variables in the data.

What, then, determines the methodological strategy for a given set of methods applied in an article? How can we tell whether an unsupervised clustering of a given dataset "generates the results of the analysis", or just prepares the data for further analysis? In the next section, I will suggest that this question can be answered by examining how theory is used in research. Conversely, we can now use the methodological strategies sketched here to investigate the roles played by theory in social scientific big data research. I will demonstrate this by looking at two examples, in which the researchers adopt different methodological strategies, although they subscribe to the same conceptions of epistemic value of big data, and use similar data sources. This way, the roles played by theory in guiding the methodological choices can be examined.

7 Roles of theory

The first example I will examine contrasts the analysis of data by unsupervised clustering in (Bode et al. 2015) with an analysis of data by supervised classification in (Brandtzaeg 2017).

Bode et al. (2015) study strategic hashtag use in online political clusters on Twitter. The stated aim of the research is to "examine whether social networking supports the maintenance of partisan divides, deeper political realignments, or partisan decoupling" (Bode et al. 2015, 151). The article draws its theoretical background from the Habermasian theory of the public sphere.

The study is based on a set of nearly nine million tweets from 23,466 Twitter users who followed candidates during the 2010 U.S. gubernatorial election (Bode et al. 2015, 152-153). The authors conducted a k-means clustering of the Twitter users based on their hashtag use, and used Multidimensional Scaling to represent the emerging clusters in two dimensions. An interpretation of the clusters was then formulated on the basis of hashtag use in them. (Bode et al. 2015, 153-159) Five different clusters resulted from the k-means clustering, and these were used directly to argue that

> ...citizens' political expression on Twitter reflects a multidimensional space...This represents a complicated digital world of multiple public spheres, with specific issues or ideas allowing individuals to coalesce into fluid and ad hoc discursive groupings that exist in addition to the Left-Right continuum. (Bode et al. 2015, 159-160)

Thus, the authors conclude that "solely Left-Right distinctions, while useful in some ways, inadequately describe political behavior" (Bode et al. 2015, 159).

The choice of unsupervised methods is motivated by arguing that it allows an understanding of the political behavior of Twitter users to emerge "organically", without "researchers imposing a known spectrum of understanding...on their actions" (Bode et al.

Pure unsupervised	**Unsupervised with supervised**
Analyzing data by clustering	Analyzing data by clustering Testing the reliability of clustering methods
Supervised with unsupervised	**Pure supervised**
Clustering data for analysis Testing the reliability of clustering methods Testing dependencies between variables in data Analyzing data by classification Clarifying the results of supervised analysis	Analyzing data by classification Testing the reliability of classification methods Classifying or measuring data for analysis Testing dependencies between variables in data

Table 4: Methodological strategies.

2015, 152). This is moreover held to be a novelty and advantage of the adopted approach with respect to earlier studies of online political clusters (Bode et al. 2015, 152). Thus, it seems that the authors' conception of the epistemic value of big data is that it enables them to study political behavior in a natural setting. In addition to this, scale is also argued to be epistemically valuable, in that the "larger, more complete sample" used as data in the study enables the authors to "better study online political clusters and their strategic practices" (Bode et al. 2015, 152).

Brandtzaeg (2017), on the other hand, is a study about gender inequality in civic participation on Facebook. This study used Facebook like behavior data including user demographic information, provided by the big data analysis application "Wisdom", to compare the proportions of males and females that like Facebook Pages associated with civic engagement in 10 different countries (Brandtzaeg 2017, 109-111). In addition, the study used world demographic data from the United Nations. However, the function of this data was primarily to provide a benchmark comparison for the Wisdom data demographics (Brandtzaeg 2017, 111). The aim of the study was to examine a hypothetical expectation derived from earlier research that "Facebook is expected to facilitate more equal participation in civic engagement across genders and countries" (Brandtzaeg 2017, abstract; see also 105-109).

The analysis methods chosen are explicitly stated to be descriptive, meaning that the analysis proceeds "without regard to causal hypotheses" (Brandtzaeg 2017, 112). The approach is basically one of comparing the gender distributions in liking activity. Thus, the methodological strategy chosen in the article can be described as an analysis of data by classification. That is, the author compares directly the proportions of males and females liking a selection of pages, classified in terms of different kinds of civic engagement.

