Editorial

Big Data & Society
July–December 2015: 1–11
© The Author(s) 2015
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/2053951715613810
bds.sagepub.com
SAGE

Ontologies, methodologies, and new uses of Big Data in the social and cultural sciences

Robin Wagner-Pacifici[1], John W Mohr[2] and Ronald L Breiger[3]

Abstract

In our Introduction to the *Conceiving the Social with Big Data* Special Issue of *Big Data & Society*, we survey the 18 contributions from scholars in the humanities and social sciences, and highlight several questions and themes that emerge within and across them. These emergent issues reflect the challenges, problems, and promises of working with Big Data to access and assess the social. They include puzzles about the locus and nature of human life, the nature of interpretation, the categorical constructions of individual entities and agents, the nature and relevance of contexts and temporalities, and the determinations of causality. As such, the Introduction reflects on the contributions along a series of binaries that capture the dualities and dynamisms of these themes: Life/Data; Mind/Machine; and Induction/Deduction.

Keywords

Abduction, Big Data, computational social sciences, digital humanities, field theory, mind/machine, ontology, text mining, topic models

This issue of *Big Data & Society* contains 18 essays by scholars from across the social sciences and humanities. These scholars were invited to reflect upon experiences they have had in working with Big Data that led them to confront their own implicit assumptions about the nature of *the social* and the sometimes contrasting assumptions that are embedded within the methodological practices of the computational sciences. Our goal in inviting these essays was to highlight some of the ways that Big Data is having an impact on the social and cultural sciences, but to do so in a manner that moves beyond the spectacle of new visualizations and empirical findings from massively large datasets to ask a set of deeper questions about how encounters with Big Data are unsettling the intellectual foundations of the social scientific and humanistic disciplines. To this end we asked some of the leading researchers from across that broad arena to reflect on how their turn to Big Data led them to reexamine their own foundational expectations about *how to study the social*.

The contributors have done exactly what we asked, and much more. They have described their own and others' encounters with the new computational premises of Big Data. They have linked these observations to experiences of empirical research while writing in a theoretically engaged and reflexive manner about how new methodological and ontological situations are changing the contours of their fields. Some have commented on the ways in which their own disciplinary assumptions were disrupted, transformed, or expanded by their move into studies of Big Data. Others use the occasion to issue a warning about how Big Data carries its own powerful (and often pernicious) assumptions about the nature of *the social*, assumptions that often need to be excavated and compensated for prior to employing the

[1] Department of Sociology, The New School, New York, NY, USA
[2] Social Sciences & Media Studies, University of California, Santa Barbara, CA, USA
[3] School of Sociology, University of Arizona, Tucson, AZ, USA

Corresponding author:
John W Mohr, Department of Sociology, University of California, Santa Barbara, CA, USA.
Email: mohr@soc.ucsb.edu

data reliably. Finally, many of our contributors approvingly describe how the shift toward Big Data may lead to radical transformations in the academic fields in which they work.

In reading our authors' responses to the queries that we posed, we learned that there are a great many ways in which working with Big Data has led to a challenging of the core assumptions that undergird scholarly work in the social sciences and humanities. In many ways this amounts to an ontological challenge. Here we think of an ontology in the same sense that Martin Ruef (1999) intended when he described ontologies as ''systems of categories, meanings and identities within which actors and actions are situated' (p. 1403).[1] In the case of the shift to the analysis of Big Data, there are many constituent elements that make up the complex ontology of *the social*—entities, agents, acts, causes, meanings, temporalities, and contexts. We find that many of these are being actively renegotiated in some way as scholars recalibrate to adjust to a new style of science where different scales of analysis are being used and new kinds of social situations are being measured. Equally important, however, are the challenges to deeply held assumptions about the nature of scholarly research. In the essays that follow, it is not uncommon to find personal accounts of how authors' encounters with Big Data have transformed their conceptions of scientific practice itself. We think of these experiences as expressions of a particular kind of duality relationship, what we call a duality of ontology and methodology. Simply stated, scholars' understandings of *the social* influence how they study Big Data at the same time as how they study Big Data influences what they think about the nature of *the social*.[2]

