Acta Baltica Historiae et Philosophiae Scientiarum
Vol. 8, No. 1 (Spring 2020)
DOI : 10.11590/abhps.2020.1.02

‘End of Theory’ in the Era of Big Data: Methodological Practices and Challenges in Social Media Studies

Anu Masso

Ragnar Nurkse Department of Innovation and Governance,
Tallinn University of Technology
Akadeemia tee 3,
Tallinn 12618, Estonia
Institute of Social Studies,
University of Tartu
E-mail: anu.masso@taltech.ee

Maris Männiste

Institute of Social Studies,
University of Tartu
Lossi 36,
Tartu 51003, Estonia
E-mail: maris.manniste@ut.ee

Andra Siibak

Institute of Social Studies,
University of Tartu
Lossi 36,
Tartu 51003, Estonia
E-mail: andra.siibak@ut.ee

Abstract: Emerging digital data sources provide opportunities for explaining social processes, but also challenge knowledge production practices within social sciences. This article contributes to the ‘end of theory’ discussions, which have intensified in the social sciences since the widening practice of big data and computational methods. Adopting a systematic literature review of 120 empirical articles through a combined quantitative and qualitative approach, this article strives to contribute to the ongoing discussions on the epistemological shifts in social media big data (SMBD) studies. This study offers an insight into the development of analytical methods and research practices in SMBD studies during their rapid growth period in 2012–2016. The study findings only partially revealed the 'end of theory' claim: the problem setting of the studies is rather weakly related to theory, often neither hypothesis nor research questions are formulated on the basis of previous theories or research. However, this relatively weak relatedness to theory has not led to the descriptive type of inference, but rather exploratory, or predictive ways of reasoning. Instead of enabling predictions in social science research, SMBD raises issues of understanding the causes and effects in predictions for evaluating the social mechanisms of global disruptions. Developing 'human research machines' that exploit the cognitive resources of individuals should not be the aim of SMBD production. The outcome should be to recognise that the cognitive abilities of researchers, access to data, and developing novel methods are necessary for evaluating the global impact of social behaviour.

Keywords: *big data, computational social science, digital methods, end of theory, social media, social science methodology*

Introduction

The global move towards digital technologies has intensified the discussions about knowledge production and the ways social sciences are practised. Digital data are becoming significant sources for explaining social processes and for managing crises. Social media posting, mobile phone interactions, or self-tracking with wearable devices are only some of the examples of data produced during everyday activities. As new media technologies offer a wide array of novel data sources, these data are creating the illusion for being the new sources of social truth and are believed to provide new opportunities for grasping social complexities and predicting social disruptions. Therefore, big data research is often seen as a means for addressing the complexities of handling data, rather than seeing the complex relation data has with the world that the society assumes it presents (Ho, 2020).

Digital data are not there just to be collected and analysed (Puschmann & Burgess, 2014; boyd & Crawford, 2012; Marres & Gerlitz, 2015) but are created and collected through exploitation of cognitive resources of humans using these digital tools (Mühlhoff, 2019). These digital data also have significant consequences on the implementation of computational tools and the implications related to knowledge in an increasingly datafied world (Dalton & Thatcher, 2014; Neff *et al.*, 2017). Studies have warned that the spreading "data revolution discourse" (Resnyansky, 2019) may reproduce previously dominant data practices and therefore hinder the integration of social science knowledge into big data analysis or even delay further innovations in methods.

In relation to the spread of the computational approach, several critical statements have been made that question the ways social reality is studied, and knowledge is created in big data research, like 'end of theory' (Anderson, 2008) or 'descriptive empiricism' claims (Kitchin, 2014). According to this criticism, the previously dominating approach to science, which was formulating testable hypotheses, testing models, confirming or falsifying theoretical models, is assumed to become obsolete. Instead, as Anderson (2008) argues, the data-driven approach to large-scale data gains importance where analysis of correlative relationships is preferred without explaining the underlying mechanisms of these relationships. As a result of these developments and as assumed by this criticism, the role of theory in big data (henceforth BD) studies may significantly diminish. Consequently, BD will define the main elements of a study's design, such as social categories under consideration, subjects and sample, and therefore may introduce shifts in the ways social sciences are practised.

Although there are ongoing discussions in the theoretical literature on the shifts in knowledge production in big data studies (see, e.g., Kitchin, 2014; Olteanu *et al.*, 2019), there are only a few empirical studies examining these emerging data practices in relation to BD and computational methods. Studies on the changing research practices within the social sciences have revealed shifts in data sources, data collection and analysis methods (e.g., starting with Breiman, 2001) as well as methodological shifts related to the emergence of BD (Zimmer & Proferes, 2014; Sivarajah *et al.*, 2017). A previous study has examined in greater detail user perspectives on social media data mining practices (Kennedy *et al.*, 2017) or the roles and implication of research tools on the analysing of data and the knowledge that can be achieved from that data (Weltevrede, 2016). Ongoing theoretical discussions highlight the epistemological shifts and challenges in relation to the data turn (Symons & Alvarado, 2016; Halford & Savage, 2017; Slota *et al.*, 2020) and the potentials and limits of datafied knowledge production (Fuchs, 2017; Thylstrup *et al.*, 2019; Hargittai, 2020). As far as we know, there has not been any systematic empirical exploration of research practices in the field of social media big data (SMBD) studies. This study aims to fill this gap in the corpus of SMBD research. We devised a systematic literature review method to reveal in detail the research practices in empirical articles using SMBD, in order to highlight the trends and variations in knowledge production. This study systematically explains the research practices which emerged during the most rapid growth period of SMBD studies, in 2012–2016. We incorporated ‘end of theory’ in the article’s title as a “plea” that is often used in discussions about the shifts within social sciences towards data-driven methods. The article strives to contribute to these ongoing discussions on the epistemological shifts and knowledge production in social sciences, through mapping the most used data practices in academic SMBD research.

Big data in social sciences

Approaches to big social data

In this study, we start from the data studies’ perspective of understanding BD as a socio-cultural phenomenon (Dalton & Thatcher 2014; Iliadis & Russo, 2016). This approach emphasises the shifts in knowledge production in relation to BD within the social sciences, such as turning attention to the opportunities, as well as to the critical issues like bias, limitations, and research ethics. Some discussions about BD have seen it as a shift from designed data to organic and often human-generated data within the social sciences (Shah *et al.*, 2015; O’Brien *et al.*, 2015; Schäfer & van Es, 2017). Other approaches have emphasised that SMBD are systematically and purposefully structured by social media platforms or ideologies (Weltevrede, 2016; Puschmann & Burgess, 2014). These studies disclaim that SMBD are an ontology, but instead claim that they are a way of purposefully manipulating the world and therefore leading to a significant reduction of the ways knowledge is created (Bowker, 2014).

Therefore, discussions on the implicit ideology of BD’s origin materialise from the tension between the newly emerged computational research paradigm that, on the one hand, sees BD as a possible resource and, on the other hand, as an approach assuming the socially constructed nature of data (Puschmann & Burgess, 2014). Therefore, novel data is creating new cultural phenomena, which are being expressed in new data cultures or habitual practices in knowledge production in the field of social sciences, where SMBD is gaining its meaning as data, only through the systematic ways it is produced and handled.

Our study focuses on SMBD, which is often used as synonymous to BD, since social media are widely intertwined with the everyday lives of individuals, as well as the relative accessibility of these data for research purposes of modelling social interactions and behaviour (Olshannikova *et al.*, 2017). Still, SMBD focuses on three particular topics within this dialectical relationship: digital self-representation, technology-mediated communication data, and digital relationships data (Olshannikova *et al.*, 2017). The methods used to mine SMBD have been assumed to offer good alternatives to the shortcomings of other data sources, data collection and interpretation methods, such as moderate response rates to surveys (Goyder *et al.*, 1985), non-representativeness of phone surveys (Szolnoki & Hoffmann, 2013), or the high cost of representative surveys (Spitz *et al.*, 2006), etc. In summary, the emergence of SMBD has been assumed as a potential opportunity for finding solutions to unanswered social issues. However, new research tools and the constructed nature of these data may lead to several shifts in research practices.