As in the case of Bode et al. (2015), the conceptions of epistemic value of big data adopted in the article are those of natural behavior and scale. For instance, Brandtzaeg argues that "the results and measurements of the big data analytics that we used in this study go beyond the limited samples used in many other studies on Facebook" (Brandtzaeg 2017, 119), allowing the investigation of "gender differences in civic engagement on Facebook in a cross-country perspective, which is rare" (Brandtzaeg 2017, 119). Further, the big data approach chosen is argued to to be "more advantageous than sample-based surveys of what people think they did" (Brandtzaeg 2017, 109). Facebook like behavior data, in contrast to survey data, provides a means of studying "what people do", distinguished from "what people say" (Brandtzaeg 2017, 109). Thus, big data provides access to people's natural behavior, not mediated by the biasing medium of language.

However, in contrast to Bode et al. (2015), the study of Brandtzaeg (2017) is based entirely on supervised methods, despite their endorsing the same conceptions of epistemic value of big data and both basing their investigation on social media data, albeit from different platforms. Why would this be the case?

I suggest that the explanation here is that the articles start out from different theoretical backgrounds, which suggest different kinds of questions. In their research, Bode et al. (2015, 151) draw on a theoretical background which suggests that the public sphere is polarized. The primary theoretical aim in this study is to demonstrate that political networks on Twitter are more varied than a simple Right-Left division would imply, and to examine the strategic use of hashtags in these different network communities. Adopting a preconceived classification scheme, with which to describe the Twitter network, might in fact be detrimental towards the achievement of this aim, as it would not reveal communities unrecognized by the classification. The strategy adopted in this paper is thus targeted at *showing that the previous theoretical conceptions are not the only option for studying the political communities on Twitter*.

By contrast, the study of Brandtzaeg (2017) is based on a rather clear hypothetical expectation concerning Facebook. The main theoretical aim of the article is to show that this expectation is misguided. A suitable strategy for doing so is to start out from concepts developed in previous research, and to demonstrate that when analyzed using those concepts, the data does not support the hypothetical expectation. Thus, the strategy in this paper seeks to *demonstrate a fact using the concepts adopted in previous research*.

I therefore suggest that there are two different roles of theory at play here. First, for Bode et al. (2015), previous theory serves to provide a conceptual foil, which is shown to be only partially correct using a methodological strategy that is not dependent on the concepts derived from that theory. Second, in Brandtzaeg (2017), previous theory served to provide concepts which sanction and determine the investigation of gender disparities in Facebook civic engagement. The aim here is not to scrutinize the concepts themselves, but rather to use them in the analysis of data, in order to derive meaningful results. Here, the results bear directly on unresolved questions suggested by the theoretical concepts adopted, but do not demonstrate that the data could be interpreted using another theoretical perspective.

Let us now move on to examining the second example. Here, the contrast will be between the methodological strategies of using unsupervised methods for analysis with supervised methods for reliability testing (Jungherr 2014), and the use of supervised methods for the analysis of an unsupervised clustering of data (Wells and Thorson 2017).

In his study of the temporal dynamics of political discussion between Twitter and news media, Jungherr (2014) used Principal Component Analysis (PCA) to analyze the data by clustering it. To do this, he combined datasets documenting the daily mentions of political actors in Twitter, printmedia and television broadcasts, during the three months preceding the federal election of Germany in 2009 (Jungherr 2014, 243-244).

Jungherr's starting point was Andreq Chadwick's theory of the hybrid media system, which he used , together with earlier research on the logic of political coverage on Twitter and traditional media, to formulate two research questions regarding the dynamics between different media (Jungherr 2014, 239-243). The first of these asked "whether the dynamics of mentions of political actors in different media follow the same temporal patterns" (Jungherr 2014, 243). The second question investigated whether the content of popular Twitter tweets corresponded to traditional media content (Jungherr 2014, 251). The first question is addressed using PCA, while the second is investigated by manually analyzing the content of 100 popular tweets.