The methodological shift in social sciences

In social sciences, the main response to the arrival of BD has been the emergence of computational social science as a new sub-discipline (see, e.g., Mason *et al.*, 2014; Keuschnigg *et al.*, 2017). Originating from the post-positivist approaches, its focus ranges from information extraction algorithms to computer simulation models. Within the field of computational social science, innovations in relation to BD analysis were initially criticised, because they were first and foremost present amongst data analysis techniques and tools (e.g., He *et al.*, 2015; Park *et al.*, 2015). There are examples which demonstrate how new analysis methods related to BD are implemented. For instance, developing the method applicable for dynamic and large network analyses (Lazega & Snijders, 2016), devising methods for analysing spatial and temporal dynamics of political orientations with online data (DellaPosta *et al.*, 2015), or developing simulation methods for analysing and coping with network risks (Helbing, 2013).

Some authors have been quite optimistic regarding the computational shift and related scientific practices in social sciences. For example, a dominant claim (see, e.g., Hindman, 2015) is that a high variation of computational methods are available for finding the most suitable way for answering research problems, and therefore studies originating from the computational approach can be both deductive and inductive, quantitative and qualitative, or critical and administrative (see, e.g., Tukey, 1962). In one study, for example, Halavais (2013) claims that BD is even challenging the previous hypothetico-deductive model. Indeed, an inductive leap towards an explanatory theory in social sciences (see, e.g., Knight, 2019), as well as the associated computational methods (Bengio *et al.*, 2019), have been made since new forms of data enable researchers to grasp novel kinds of complexity. The inductive logic of machine-learning methods is seen as providing a perfect fit to reality, as it does not test a hypothesis but generates it from interested appraisal of past experiences (Breiman, 2001; Bengio *et al.*, 2019). Implementation of grounded theory approach in the field of machine learning (Barberis Canonico *et al.*, 2018), which operates inductively and leads to novel conceptual explanations, is one of the examples of the inductive approaches in the field of SMBD studies. As Bengio (2019) argues, the use of BD, through combining analytical procedures and theoretical frameworks, raising and answering questions of *why* instead of *what*, has the potential to explain high-level structural phenomena and therefore challenge the established theories.

Therefore, a certain permanent search for a "third way" is inherent to the field of computational social science (Breiman, 2001; Boellstorff, 2013; Manovich, 2017). Moving away from the exclusive dependence on data models and the adoption of a more diverse set of tools without making a clear choice between hermeneutic or empirical epistemic traditions are inherent aspects to these approaches. Veltri (2017) has observed the emergence of a "new culture of statistical modelling" that could have potential in bridging the theory and data-driven approaches. These discussions question the epistemological grounds of big data studies and propose pragmatism (Eklund *et al.*, 2019) or critical realism (Törnberg & Törnberg, 2018) and call for the detailed consideration of the reality and social life of big data methods.

Several initiatives illustrate this stream towards shifts in the culture of big data methods, like a move towards a reflexive digital data analysis (van Es *et al.*, 2017), implementation of a novel method of model-based recursive partitioning (Veltri, 2017), or movement towards complementarity of predictive accuracy and interpretability (Hofman *et al.*, 2017). In the case of social media, BD emerged from cultural analytics (Manovich, 2011; 2017), and the approach suggests researchers critically question their cultural assumptions related to data instead of only using demographic generalisations. Similarly, the digital methods approach (Rogers, 2019) questions the mechanically obtained objectivity and transparency of computational methods and proposes new methods and tools in line with the new medium of social media.

In sum, a wide variety of novel tools and methods are developed within social sciences for analysing SMBD. However, this new computational shift has been often criticised.

Criticism of the methodological shift

Debates about the crisis of method have been central to social sciences since its foundation in the early 19th century (Halavais, 2013; Masson, 2017). These debates have intensified in the context of the computational shifts in social sciences and raise questions about social research practices in the platform age (Wagner-Pacifici *et al.*, 2015; Gangneux & Docherty, 2018), or education in the context of the continuing technological disruptions (Dawson, 2019). As the universal and often descriptive models used within the computational approach do not take into consideration the multiplicity of the human population and the individual meanings ascribed to these diversities (Kitchin, 2014), there is a need for shifts in the ways knowledge is created (Kitchin, 2014; McFarland *et al.*, 2016). For instance, developing a new situated, reflexive and contextually nuanced epistemology (Kitchin, 2014), implementing the merging of applied and theory-driven perspectives of forensic social science (Crawford *et al.*, 2014) and the emergence of critical realism into social science research (Schäfer & van Es, 2017).

Two significant debates have emerged in social sciences in the context of BD as suggested by Veltri (2017): the crisis of measurement and the rise of competing paradigms between traditional statistical methods and algorithmic and machine-learning approaches. Computational techniques applied to digital data sources have been seen as a solution to the data turn in social sciences (Chang *et al.*, 2014a; Keuschnigg *et al.*, 2017; Slota *et al.*, 2020). Initial critiques concerning the emergence of computational social science focused on ‘the end of theory’ argument (Anderson, 2008), claiming that correlative methods used for analysing BD tend to be descriptive rather than offering explanations or using coherent models and unified theories. A similar claim has been made by Scheinfeldt (2012), who suggests we have reached a ‘post-theoretical age’, indicating that in the age of data, there is no need for theory. Considering the above, authors have agreed that a certain ‘methodological moment’ characterises the discipline of social sciences (Cohen, 2010; Rieder, 2016) as new information sources are emerging and thus discussions about discipline building and the need to reshape research practices are timely. Other research has revealed that analytic means and techniques enabling generalisability, instead of a lack of theory (Slota *et al.*, 2020), are challenging social science research in relation to the data turn.

Recently, somewhat more nuanced arguments have emerged (Thatcher *et al.*, 2016; Resnyansky, 2019), which emphasise that increasing quantification and its asymmetrical power relations may have significant influences on the ways that the social sciences are practised. However, these recent discussions (Veltri, 2017) have been pessimistic regarding the BD shift in the social sciences and related scientific practices. Importing analysis tools and methods from “hard” sciences, the exclusive use of both algorithmic methods (Veltri, 2017), and statistical stochastic data models (Bruns, 2013), have been considered as negative consequences of the computational shift. A change in scientific practices, where no coherent models, unified theories or mechanistic explanations are used, is seen as making the formerly valid approach to science obsolete (Anderson, 2008). Problematic consequences such as an analysis leading to an irrelevant theory, questionable conclusions, or the inability to work on numerous research problems are only some of the fears that have been voiced so far (Veltri, 2017). In summary, scholars have frequently been critical regarding the BD shift in social sciences, because computational methods are often referred to be only weakly related to theory. Although a wide variety of methodological solutions have been offered in response to this criticism, these debates indicate that the emergence of SMBD may influence the ways in which social sciences are practised.

Data and method of the present study

Systematic literature review

For studying the shifts in the developments of practices and the ways scientific knowledge is produced in BD studies, we used a systematic literature review method. Close reading of empirical SMBD studies was used to explain the shifts in research practices as directly reflected by the authors of these studies. Peer-reviewed empirical articles using SMBD as the main source in the research were eligible for inclusion in our sample. Furthermore, for constructing the sample for the systematic study, we used standardised search criteria: concurrent keywords of ‘social media’ and ‘big data’, full-text articles accessible online, articles published only in peer-reviewed journals and written in English, and published in the journals listed in the Social Sciences Citation Index.

The initial sample consisted of 478 articles. Several additional exclusion criteria were implemented, so that theoretical articles having no empirical focus (n = 188), articles with popular scientific focus (n = 11), articles only mentioning the keywords of social media or big data without analytically using these data (n = 112) and others (n = 48) were omitted from the final sample. The category ‘others’ mostly included articles having no empirical focus, such as editorials, commentaries, keynotes, introductions, calls for papers, essays or interviews.

The final sample consists of 120 articles that were published between 2012 and 2016 (see Table 1 on p. 42 for an overview of the structure of the sample).[1] This study first and foremost focuses on the period of the most significant increase in the SMBD studies, in which the main shifts in the used analysis methods and knowledge-creating practices are expressed.

In addition to the main sample, we also compiled a comparative group to further explain the role of theory in social science research. The comparative group included 20 articles where the terms 'big data' and 'social media' were mentioned in the body of the text. The methods used in these articles comprise more "traditional" data and methods, e.g., individual or focus group interviews, surveys, case studies, content analysis, analysis of visuals, or the use of register data with undefined sample size. However, the authors in these articles expressed awareness of BD and related methods, since most of the articles in this group either discussed the importance of developing novel methods in BD studies or proposed novel tools or methods that had either been empirically tested on a small-scale dataset or did not have clearly defined empirical data for testing the method.[2] The detailed list of articles used in this systematic literature review study can be received upon request from the corresponding author.

A semi-structured coding schema was developed for systematically studying the analysis practices in the empirical articles using SMBD. The articles were coded through close reading of the full texts of each publication. The codes reflected the identifiable formal information that was implicitly visible for the readers or explicitly expressed by the authors of the articles. For explaining the shifts in research practices, the next step in the analysis mostly used a qualitative approach, where the thematic variation between and within particular codes, characterising single aspects of research practices, was compared systematically. Open-ended textual codes were summarised using qualitative thematic analysis techniques and the software programme Maxqda. The results of the main

[1] The relatively small size of the final sample is due to four factors. (1) The sample only included studies conducted in the field of social sciences, and therefore related fields, such as the digital humanities, were excluded. (2) Only studies defined by the authors as 'big data studies' were included and then matched to the search criteria, and therefore studies combining large-scale data with a comprehensive contextual analysis may have been excluded. (3) Only accessible full-text studies that enable comprehensive analysis of research practices were excluded. (4) The study was conducted during a period of the most rapid shifts in the methods and therefore the terminology was also changing (e.g., using the term 'API' instead of 'social media'), therefore instead of an exclusive quantification of the shifts, the qualitative approach was also used.

[2] This additional comparative group is used to reveal knowledge production in a large variety of social scientific studies, instead of exclusively testing the 'end of theory' hypothesis through comparing the main sample and the control group.

Table 1. Overview of the sample structure of the study

Code	Sub-code	Frequency	%
Data source	Twitter	74	42
	Facebook	13	17
	Other social media*	39	22
	Other sources**	49	28
	Total	175	100
Year	2012	2	2
	2013	6	5
	2014	13	11
	2015	33	28
	2016	66	55
	Total	120	100
Disciplines	Computer sciences	87	21
	Media and communication	72	18
	Geography (incl. environmental sciences, tourism studies)	58	14
	Social sciences (sociology, political sciences, public administration, information management)	54	13
	Economy	48	12
	Psychology	31	8
	Interdisciplinary centers	17	4
	Medicine	15	4
	Other	25	6
	Total	407	100

* Other social media channels include (the number of times used by the articles): Flickr (6), Wikipedia (4), Forums (3), Youtube (3), Linkedin (2), Blogs (2), online geocaching platform, Microblogs, Panoramio, Sina Microblog (www.weibo.com), tripadvisor.ee, VKontakte, IMDB reviews, monCherie, OkCupid, Baidu, Uwants, Myspace (each used once) and Brightkite [used in some articles (e.g., Jiang, 2014) was a social media channel that was closed before the publication of the articles].

** Other data sources include (number of times used): Webpage (8), traditional media (7), surveys and questionnaires (5), interviews (5), online database (4), Google Trends (3), web search engines (2), documents (2), mobile app, Chrunchbase, advertisements, amazon.com, Digg, mobile call data, taxi trajectories, Word-emotion lexicon data, stock market index scores, UN official statistics.

codes were additionally summarised quantitatively, using uni- and multivariate statistical techniques with R software environment. The coded textual data were analysed quantitatively to generalise the main differences in research practices across formal article characteristics (i.e., year of publication and disciplinary background). These differences in distributions were studied both quantitatively, using association coefficients of Cramer's V and analysis of variance test F, as well as qualitatively through detecting the code intersections.

Coding schema

The semi-structured coding schema consists of ten main categories (see Table 2 for details): (1) Background of the study (including the name of the article and journal, year of publication); (2) Disciplinary background (list of disciplines identified based on authors' affiliation); (3) Problem setting (the study having clearly formulated hypothesis, research questions, or not); (4) Relatedness of the problem setting to the theory (using a 4-point ordinal scale where 1—very weakly, 4—very strongly); (5) Sample size (numeric value); (6) Data source (including a list of 23 social media platforms); (7) Data structuration (including variants of structured, un- and semi-structured, as well as database fusion); (8) Type of inference used (descriptive, exploratory, explanatory/predictive); (9) Analysis techniques used (statistical, computational, content analysis, social media analytics, other techniques); (10) Innovation, novelty of the study as estimated by the authors of the articles (contributing to the methodology, methods, techniques, research tools, using novel data sources, substantial innovations).[3]

Previous empirical studies (Zimmer & Proferes, 2014; Sivarajah *et al.*, 2017) and theoretical approaches about dynamics in SMBD studies (boyd & Crawford, 2012; Sivarajah *et al.*, 2017) were used as starting points for formulating the quantitative codes about relatedness to the theory and type of inference. For formulating sub-categories of analysis techniques used in the analysis, previous theoretical approaches (Brock, 2015) and recent textbooks (Breiman, 2001;

[3] The relatedness to theory could also be explained with the generally prescribed formatting rules of the journals, in which the articles were published. For example, the existence and comprehensiveness of the literature review section in a published empirical research could be the result of either the guidelines formulated by a particular journal or variations between disciplines. However, this study did not reveal significant differences in theory relatedness in the journals where the articles were published. Several categories were implemented to explain the practices in knowledge production and to contribute to the 'end of theory' discussion" proposed by Chris Anderson (2008).

Kitchin & McArdle, 2016) were used as a basis of code formulation and classification. The coding schema also consisted of one qualitative open code about the methodological limits as indicated by the authors.

All the three authors of the present study were engaged in close reading of the initial ten texts of the final sample so as to formulate the formalised codes for the analysis. After that, one author carried out the main quantitative coding of the articles while the other two authors contributed in those instances when the focus of the article and the used analysis techniques and types of inferences were more ambivalently formulated and therefore more difficult to code under one particular category. After coding two-thirds of the articles (80 of 120), an inter-coder agreement was calculated, by which every code was estimated, using a 5-point Likert scale (5—very easy, 1—very difficult) concerning the ease of assigning particular codes for each category. Based on this evaluation, the code descriptions and coding were revised. Those codes that were more difficult to estimate and had lower inter-coder agreement value (high degree of difficulty) were excluded from the final analysis (e.g., the code about the structuration degree of the data). After the analysis was finished, the inter-coder agreement was calculated for randomly chosen articles (n = 11), for testing the reliability of the quantitative coding. Code existence 93.59% and code frequency in compared documents was 92.8%, segment agreement was r = .82, including average Kappa coefficient 0.81. This level of inter-coder agreement is considered to be strong in the field of quantitative content analysis within social sciences.

Results

Varieties of social media big data

The systematic literature review analysis revealed a high degree of variety of SMBD used in the empirical studies; the diversity was expressed in the data structure, data sources, as well as sample size. The variation could be partly explained by the temporal and disciplinary dynamics inherent in the use of SMBD in empirical studies.

The number of articles doubled from 2012 to 2016 (see Table 1). The rapid increase highlights the growing importance of this topic among the global academic community. Also, this growth can be related to the development of skills that the researchers did not have in previous years and thus could be viewed as one of the main obstacles to doing SMBD research.

The analysed articles used 22 social media platforms as data sources. Twitter, due to its accessibility, was not only the dominant source (appearing in 74 articles) but also the most widely used transdisciplinary source. By contrast, 12 of the 22 social media platforms were either discipline-specific (e.g., use of TripAdvisor in tourism studies) or only appeared in one article. Despite ranking second, Facebook's use in the articles as a data source was six times lower than Twitter, and three times lower than the combined 'other social media' data sources. Likewise, several other social media platforms (e.g., Flickr, Wikipedia, YouTube) were prevalent, used in one-third of the cases.

The authors of the articles studied defined and expressed SMBD in a variety of ways. For example, Yang *et al.* (2015) and Zhai *et al.* (2015) consider consumer reviews to be social media due to their user-generated content and because they were combined with social media data. In eight cases, the authors classify a webpage as a social media platform, due to the former combining various functions like social networking with customer reviews, etc. In one case, Bapna *et al.* (2016, pp. 3102, 3104) used a pseudonym to disguise the site's real name: "monCherie.com […] constitutes a typical online dating website and offers features to its users, which are common to most online dating websites".