Jungherr (2014, 243) argues that the combination of the different datasets is the key factor which enables the examination of the dynamics between social media and news media sources. Thus, big data methods enable here a flexible research setting, where different data sources are combined into a single time series, offering "a unique window into the dynamics of political actors in traditional media and on Twitter" (Jungherr 2014, 243).

In order to examine where the dynamics of mentions in different media converge and differ, a method is required that "allows for the identification of groups of variables that correlate strongly and thus might be influenced by the same underlying process" (Jungherr 2014, 245). PCA is a method that can do this, transforming the set of variables in the data into

a smaller set of new, "latent" variables, which are "often interpreted as processes driving the values of manifest variables" (Jungherr 2014, 247).

The idea here is that if the number of components produced by PCA corresponds to the number of political actors identified in the various data sources (seven), then the dynamics of all different media can be interpreted as being "driven" by the same latent variables. On the other hand, if the number of components corresponds to the number of different data sources (three), then the dynamics of different media can be interpreted as being driven by different latent variables. (Jungherr 2014, 248.) Thus, the unsupervised method is chosen here because it enables the researcher to address a research question which requires that data from multiple different sources are studied in combination, by revealing latent features that drive the dynamics that manifest in the data.

Interestingly, Jungherr (2014) emphasizes that in order to apply PCA to the data, statistical tests are required to check that the data are appropriate for this analysis. The reason for this is that the dataset used is "somewhat below the ideal observation count for PCA" (Jungherr 2014, 248). Thus, the epistemic value of scale, although not otherwise mentioned in the article, enters the discussion here as a methodological requirement. The application of computational methods, in this case, would ideally require that the dataset fulfill certain requirements pertaining to volume.

The PCA generates three clusters in total, leading Jungherr to conclude that

> The PCA shows two patterns: First, the mentions of political parties in traditional media...and on Twitter follow different dynamics. It appears fair to assess that both media types, traditional and new, follow different logics when it comes to the coverage of political events regarding parties. (Jungherr 2014, 249-250.)

This interpretation is based mainly on a short reflection of the clustering results, and Jungherr (2014, 249) himself cautions that "one is well advised not to overinterpret the results of exploratory data analysis". However, the point here is that new unsupervised methods made it possible for Jungherr to address a research question suggested by previous theory, but difficult to investigate without the flexibility in data analysis enabled by this approach. The methodological strategy adopted here was therefore aimed at *expanding the scope of theory to new areas and phenomena*.

This same aim is also present in the study of Wells and Thorson (2017), but the methodological strategy adopted here differs from that of Jungherr (2014). This study examines dependencies between variables measuring the political and news consumption of a sample of survey respondents, and the contents of the respondents' Facebook news feeds (Wells and Thorson 2017, 38-41).

Wells and Thorson start their investigation with hypotheses derived from the theoretical framework of *curated flows*, according to which "citizens now sit at the epicenter of multiple, intertwined content *flows*", suggesting that "the choice to follow political or news actors on FB will be driven by personal interest" (Wells and Thorson 2017, 36-37). Further, on the basis of previous research, Wells and Thorson (2017, 37-38) "expect that the likelihood of encountering news content in friends' posts should not be affected by a given individual's propensity for personal politics curation", and that "news delivered via strategic curation or journalistic curation will have a greater effect on knowledge...than will public affairs content encountered via social curation".

Wells and Thorson (2017, 41-44) test these hypotheses with ordinary least squares regression models. However, to investigate the contents appearing in the Facebook users' news feeds, the authors also classified the Facebook pages liked by these users into 12 different categories, and then used Principal Component Analysis to cluster the Pages to "explore the presence of underlying factors among these categories." In the end, the authors were left with seven distinct categories for the Facebook pages, produced by the PCA (Wells and Thorson 2017, 40). These categories were then used in the examination of what content appears in the survey respondents' Facebook feeds, and from which source this content originates (Wells and Thorson 2017, 41-43).