Studies using more than one data source started to increase in 2016. For example, Ngai *et al.* (2016) use seven, Liu *et al.* (2016) use five, Stephansen and Couldry (2014) use four, and Dehghani *et al.* (2016) use three sources, although this number of social media data sources also appeared in Cord *et al.* (2015). The need to use more than one SMBD source could be because the authors' research questions or hypotheses became more complex so that one data source was insufficient. Also, the research teams became more interdisciplinary, e.g., social scientists teamed up with computer scientists. Interdisciplinary teamwork enables researchers to use new technological tools and various data sources. Our analysis suggests that the most common usage of a range of data sources is to compare research results or to test tools or software. For example, Ngai *et al.* (2016) use seven data sources (Weibo, Baidu, Uwants, Twitter, Facebook, Google Search, and webpages) and illustrated, using text-mining techniques, how social media can aid in capturing useful information from a high variety of information sources. As this was a pilot study testing a system, Ngai *et al.* (2016) consider the usage of several social media platforms justified.

There was also a broad variation between how the articles defined sample size and data structure of SMBD. The sample size ranged from billions of tweets (e.g., Nguyen *et al.*, 2016) to millions of photographs (Park, 2015) to just thousands of user accounts (Durahim & Coşkun, 2015). Furthermore, the units of analysis differed. Griffin (2015) used kilometres, Isari *et al.* (2016) chose individual system logs, Cord *et al.* (2015) geocaches, Goulden *et al.* (2017) unique keywords, Arribas-bel *et al.* (2016) accommodation check-ins and Kolliakou *et al.* (2016) complete sentences of text. Some articles that aimed to test a certain method or tool without content-related results did not indicate a sample size (see, e.g., Aramo-Immonen *et al.*, 2016).

The structure of the data also varied significantly between three types: structured (16%), unstructured (34%), and semi-structured (39%). Structured data involves quantification such as the numerical data about tweets, whereas unstructured refers to non-numerical data such as the text in tweets (e.g., Dehghani *et al.*, 2016). The most popular form was semi-structured, combining structured and unstructured data, such as geo-locations referred to in tweets as well as the textual data in those tweets. Unstructured data often needed to undergo further data analysis techniques to give them any sense or meaning.

Theory relatedness and type of inference

For estimating the developments in scientific practices and the ways scientific knowledge is produced in SMBD studies, we examined the relatedness to theory, type of inference and research design practised in the sampled studies.

We used a 4-point ordinal scale to estimate relatedness to theory, as being either strong or weak, based on the explicit expressions of the authors. As Figure 1a (see p. 48) indicates, more than half of the articles were weakly related to theory, i.e., the formulation of the research questions and hypotheses were not directly based on either theoretical assumptions or empirical studies. In several cases, when problem setting was weakly related to theory, the authors indicated there were no previous studies available in the field (e.g., Kim *et al.*, 2016), and therefore the purpose of the study was innovative (Kern *et al.*, 2014). For example, Williams and Burnap (2016, p. 212) state that the study was a pioneer in the field: “this paper represents the first criminological analysis of an online social reaction to a major crime event”.

Less than half of the articles were coded as being strongly related to theory—i.e., the authors directly stated that the same topics had been studied before or the formulated research questions were raised from previous studies (e.g., Shelton, 2014). For example: “Drawing from these earlier findings, we adopted the term ‘influentials’ to categorise people with extraordinary influence, such as public figures or celebrities" (Araujo *et al.*, 2017, p. 500). Our analysis of associations indicates a tendency that the relatedness to theory even decreased over the period analysed because in 2012–2015 about half of the articles were "rather" or "very strongly" related to theory but in 2016 that figure had fallen to a third.

At the same time, relatedness to theory could also vary across disciplines. Our analysis indicates that a slightly stronger relatedness to theory was expressed by authors who have a background in media and communication studies (Cramer's V = .240, p < .001), whereas a computer scientist's articles displayed a weaker relatedness to theory (Cramer's V = .168, p = .067, this relationship is statistically marginal). Also, the qualitative code intersections analysis revealed some tendencies that relatedness to theory is connected to the discipline. Indeed, articles written by computer scientists are the weakest in the context of relatedness to theory (e.g., 67% of aggregate articles in 2016). By contrast, articles written by either or both media and communication scientists display a generally stronger relatedness to theory. Nevertheless, we found no significant differences comparing the main article sample with those studies not practising SMBD (the relatedness was only one per cent stronger in the case of the main sample, and the relatedness was one per cent weaker in the comparison group).

Besides studying relatedness to theory, we also examined how the research problems were formulated. Our analysis reveals that in one-third of the articles, the authors formulated open research questions and in about one fifth the authors developed clear hypothetical statements based on theoretical or empirical assumptions formulated in previous studies or based on the initial analysis of their empirical data. As Figure 1b indicates, the authors in almost half of the articles did not formulate any research questions or hypotheses. Instead, the authors formulated general research aims or assumptions. In nine articles, the authors formulated both research questions and hypotheses. Based on the qualitative analysis of the articles, we can assume that the authors tried to postulate their hypotheses based on an initial data analysis. However, checking the validity of those hypotheses may actually take longer than expected because scholars, across disciplines, need time to develop skills or tools suitable for the intended analysis using SMBD, although tools for scraping and analysing data from one social media channel cannot always be implemented on other social media platforms. In the context of formulating a research problem, we did not find any significant differences between the main sample of the study and the comparative group, not using SMBD.

As Figure 1c shows, most of the articles in our sample were exploratory (45%) or explanatory/predictive (48%) in their character, whereas in a rather marginal

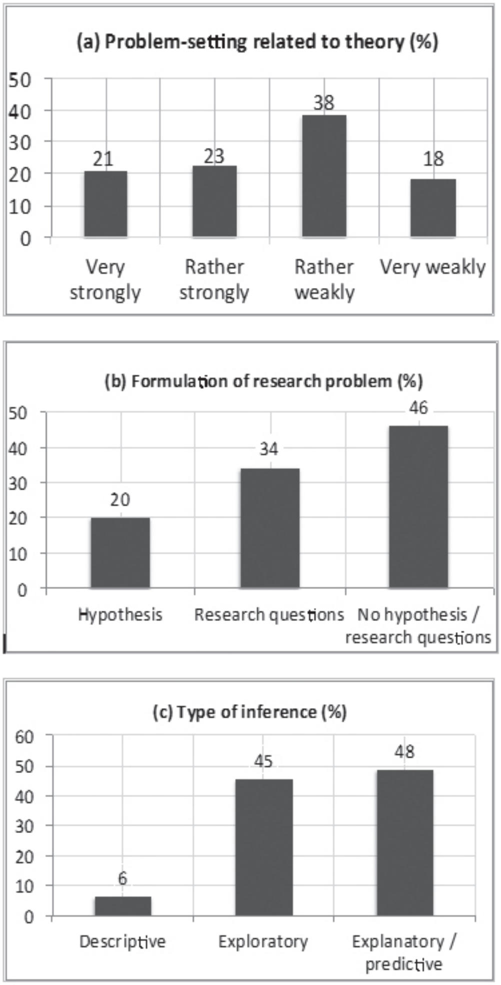

Figure 1. Criteria for estimating developments in social media big data studies

number of cases the articles were coded as descriptive (6%). In the latter context, the authors explicitly indicated their use of the descriptive type of inference (e.g., the descriptive statistics function in particular analysis software was used) (Lewis, 2013). In other cases, the authors explicitly indicated that the type of inference used in the study was exploratory (e.g., Kalyanam *et al.*, 2016) or explanatory (e.g., Arribas-Bel *et al.*, 2015). For example, Jung (2014, p. 52) states: “I use examples from an exploratory case study of geo-tweets in King County, WA, to demonstrate how code clouds can be applied to the production of meanings through qualitative geo-visualisation”. Explanatory or predictive studies were clearly distinct because the authors tried to either explain certain phenomena or demonstrate the usage of a particular method or tool. For example, Arribas-Bel *et al.* (2015, p. 231) writes: “We show that analysis of geo-referenced tweets can shed significant light on physical aspects of the city and on the spatial distribution of urban functions.”

A comparison of the main sample with the comparative group revealed some differences regarding the type of inference. Compared to the main sample, the comparative group displayed descriptive type of inference somewhat more often (13%). In contrast, the main sample displayed predictive or explanatory inference more often (10%) than the comparative group, using more traditional data and related methods. This indicates that novel data sources may offer new opportunities for predicting social processes and explaining unanswered questions.

Analysis methods practised

In order to study the developments in scientific practices and in the ways scientific knowledge is produced in SMBD studies in greater detail, we had a closer look at the analysis methods used in the studies. The methods were estimated and based on the explicitly expressed descriptions in the articles, and subsequently coded using a predefined list of possible methods.

As Figure 2a (see p. 50) indicates, the empirical SMBD studies forming our sample used various research methods to a similar extent. However, classical statistical methods including univariate descriptive techniques (mentioned in 25 articles, e.g., in the form of percentages or mean values), simple multivariate (22 articles, like in the form of crosstabs and association coefficients) as well as advanced multivariate techniques (mentioned in 22 articles, including techniques like regression, factor and cluster analysis, etc.) were dominant (58%; see also Table 2 for a more detailed explanation of the codes used in the analysis). Other applied analysis techniques were computational (42%) and content analysis (44%), such as mainly text structuring based on pre-defined categories (e.g., dictionary approach). Various social media analysis techniques, for example, more automatised techniques (e.g., community detection, opinion mining, etc.) were used to a lesser extent (27%).

Most of the articles displayed the combined use of several analysis techniques, involving statistical and computational ones. For example, Kern *et al.* (2014) combined linear regression with a computational linguistic approach, and Lipizzi *et al.* (2016) wrote that they applied clustering analysis to extract word clusters that potentially correspond to topics in the conversation, but also developed their own Python scripts to automate the analysis process. Other authors used mixed method approaches to show the possibilities for future research related to BD. Shelton *et al.* (2014, p. 178) emphasise: “A quantitative mapping of tweet density ultimately stops short of understanding the complex and polymorphous geographies of such data without also performing a qualitative analysis of the actual tweets and the context in which they are produced”. In addition to the wide variety of quantitative methods and the combination of qualitative methods, the authors used several other topic-specific techniques. For example, Diaconita (2016) used the advanced geo-statistical procedure kriging (Gaussian process regression); Kim *et al.* (2016) used the ARIMA model that captures a

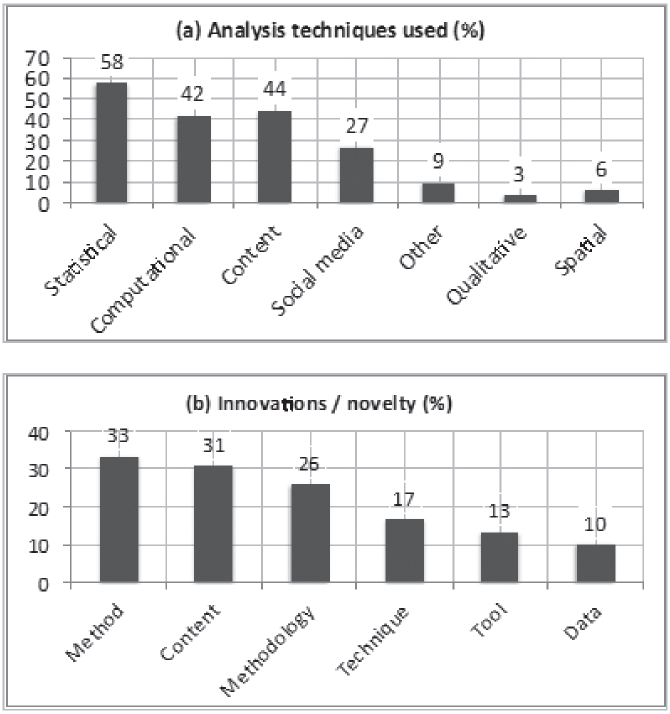

Figure 2. Criteria for estimating developments in social media big data studies

suite of different standard temporal structures in time series data, and Chu *et al.* (2016) used trajectory mining.

The disciplinary backgrounds of the authors of the SMBD studies indicate that computer sciences, followed by 'media and communication', are dominant (see Table 1 on p. 42)[4]. In 17 articles, the authors came from interdisciplinary centres or laboratories (see Table 1). The reasons are multiple and based on the awareness that BD analysis requires a range of discipline-specific skills, although some authors developed and displayed their own analysis tools. A qualitative analysis of code intersections revealed that interdisciplinary teams, compared to mono-discipline teams, used computational techniques somewhat more often. Also, quantitative analysis revealed statistically significant associations between the use of computational techniques and interdisciplinary teams ($F = 64.52$, $p < .001$). Certain temporal dynamics were also visible in the SMBD studies, so that over time a slightly higher disciplinary variation was visible in both the

[4] In the next step of the analysis the disciplinary differences were primarily analysed by comparing media studies and computer sciences as the largest groups represented in this study.

qualitative analysis of code intersections as well as in a quantitative analysis of associations (Cramer's V = .283, p < .01). The growing disciplinary variation could be the result of increasing data competencies but could also be explained by the integration of discipline inherent methods with techniques used in computer sciences.

We also examined in greater detail the innovations that the authors expressed in the articles. The innovation and novelty category contained six sub-categories: methodology, method, technique, tool, data source, and proposed innovations related to content or theory. Each article could appear in several categories and multiple times, depending on how many times the authors pointed out something different, which indicated some kind of novelty or innovation in the study.

Figure 2b indicates that the novelty of the article was linked to the methods used (33%). For example: "We propose a novel method that takes advantage of the global structure of social interactions to alleviate the opinion classification problem in a collective manner" (Li *et al.*, 2016, p. 988). Similarly, in several cases, the innovations were content-related (31%), for example, Neuman (2014, p. 211) emphasised: "We have aspired to demonstrate that [...] big data can serve to refine how the questions themselves are formulated". The analysis of associations also revealed that content-related innovations are increased slightly throughout the four-year period (Cramer's V = .333, p < .001).

The authors also quite often expressed that their articles contributed to research through methodological innovations in a particular study field (26%). For example, Duvanova *et al.* (2016) indicate that their methodological contribution was to integrate BD in the study of mass attitudes and social behaviour, and Arazy *et al.* (2016, p. 805) write: "studies in the area [online production communities] rarely examine clustering reproducibility and assume, rather than validate, that clustering results represent natural groupings in the data". However, instead of focusing on certain ontological shifts, the authors in our sample expressed the belief that greater methodological shifts would occur after new methods and analysis techniques have been developed and implemented.

In addition, several new tools that were developed by the authors were referred to as innovations in the research (13%). For example, Kim (2015) developed an automated software application, the News Diffusion Tracker, that enables researchers to conduct two real-time data-mining tasks concurrently—importing data through an application programming interface and crawling the necessary information from the web page. In some articles, however, the authors only indicated that they had developed a software program but did not name it and described it through the purpose of the particular tool. For example, "this paper analyses the data shadows of Hurricane Sandy through a specially designed software program that collects all geocoded tweets worldwide through the Twitter API" (Shelton *et al.*, 2014, p. 170). Innovation and novelty related to a data source could also occur with the usage of several data sources, e.g., "most of these studies focus on single-source data and do not take into account the fusion of multisource data" (Liu *et al.*, 2015, p. 516). The data source in the broader sense of the word may not have been that innovative, as Obholzer and Daniel (2016, p. 402), for example, used Twitter data, but the authors referred to the use of a particular dataset as "a novel dataset of MEP activity on Twitter [...] we are able to trace the evolution of traditional modes of campaigning".

Discussion and implications

The starting point of this study was the shift induced by digital technology in the ways that social sciences are practised. More broadly, the article aimed to contribute to discussions about the possible move towards the 'end of theory' era raised by Anderson (2008) and discussed in later studies (Kitchin, 2014; Wagner-Pacifici *et al.*, 2015; Resnyansky, 2019; Slota *et al.*, 2020). These approaches warn that due to the emergence of large-scale data sources, the previously dominant approach to science that involves formulating testable hypotheses, testing models, and validating theoretical models is becoming obsolete. Therefore, whereas digital data are believed to be a significant source for explaining social processes, and for managing crises, these data are also challenging the research practices and ways of knowledge production.