The authors argue that the epistemological novelty in their work stems from the possibility provided by big data to combine survey data, which "pluck individuals from their social context" with granular

social media data, offering "the possibility of placing individuals in their (partial) networked context" (Wells and Thorson 2017, 35). According to the authors, the central thesis of the curated flows framework (that content flows are unique) "jeopardize the validity of traditional survey techniques", but luckily "big data suggests the possibility of capturing snapshots of the information experiences of a given individual" (Wells and Thorson 2017, 37).

Moreover, this flexibility of research settings is made possible by the scale of data, enabling the researchers to construct networks large enough to represent the social contexts of the individuals studied (Wells and Thorson 2017, 35). Thus, here the value of big data is conceived of as flexibility of research settings enabled by the scale of data.

As in the case of Jungherr (2014), in Wells and Thorson (2017) Principal Component Analysis is used to cluster a dataset of social media data. Both articles subscribe to the epistemic value of big data being due to the flexible research settings and scale of research they enable. The role played by unsupervised clustering differs from that of Jungherr (2014), however. In this latter case, PCA serves only to cluster Facebook Pages for further analysis of the data using regression models. As with the first example above, we can now ask, why do the methodological strategies in these two articles differ?

With respect to this second example, I suggest that the role played by theory is similar in both of the articles, namely, that of providing a conceptualization in terms of which the data is investigated. However, in both of these cases, the authors argue that without the possibility of combining different datasets provided by big data, addressing the questions derived from the theories used would not be possible. Thus, the theories used in these articles suggest novel phenomena which require particular methodological strategies in order for them to be amenable for study. That is, the role of theory here is to guide methodological exploration, by suggesting how combinations of new data sources and methods might yield meaningful results.

This reading is supported by Wells and Thorson's (2017, 47) observation that big data research is often criticized of being "heavily descriptive and theoretically light". Their stated motivation for using Facebook data in combination with survey responses is that the "approach of combining social media trace data with conventional methods is a useful way to ground the emerging big data perspective on data we better understand". Therefore, combining Facebook data with survey data aided the authors "in making a theoretically informed contribution" (Wells and Thorson 2017, 47). It seems then that the role of theory here is to guide the adoption of novel methods and use of new data sources, and the aim of the methodological strategy adopted in the article is targeted at accomplishing this aim.

This observation can also explain differences between the methodological strategies followed by Jungherr (2014) and Wells and Thorson (2017). Jungherr (2014, 243), too, holds that the flexible research setting enabled by big data is epistemically valuable because it allowed him to investigate theoretical questions that would have previously been difficult to address. However, Jungherr (2014, 249) explicitly claims that the unsupervised analysis he conducted is "exploratory", and that the results of such an analysis should be interpreted with caution. Therefore, the methodological strategy of using supervised methods for generating the analysis results, adopted in Wells and Thorson (2017), seems to enable more reliable results regarding the new theoretical questions addressed.

8 A radical change in epistemology?

In this paper, I have proposed a classification scheme constructed on the basis of 29 social scientific research articles that use big data to empirically study a phenomenon.

I started out with two research questions investigating the ways in which new computational methods are used in the social sciences, and the role that theory plays in research with these methods. Basing on Kitchin's (2014) idea of data-driven science, I identified these two questions as central for the present state of the epistemology of big data.

The classification scheme proposed here is in-

tended to help address these questions and to ground the examination of epistemological questions pertaining to the use of big data on the way in which new data sources and methods are actually used in the social sciences.

Using the classification scheme, I identified four different methodological strategies employed in the articles, which are based on the use of unsupervised and supervised methods of analysis in different combinations. Further, I examined how these methodological strategies are used in four different articles and, argued that doing so can help us discern different roles that theory plays in social scientific big data research.

Basing on the examined sample of articles, the idea that big data would bring about a radical shift in the epistemology of social science seems premature. In all of the examined articles, although some were based on data exploration using unsupervised methods, theory served the crucial role of suggesting research questions which inform the methodological strategies chosen. What seemed more a novelty in these cases was the way in which supervised and unsupervised methods were used in concert with each other.