The article strived to contribute to these discussions by conducting a systematic literature review based on empirical articles using social media big data (SMBD). This article relies on the empirical research published between 2012 and 2016, the period of rapid increase of SMBD studies. During this period, the principles of big data studies were developed and tested and have formed the foundations for further developments and discussions on knowledge production using SMBD.

This systematic literature review revealed significant discrepancies in how the authors of the articles in our sample understood SMBD both in terms of the definitions, data structure and data size used. These results confirm those of previous studies about variations in SMBD definitions (Gupta *et al.*, 2012; Lupton, 2015), related characteristics (Kitchin & McArdle, 2016), and the emerging shifts in knowledge production and data practices (Iliadis & Russo, 2016). Therefore, SMBD studies constitute not just a potential source for managing abstract social complexity but are employed to brand its power to control (un)predictable social crises. Also, SMBD studies highlight the potential to grasp the variety of social reality available in granular SMBD, and to evaluate the unknown social mechanisms of these disruptions. Therefore, as this original empirical study revealed, the SMBD are not operating exclusively as a source for either or both speeding and scaling up social science's knowledge-making (data as *big*), but extending the potentiality grasping variety of social reality (data as *social*).

The study results indicate that the relationship between the analysed articles and theory were more often weak than strong. Although about half of the analysed articles clearly defined either or both research questions and hypotheses, the remainder have not formulated either of those. A marginal number of studies expressed a descriptive type of inference, as opposed to exploratory, or explanatory/predictive ways of reasoning expressed in most studies. Therefore, although the relatedness to theory was not necessarily strong in the analysed SMBD studies in 2012–2016, our study revealed that it has not led to the widespread use of descriptive ways of inference that Anderson (2008) proposes.

However, our study revealed that the SMBD studies have slightly more often used predictive and explanatory types of inference, compared to more traditional social science studies that do not use BD. Thus, novel data sources may enable researchers to explain and predict social phenomena that were previously not feasible. However, since some articles clearly do lack either a hypothesis or theory, this study shows that further discussions are needed about the role of theory in SMBD studies and in social science methods using large-scale data. SMBD enables predictions in social scientific research and raises questions of understanding, causation evaluating and resolving issues of the underlying mechanisms of social phenomena. Törnberg & Törnberg (2018) suggest that there are alternative approaches to mainstream positivist causation are both theoretical and, as Bengio *et al.* (2019) argue, methodologically introduced. Nevertheless, these are not yet rooted in the social science knowledge-making as highlighted in the initial SMBD studies. Failures in predictive modelling lead to global disruptions like the pandemic coronavirus that emerged in spring 2020, which is an example that would benefit from an evaluation of the causes and effects of social behaviour, based on the SMBD.

The study indicates that the empirical SMBD studies conducted in 2012–2016 are built on the computational approach, which BD studies initially proposed (Chang *et al.*, 2014; Keuschnigg *et al.*, 2017), combined with alternatives that emerged in that period and advanced to full use later, such as digital methods (Rogers, 2013; 2019) and cultural analytics (Manovic, 2017). Our study highlighted that, during 2012–2016, interdisciplinary teams often used the novel computational methods; that the frequency of usage has slightly increased over time. However, traditional and classical social science methods, including qualitative and quantitative, as well as mixed methods approaches, have been used alongside computational methods. Therefore, a prerequisite of both making sense of SMBD and of evaluating the global impact of social behaviour using SMBD are computational skills as well as careful consideration of the socio-cultural context of the collected data and the research phenomena.

The analysis highlighted that content-related innovations, besides proposing novel analysis techniques, methods, data, etc. when analysing SMBD, are on the increase. The availability of an increasing number of data analysis tools combined with the formation of interdisciplinary research teams has enabled previously unresolved research questions to be answered. Studies have shown that a variety of research features are needed to resolve unanswered research problems, like innovations in data analysis techniques, tools and methods (He *et al.*, 2015; Park *et al.*, 2015; Bengio *et al.*, 2019) as well as novel methodologies (Breiman, 2001; Boellstorff, 2013; Manovich, 2017). However, future developments in social scientific theory-building in relation to BD sources are still wide open.

Scholars of preliminary studies (see Halavais, 2013; Veltri, 2017) propose that the growth of machine-learning methods may provide new perspectives for theory development within the discipline of social science. Recent research has been rather critical regarding both the implementation of machine-learning methods (Bengio *et al.*, 2019) and the detection of patterns in the phenomena without explaining the underlying mechanisms. Previous studies have indicated that the cognitive abilities of the data subjects are exploited (Mühlhoff, 2019) for collecting the SMBD. As this empirical study has revealed, further methodological development of SMBD studies assumes the cognitive abilities of the researchers for detecting the social disruptions evident in the data, and for being aware of the human biases in the data.

The turn to a computational social science risks developing passive 'human research machines', where despite the explosive growth of computational methods, researchers have limited opportunities to resolve societal problems. Instead, social science research is challenged to move towards human researchers having an active role in societies in detecting and evaluating the results of machines. This notion follows two assumptions. First, access to granular data is a necessary pre-condition. Second, introducing methods in social sciences enabling the explanation of the mechanisms of cause and effect will be used instead of implementing machine-learning methods exclusively focusing on pattern recognition (see, e.g., Begio *et al.*, 2019).

Future studies should consider several additional factors when empirically explaining the shifts in knowledge production in SMBD studies within the social sciences. As this article focused on the short period (2012–2016) of rapid shifts in SMBD studies, the study covering a longer time period would reveal if and how the tendencies and revealed variations in knowledge production in SMBD studies would develop further. We assume that the implementation of the European General Data Protection Regulation (Regulation (EU)2016/679) and data scandals like the one involving Cambridge Analytica (2018) may have influenced the research practices of SMBD studies. Temporally extending the sample would enable researchers to examine these changes. Future research should also consider in greater detail interdisciplinary differences, in order to evaluate the understandings of cause and effect in empirical studies and theoretical approaches in social science research.

The study enabled us to contribute to the discussions about the practice of using SMBD in knowledge production and calls for further discussions on the role of theory in digital social research and alternative understandings and practices of causation in the era of SMBD. This will assure an explanation of unknown mechanisms of realities as constructed in social media and revealed in social science research.

Acknowledgements

This research was financed by the Development Program of Tallinn University of Technology (2014-2020.4.01.16-0032), the European Commission (Marie Curie 626601), and the Estonian Research Council (IUT20-38). We also thank Stefan

Wehrli, head of the Decision Science Laboratory, for assistance with systematic literature analysis method, and Andreas Diekmann, professor of sociology at ETH Zürich for comments that greatly improved the manuscript. We are also grateful to Marcus Denton of Derettens OÜ for language editing the article.

References

Anderson, C. (2008), 'The end of theory,' *Wired*, vol. 16, no. 7, p. 108.

Bengio, Y.; Deleu, T.; Rahaman, N.; Ke, R.; Lachapelle, S.; Bilaniuk, O.; Goyal, A. & Pal, C. (2019), *A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms*, ArXiv Preprint, arXiv: 1901.10912.

Barberis Canonico, L.; Mcneese, N. J. & Duncan, C. (2018), 'Machine learning as grounded theory: human-centered interfaces for social network research through artificial intelligence,' *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 62, no. 1, pp. 1252–1256. https://doi.org/10.1177/1541931218621287

Boellstorff, T. (2013), 'Making big data, in theory,' *First Monday*, vol. 18, no. 10. https://doi.org/10.5210%2Ffm.v18i10.4869

Bowker, G. C. (2013), 'Data flakes: an afterword to '"Raw Data" Is an Oxymoron',' in L. Gitelman (ed.) *"Raw Data" Is an Oxymoron*, Cambridge: MIT Press. Retrieved from http://www.ics.uci.edu/*vid/Readings/bowker_data_flakes.pdf [accessed Feb 2020]

Bowker, G. C. (2014), 'Big data, big questions: the theory/data thing,' *International Journal of Communication*, vol. 8, no. 2043, pp. 1795–1799. Retrieved from https://ijoc.org/index.php/ijoc/article/view/2190/11568, 5 [accessed Feb 2020]

boyd, d. & Crawford, K. (2012), 'Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon,' *Information, Communication & Society*, vol. 15, no. 5, pp. 662–679. https://doi.org/10.1080/1369118X.2012.678878

Breiman, L. (2001), 'Statistical modeling: the two cultures,' with comments and a rejoinder by the author, *Statist. Sci.*, vol. 16, no. 3, pp. 199–231. https://doi.org/10.1214/ss/1009213726

Brock, A. (2015), 'Deeper data: a response to boyd and Crawford,' *Media, Culture & Society*, vol. 37, no. 7, pp. 1084–1088. https://doi.org/10.1177/0163443715594105

Bruns, A. (2013), 'Faster than the speed of print: reconciling 'big data' social media analysis and academic scholarship,' *First Monday*, vol. 18, no. 10. https://doi.org/10.5210/fm.v18i10.4879

Chang, R. M.; Kauffman, R. J. & Kwon, Y. (2014), 'Understanding the paradigm shift to computational social science in the presence of big data,' *Decision Support Systems*, vol. 63, pp. 67–80. https://doi.org/10.1016/j.dss.2013.08.008

Cohen, P. (2010), 'Digital keys for unlocking the humanities' riches,' *New York Times*, 17 November.

Crawford, K.; Miltner, K. & Gray, M. L. (2014), 'Critiquing big data: politics, e-ethics, epistemology,' Special section introduction, *International Journal of Communication*, vol. 8, pp. 1663–1672.

Dalton, C. & Thatcher, J. (2014), 'What does a critical data studies look like, and why do we care?' *Society & Space*, 12 May. Retrieved from http://societyandspace.org/2014/05/12/what-does-a-critical-data-studies-look-like- and-why-do-we-care-craigdalton-and-jim-thatcher

Dawson, M. (2019), 'Algorithmic culture, networked learning and the technological horizon of theory,' *Technology, Pedagogy and Education*, vol. 28, no. 4, pp. 463–472. https://doi.org/10.1080/1475939X.2019.1643780

DellaPosta, D.; Shi, Y. & Macy, M. (2015), 'Why do liberals drink lattes?' *American Journal of Sociology*, vol. 120, no. 5, pp. 1473–1511. https://doi.org/10.1086/681254

Eklund, L.; Stamm, I. & Liebermann, W. K. (2019), 'The crowd in crowdsourcing: crowdsourcing as a pragmatic research method,' *First Monday*, vol. 24, no. 10. https://doi.org/10.5210/fm.v24i10.9206

Goyder, J. & Leiper, J. M. (1985), 'The decline in survey response: a social values interpretation,' *Sociology*, vol. 19, no. 1, pp. 55–71. https://doi.org/10.1177/0038038585019001006

Gupta, R.; Gupta, H. & Mohania, M. (2012), 'Cloud computing and big data analytics: What is new from database perspective?' in *Data Analytics, Proceedings of First International Conference*, BDA 2012, New Delhi, India, December, pp. 42–61. https://doi.org/10.1007/978-3-642-35542-4_5

Halavais, A. (2013), 'Home made big data? Challenges and opportunities for participatory social research,' *First Monday*, vol. 18, no. 10. https://doi.org/10.5210/fm.v18i10.4876

Hargittai, E. (2020), 'Potential biases in big data: omitted voices on social media,' *Social Science Computer Review*, vol. 38, no. 1, pp. 10–24. https://doi.org/10.1177/0894439318788322

He, W.; Wu, H.; Yan, G.; Akula, V. & Shen, J. (2015), 'A novel social media competitive analytics framework with sentiment benchmarks,' *Information & Management*, vol. 52, pp. 801–812. https://doi.org/10.1016/j.im.2015.04.006

Helbing, D. (2013), 'Globally networked risks and how to respond,' *Nature*, vol. 497, no. 7447, pp. 51–59. https://doi.org/10.1038/nature12047

Hindman, M. (2015), 'Building better models: prediction, replication, and machine learning in the social sciences,' *The Annals of the American Academy of Political and Social Science*, vol. 659, no. 1, pp. 48–62. https://doi.org/10.1177/0002716215570279

Hofman, J. M.; Sharma, A. & Watts, D. J. (2017), 'Prediction and explanation in social systems,' *Science*, vol. 355, no. 6324, pp. 486–488. https://doi.org/10.1126/science.aal3856

Iliadis, A. & Russo, F. (2016), 'Critical data studies: an introduction,' *Big Data & Society*, vol. 3, no. 2, pp. 1–7. https://doi.org/10.1177/2053951716674238

Kennedy, H.; Elgesem, D. & Miguel, C. (2017), 'On fairness: user perspectives on social media data mining,' *Convergence*, vol. 23, no. 3, pp. 270–288. https://doi.org/10.1177/1354856515592507

Keuschnigg, M.; Lovsjö, N. & Hedström, P. (2017), 'Analytical sociology and computational social science,' *Journal of Computational Social Science*, vol. 1, pp. 3–14. https://doi.org/10.1007/s42001-017-0006-5

Kitchin, R. (2014), 'Big data, new epistemologies and paradigm shifts,' *Big Data & Society*, vol. 1, no. 1, pp. 1–12. https://doi.org/10.1177/2053951714528481

Kitchin, R. & McArdle, G. (2016), 'What makes big data, big data? Exploring the ontological characteristics of 26 datasets,' *Big Data & Society*, vol. 3, no. 1, pp. 1–10. https://doi.org/10.1177/2053951716631130

Knight, W. (2019), 'An AI pioneer wants his algorithms to understand the 'why',' *Wired*, 10 August. Retrieved from https://www.wired.com/story/ai-pioneer-algorithms-understand-why/?fbclid=IwAR0YmGkLKgCuqQmllBldZUtG66HFhoQYVztTeF3Rmx2hB9F9tCNlLg0Pybc&mbid=social_facebook&utm_brand=wired&utm_campaign=falcon&utm_medium=social&utm_social-type=owned&utm_source=facebook [accessed Mar 2020]

Lazega, E. & Snijders, T. A. B. (2016), *Multilevel Network Analysis for the Social Sciences: Theory, Methods and Applications*, 1st ed., Methodos Series, Methodological Prospects in the Social Sciences: vol. 12. Retrieved from http://sfx.ethz.ch/sfx_locater?sid=ALEPH:EBI01&genre=book&isbn=9783319245201 [accessed Mar 2020] https://doi.org/10.1007/978-3-319-24520-1

Lupton, D. (2015), 'The thirteen Ps of big data,' *This Sociological Life*, A blog by sociologist Debora Lupton. Retrieved from https://simplysociology.wordpress.com/2015/05/11/the-thirteen-ps-of-big-data/ [accessed Mar 2020]

Manovich, L. (2011), 'Trending: the promises and the challenges of big social data,' in M. K. Gold (ed.) *Debates in the Digital Humanities*, Minneapolis, MN: University of Minnesota Press, pp. 460–475. https://doi.org/10.5749/minnesota/9780816677948.003.0047

Manovich, L. (2017), 'Cultural analytics, social computing and digital humanities,' in M. T. Schäfer & K. van Es (eds.) *The Datafied Society Studying Culture through Data*, Amsterdam: Amsterdam University Press, pp. 55–69. https://doi.org/10.1515/9789048531011-006

Marres, N. & Gerlitz, C. (2015), 'Interface methods: renegotiating relations between digital social research, STS and sociology,' *Sociological Review*, vol. 64, pp. 21–46. https://doi.org/10.1111/1467-954X.12314

Mason, W.; Vaughan, J. W. & Wallach, H. (2014), 'Computational social science and social computing,' *Machine Learning*, vol. 95, no. 3, pp. 257–260.