As was already noted above, this article is based on a sample that is not representative of the field of social scientific big data in general. Therefore, the classification of the articles should only be taken as an initial step towards developing an epistemological account that is based on an examination of how big data are used in social scientific research. Ultimately, the aim of this endeavor is to construct a conceptual scheme that could account for the multitude of different methodological strategies at play in social scientific research with big data. However, doing so would require that the uses of big data be reviewed in a systematic manner.

Also, the examination of the roles of theory here could not control for the possible influence of various pragmatic factors that might have been at play in the choice of different methodological strategies. Moreover, it is unlikely that a study aiming to examine the influence of such factors could be based solely on published research articles, which are likely to contain polished accounts of the research process. However, an account aiming to flesh out the respective roles of methods and theories in big data research should also pay attention to how pragmatics shape method choice and theoretical approaches. This remains a topic for future research.

References

Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete — WIRED. *Wired*.

Bode, L., Hanna, A., Yang, J., and Shah, D. V. (2015). Candidate Networks, Citizen Clusters, and Political Expression: Strategic Hashtag Use in the 2010 Midterms. *The ANNALS of the American Academy of Political and Social Science*, 659(1):149–165.

Bowker, G. C. (2014). Big Data, Big Questions— The Theory/Data Thing. *International Journal of Communication*, 8(0):5.

boyd, d. and Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, 15(5):662–679.

Brandtzaeg, P. B. (2017). Facebook is no "Great equalizer". *Social Science Computer Review*, 35(1):103–125.

Burrows, R. and Savage, M. (2014). After the crisis? Big Data and the methodological challenges of empirical sociology. *Big Data & Society*, 1(1):205395171454028.

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

Colleoni, E., Rozza, A., and Arvidsson, A. (2014). Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. *Journal of Communication*, 64(2):317–332.

Crawford, K. (2013). The Hidden Biases in Big Data. *Harvard Business Review*.

Deutschmann, E. (2016). The spatial structure of transnational human activity. *Social Science Research*, 59:120–136.

Emery, S. L., Szczypka, G., Abril, E. P., Kim, Y., and Vera, L. (2014). Are You Scared Yet? Evaluating Fear Appeal Messages in Tweets About the Tips Campaign. *Journal of Communication*, 64(2):278–295.

Felt, M. (2016). Social media and the social sciences: How researchers employ Big Data analytics. *Big Data & Society*, 3(1):205395171664582.

Floridi, L. (2012). Big Data and Their Epistemological Challenge. *Philosophy & Technology*, 25(4):435–437.

Freelon, D., Lynch, M., and Aday, S. (2015). Online Fragmentation in Wartime: A Longitudinal Analysis of Tweets about Syria, 2011-2013. *The ANNALS of the American Academy of Political and Social Science*, 659(1):166–179.

Frické, M. (2015). Big data and its epistemology. *Journal of the Association for Information Science and Technology*, 66(4):651–661.

Giglietto, F. and Selva, D. (2014). Second Screen and Participation: A Content Analysis on a Full Season Dataset of Tweets. *Journal of Communication*, 64(2):260–277.

Gil de Zúñiga, H. and Diehl, T. (2017). Citizenship, Social Media, and Big Data. *Social Science Computer Review*, 35(1):3–9.

Guggenheim, L., Jang, S. M., Bae, S. Y., and Neuman, W. R. (2015). The Dynamics of Issue Frame Competition in Traditional and Social Media. *The ANNALS of the American Academy of Political and Social Science*, 659(1):207–224.

Harlow, L. L. and Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods*, 21(4):447–457.

Hey, T., Tansley, S., and Tolle, K. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*.

Iliadis, A. and Russo, F. (2016). Critical data studies: An introduction. *Big Data & Society*, 3(2).

Jones, N. M., Wojcik, S. P., Sweeting, J., and Silver, R. C. (2016). Tweeting negative emotion: An investigation of Twitter data in the aftermath of violence on college campuses. *Psychological Methods*, 21(4):526–541.

Jungherr, A. (2014). The Logic of Political Coverage on Twitter: Temporal Dynamics and Content. *Journal of Communication*, 64(2):239–259.

Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., and Hooker, G. (2009). Data-intensive Science: A New Paradigm for Biodiversity Studies. *BioScience*, 59(7):613–620.

Keuschnigg, M., Bader, F., and Bracher, J. (2016). Using crowdsourced online experiments to study context-dependency of behavior. *Social Science Research*, 59:68–82.

King, G. (2016). Preface: Big Data Is Not About The Data! In R. Michael Alvarez, editor, *Computational Social Science: Discovery and Prediction*. Cambridge University Press.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*.

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

Kwon, K. H. and Cho, D. (2017). Swearing Effects on Citizen-to-Citizen Commenting Online. *Social Science Computer Review*, 35(1):84–102.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. (2009). Computational Social Science. *Science*, 323.

Maireder, A., Weeks, B. E., Gil de Zúñiga, H., and Schlögl, S. (2017). Big Data and Political Social Networks. *Social Science Computer Review*, 35(1):126–141.

Mayer-Schönberger, V. and Cukier, K. (2013). *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. An Eamon Dolan book. Houghton Mifflin Harcourt.

Olshannikova, E., Olsson, T., Huhtamäki, J., and Kärkkäinen, H. (2017). Conceptualizing Big Social Data. *Journal of Big Data*, 4(3).

Park, J., Baek, Y. M., and Cha, M. (2014). Cross-Cultural Comparison of Nonverbal Cues in Emoticons on Twitter: Evidence from Big Data Analysis. *Journal of Communication*, 64(2):333–354.

Park, S., Lee, J., Ryu, S., and Hahn, K. S. (2015). The Network of Celebrity Politics: Political Implications of Celebrity Following on Twitter. *The ANNALS of the American Academy of Political and Social Science*, 659(1):246–258.

Parks, M. R. (2014). Big Data in Communication Research: Its Contents and Discontents. *Journal of Communication*, 64(2):355–360.

Rains, S. A. and Brunner, S. R. (2015). What can we learn about social network sites by studying Facebook? A call and recommendations for research on social network sites. 17(1):114–131.

Reed, P. J., Spiro, E. S., and Butts, C. T. (2016). Thumbs up for privacy?: Differences in online self-disclosure behavior across national cultures. *Social Science Research*, 59:155–170.

Rieder, B., Abdulla, R., Poell, T., Woltering, R., and Zack, L. (2015). Data critique and analytical opportunities for very large Facebook Pages: Lessons learned from exploring "We are all Khaled Said". *Big Data & Society*.

Rogers, R. (2013). Situating Digital Methods. *Digital Methods*, pages 1–17.

Russell Neuman, W., Guggenheim, L., Mo Jang, S., and Bae, S. Y. (2014). The Dynamics of Public Attention: Agenda-Setting Theory Meets Big Data. *Journal of Communication*, 64(2):193–214.

Shah, D. V., Cappella, J. N., and Neuman, W. R. (2015a). Big Data, Digital Media, and Computational Social Science: Possibilities and Perils. *The ANNALS of the American Academy of Political and Social Science*, 659(1):6–13.

Shah, D. V., Hanna, A., Bucy, E. P., Wells, C., and Quevedo, V. (2015b). The Power of Television Images in a Social Media Age: Linking Biobehavioral and Computational Approaches via the Second Screen. *The ANNALS of the American Academy of Political and Social Science*, 659(1):225–245.

Shaw, A. and Hill, B. M. (2014). Laboratories of Oligarchy? How the Iron Law Extends to Peer Production. *Journal of Communication*, 64(2):215–238.

Soroka, S., Young, L., and Balmas, M. (2015). Bad News or Mad News? Sentiment Scoring of Negativity, Fear, and Anger in News Content. *The ANNALS of the American Academy of Political and Social Science*, 659(1):108–121.

Stanley, C. and Byrne, M. D. (2016). Comparing vector-based and Bayesian memory models using large-scale datasets: User-generated hashtag and tag prediction on Twitter and Stack Overflow. *Psychological Methods*, 21(4):542–565.

Symons, J. and Alvarado, R. (2016). Can we trust Big Data? Applying philosophy of science to software. *Big Data & Society*, 3(2).