Masson, E. (2017), 'Humanistic data research: an encounter between epistemic traditions,' in M. T. Schäfer & K. van Es (eds.) *The Datafied Society Studying Culture through Data,* Amsterdam: Amsterdam University Press, pp. 25–37. https://doi.org/10.1007/s10994-013-5426-8

McFarland, D.; Lewis, K. & Goldberg, A. (2016), 'Sociology in the era of big data: the ascent of forensic social science,' *The American Sociologist*, vol. 47, no. 1, pp. 12–35. https://doi.org/10.1007/s12108-015-9291-8

Mühlhoff, R. (2019), 'Human-aided artificial intelligence: or, how to run large computations in human brains? Toward a media sociology of machine learning,' *New Media & Society*, 6 November. https://doi.org/10.1177/1461444819885334

Neff, G.; Tanweer, A.; Fiore-Gartland, B. & Osburn, L. (2017), 'Critique and contribute: a practice-based framework for improving critical data studies and data science,' *Big Data*, vol. 5, no. 2, pp. 85–97. https://doi.org/10.1089/big.2016.0050

O'Brien, D. T.; Sampson, R. J. & Winship, C. (2015), 'Ecometrics in the age of big data: measuring and assessing "broken windows" using large-scale administrative records,' *Sociological Methodology*, vol. 45, no. 1, pp. 101–147. https://doi.org/10.1177/0081175015576601

Olshannikova, E.; Olsson, T.; Huhtamäki, J. & Kärkkäinen, H. (2017), 'Conceptualizing big social data,' *Journal of Big Data*, vol. 4, no. 1, pp. 1–19. https://doi.org/10.1186/s40537-017-0063-x

Olteanu, A.; Castillo, C.; Diaz, F. & Kıcıman, E. (2019), 'Social data: biases, methodological pitfalls, and ethical boundaries,' *Frontiers of Big Data*, vol. 2, no. 13. https://doi.org/10.3389/fdata.2019.00013

Park, G.; Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Kosinski, M.; Stillwell, D. J.; … Seligman, M. E. P. (2015), 'Automatic personality assessment through social media language,' *Journal of Personality and Social Psychology*, vol. 108, no. 6, pp. 934–952. https://doi.org/10.1037/pspp0000020

Puschmann, C. & Burgess, J. (2014), 'Big data, big questions: metaphors of big data,' *International Journal of Communication*, vol. 8, pp. 1690–1709.

Regulation (EU)2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (Data Protection Directive), *OJ* L119, 4.5.2016, implementation date 25.5.2018.

Resnyansky, L. (2019), 'Conceptual frameworks for social and cultural Big Data analytics: answering the epistemological challenge,' *Big Data & Society*, vol. 6, no. 1. https://doi.org/10.1177/2053951718823815

Rieder, B. (2016), 'Big data and the paradox of diversity,' *Digital Culture & Society*, vol. 2, no. 2. https://doi.org/10.14361/dcs-2016-0204

Rogers, R. (2013), *Digital Methods*, Cambridge, Ma: MIT. https://doi.org/10.7551/mitpress/8718.001.0001

Rogers, R. (2019), *Doing Digital Methods*, Los Angeles: Sage.

Schäfer, M. T. & van Es, K., eds. (2017), *The Datafied Society: Studying Culture Through Data*, Amsterdam: Amsterdam University Press. https://doi.org/10.5117/9789462981362

Scheinfeldt, T. (2012), 'Sunset for ideology, sunrise for methodology,' in M. K. Gold (eds.) *Debates in the Digital Humanities*, Minneapolis, MN & London: University of Minnesota Press, pp. 124–126. https://doi.org/10.5749/minnesota/9780816677948.003.0014

Shah, D. V.; Cappella, J. N. & Neuman, W. R. (2015), 'Big data, digital media, and computational social science,' *The Annals of the American Academy of Political and Social Science*, vol. 659, no. 1, pp. 6–13. https://doi.org/10.1177/0002716215572084

Sivarajah, U.; Kamal, M. M.; Irani, Z. & Weerakkody, V. (2017), 'Critical analysis of Big Data challenges and analytical methods,' *Journal of Business Research*, vol. 70, pp. 263–286. https://doi.org/10.1016/j.jbusres.2016.08.001

Slota, S. C.; Hoffman, A. S.; Ribes, D. & Bowker, G. C. (2020), 'Prospecting (in) the data sciences,' *Big Data & Society*, vol. 7, no. 1. https://doi.org/10.1177/2053951720906849

Spitz, G.; Niles, F. L. & Adler, T. J. (2006), *Web-based Survey Techniques*, Transit Cooperative Research Program (TCRP) Synthesis 69, Washington, DC: Transportation Research Board, pp. 1–104. https://doi.org/10.17226/14028Transportation Research Board.

Szolnoki, G. & Hoffmann, D. (2013), 'Online, face-to-face and telephone surveys—comparing different sampling methods in wine consumer research,' *Wine Economics and Policy*, vol. 2, no. 2, pp. 57–66. https://doi.org/10.1016/j.wep.2013.10.001

Thatcher, J.; O'Sullivan, D. & Mahmoudi, D. (2016), 'Data colonialism through accumulation by dispossession: new metaphors for daily data,' *Environment and Planning D: Society and Space*, vol. 34, no. 6, pp. 990–1006. https://doi.org/10.1177/0263775816633195

Thylstrup, N. B.; Flyverbom, M. & Helles, R. (2019), 'Datafied knowledge production: introduction to the special theme,' *Big Data & Society*, vol. 6, no. 2. https://doi.org/10.1177/2053951719875985

Tukey, J. W. (1962), 'The future of data analysis,' *The Annals of Mathematical Statistics*, vol. 33, no. 1, pp. 1–67. https://doi.org/10.1214/aoms/1177704711

van Es, K.; Lopez Coombs, N. & Boeschoten, T. (2017), 'Towards a reflexive digital data analysis,' in Schäfer & van Es (eds.) *The Datafied Society Studying Culture through Data*, Amsterdam: Amsterdam University Press, pp. 171–183. https://doi.org/10.1515/9789048531011-015

Veltri, G. A. (2017), 'Big Data is not only about data: the two cultures of modelling, *Big Data & Society*, vol. 4, no. 1, pp. 1–6. https://doi.org/10.1177/2053951717703997

Weltevrede, E. J. T. (2016), *Repurposing Digital Methods: The Research Affordances of Platforms and Engines*, PhD thesis. Retrieved from http://dare.uva.nl/personal/pure/en/publications/repurposing-digital-methods-the-research-affordances-of-platforms-and-engines(aaaa9bb3-8647-41df-954c-2bb1e9f15d77).html [accessed Mar 2020]

Zimmer, M. & Profes, N. J. (2014), 'A topology of Twitter research: disciplines, methods, and ethics,' *Aslib Journal of Information Management*, vol. 66, no. 3, pp. 250–261. https://doi.org/10.1108/AJIM-09-2013-0083

Anu Masso is an associate professor of big data in social sciences at Ragnar Nurkse Department of Innovation and Governance, Tallinn University of Technology, and a senior researcher at the Institute of Social Studies, University of Tartu. Her research focuses on the theory of social transformations, spatial mobility and socio-cultural consequences of big data. Her recent work concerns misconceptions regarding social diversities in data technologies. She is known for her extensive work on social science methods and methodologies. Her publications include journal articles, e.g., in *European Journal of Cultural Studies*, *European Societies*, *Geopolitics*, *Journal of Baltic Studies*, *Journal of Ethnic and Migration Studies*, *Information Systems Frontiers and Population*, *Space and Place*, and *The Routledge International Handbook of European Social Transformations* (co-edited with Peeter Vihalemm and Signe Opermann, 2018).

Maris Männiste is a doctoral student and information systems assistant at the Institute of Social Studies, University of Tartu, Estonia. Her research interests include perceptions of privacy, datafication practices, peoples' attitudes, and everyday practices related to personal informatics systems (e.g., self-tracking, self-monitoring, but also monitoring children). Her current research focuses on Estonian pioneer data experts' attitudes, experiences, and ideals of algorithmic governance.

Andra Siibak is a professor of media studies and a program director of the Media and Communication doctoral program at the Institute of Social Studies, University of Tartu, Estonia. Her main field of research has to do with the opportunities and risks surrounding internet use, social media usage practices, datafication of childhood, privacy, and new media audiences. She has published more than 80 peer-reviewed articles in international journals and edited collections on young people's practices online: e.g., self-presentation on social media; teacher/parental/sibling mediation of young people's internet use; privacy strategies and imagined audiences on social media; touch-screen usage of toddlers; digital literacies; etc.