Vargo, C. J., Guo, L., McCombs, M., and Shaw, D. L. (2014). Network Issue Agendas on Twitter During the 2012 U.S. Presidential Election. *Journal of Communication*, 64(2):296–316.

Vargo, C. J. and Hopp, T. (2017). Socioeconomic Status, Social Capital, and Partisan Polarity as Predictors of Political Incivility on Twitter. *Social Science Computer Review*, 35(1):10–32.

Welles, B. F. and Contractor, N. (2015). Individual Motivations and Network Effects: A Multilevel Analysis of the Structure of Online Social Relationships. *The ANNALS of the American Academy of Political and Social Science*, 659(1):180–190.

Wells, C. and Thorson, K. (2017). Combining Big Data and Survey Techniques to Model Effects of Political Content Flows in Facebook. *Social Science Computer Review*, 35(1):33–52.

Westlake, B. G. and Bouchard, M. (2016). Liking and hyperlinking: Community detection in online child sexual exploitation networks. *Social Science Research*, 59:23–36.

Wheeler, G. (2017). Machine Epistemology and Big Data. In *The Routledge Companion to Philosophy of Social Science*.

Zhu, Q. (2017). Citizen-Driven International Networks and Globalization of Social Movements on Twitter. *Social Science Computer Review*, 35(1):68–83.

We are the leading scholarly society concerned with the research and teaching of political science in Europe, headquartered in the UK with a global membership. About Us. Contact Us.Â Epistemologies in Practice: A Review of the Uses of Big Data in the Political and Social Sciences. View Paper Details. Information, Communication, Digitization, and Datafication: Four Analytical Stages in Researching Social Movements and Media. View Paper Details. The European Consortium For Political Research. GeoJournal DOI 10.1007/s10708-014-9599-x Rethinking big data in digital humanitarianism: practices, epistemologies, and social relations Ryan Burns Springer Science+Business Media Dordrecht 2014 Abstract Spatial technologies and the organizations Keywords Big Data Digital humanitarianism around them, such as the Standby Task Force and Geoweb Critical GIS Social media Critical Ushahidi, are increasingly changing the ways crises technology studies.Â Within digital human- itarianism, Big Data has featured strongly in recent efforts to improve digital humanitarian work. This Introduction shift toward social media and other Big Data sources has entailed unexamined assumptions about techno- With the advent of Big Social epistemology is the subfield of epistemology that addresses the way that groups, institutions, or other collective bodies might come to acquire knowledge. 2. The Nature of Propositional Knowledge. Having narrowed our focus to propositional knowledge, we must ask ourselves what, exactly, constitutes knowledge.Â We have noted that knowledge should not involve luck, and that Gettier-type examples are those in which luck plays some role in the formation of a justified true belief. In typical instances of knowledge, the factors responsible for the justification of a belief are also responsible for its truth. For example, when the clock is working properly, my belief is both true and justified because itâ€™s based on the clock, which accurately displays the time. Political science is practiced mostly by people who are either materialists or who donâ€™t think about big questions of meaning and existence. To be blunt, politi.Â Keywords: Ontology, Epistemology, Cosmology, philosophy of science, social sciences, Kuhn, Feyerabend, Popper. Suggested Citation: Suggested Citation. Betti, Daniel, Political Science: Ontology and Epistemology in a Self-Undermining Philosophy of Science (March 15, 2010). Available at SSRN: https://ssrn.com/abstract=2902867 or http://dx.doi.org/10.2139/ssrn.2902867. Daniel Betti (Contact Author). In this paper I will discuss Big Data as a suite of new methods for social and political research. I will start by tracing a genealogy of the idea that machine can perform better than human beings in managing extremely huge quantity of data, and that the quantity of information could change the quality of the interrogation posed to those data.Â This is an especially important feature of the epistemology of Big Data. In â€œErrorâ€ section we explain the main characteristics of error detection and correction along with the relationship between error and path complexity in software. In this section we provide an overview of conventional statistical methods for error detection and review their limitations when faced with the high degree of conditionality inherent to modern software systems. View.