In this introductory essay we provide a short summary of the main issues we have encountered in these papers. Two fundamental questions keep coming into focus for us as we work across these texts. One involves the nature of the relationship between where and how human life happens and what exactly Big Data can tell us about that. A second overarching question concerns the nature of interpretation—just what is it? What should it consist of? When should it come into play in Big Data research? Both of these issues make appearances in many of these essays.[3] As a way to unpack these matters more systematically we focus here on three analytic binaries: (1) Life/Data, (2) Mind/Machine, and (3) Induction/Deduction. We use these binaries in a bid for analytic clarity but also because the binary juxtapositions reflect our own theoretical and methodological propensity to pay close attention to those places within these essays where two domains or types or levels of social life intersect, articulate and co-constitute each other. Finally we will briefly discuss five of the ontological categories that we find being actively negotiated in this intellectual space: (1) what is a thing? (2) what is an agent? (3) what is time? (4) what is context? And (5) what is causality? As we point out in our review of these issues below, the authors of this special issue not only address these questions and portray problems with the current state of the art, but in varying degrees they also propose solutions and remediations.

Binary #1: Life/Data

One of the key issues foregrounded again and again in these essays is how Big Data is able to, or is assumed to be able to, stand in for social life itself. As against these kinds of extrapolations, several authors question the idea that Big Data offers us a neutral presentation of social reality.

In her article, ''Small Decisions with Big Impact for Data Analytics,'' Jana Diesner demonstrates that, against the conventional wisdom in many quarters (including much of sociology), failure to pre-process data in Big Data analysis can have major negative consequences. Before the Big Data are able to (as some have put it, though not this author) ''speak for themselves,'' the analyst must confront issues concerning data quality and also how results are to be contextualized. This requires ''a deep understanding of content domains, their features, and trustworthy findings.'' As reflected in the essay's title, small decisions concerning data construction and data preparation, decisions that are often not given careful attention and about which there are few or no ''best practices,'' can have enormous (often undesired) impact on the results of Big Data analysis.[4]

Julia Adams and Hannah Brückner provide a classic example of the problem in their article, ''Wikipedia, Sociology, and the Promise and Pitfalls of Big Data''. They report on some of the ways that the social system operating at the heart of the Wikipedia website can promote systematic distortions in the image of the social world. Adams and Brückner explore the inner politics of Wikipedia and show that there are many biases in the production process and in the database itself. They point to their study of the underrepresentation of women scholars in Wikipedia's catalogue of ''living American sociologists'' as one example.

Patrick Park and Michael Macy (''The Paradox of Active Users'') give examples of the types of distortions that can occur in the analysis of Twitter data. Their essay focuses on how the presence of more active users creates a series of systematic biases in the way Twitter information is created and subsequently analyzed. Daniel McFarland carries this theme over in his essay, ''Big Data and the Danger of Being Precisely Inaccurate,'' which is concerned with how very large datasets, especially "found" data resulting e.g. from online observations of participants on the Internet (in contrast to data that comes into existence as "representative" of a population as survey or census), can lead to activity bias and to biases resulting from treating "found" data as if it were a census. Due to the large size of the (big) dataset, almost all findings are statistically "significant" in the conventional sense, thus leading to McFarland's characterization of the enterprise as being "precisely inaccurate".[5] He also describes his progress in using data segmentation techniques to fix this. His goal is to show "how nonhomogeneous subpopulations of Internet participants might be identified in order to understand how various distinctive groups of participants compose the internet."

Kevin Lewis ("Three Fallacies of Digital Footprints") writes about this as a general problem afflicting any study of our so-called "digital footprints." Lewis identifies three critical issues that plague the field (he describes them as ontological fallacies)—the assumptions that these data contain information about entire populations; that they record "naturalistic" behavior; and that they can be analyzed without consideration of context. Ryan Shaw goes a step further. In his article, "Big Data and Reality," he argues that if we are to make use of Big Data to study social life then we need to appreciate that the lens will be warped in specific ways by the particular compressions that software systems impose upon social life as they transform it into Big Data. This is especially complicated, Shaw reminds us, by the fact that these software environments are layered, dynamic and evolving practical affordance systems.[6] Shaw puts out a call for a reflexive social history of Big Data that recognizes and appreciates how the materiality of data is produced out of this dynamic interplay of social life and social software.

Binary #2: Mind/Machine

Another theme that emerges repeatedly concerns how human readers and computational techniques of reading provide interpretations of meaning. Following on this are questions about just what is the best interpretive practice or technique to use for Big Data research or, as Ronald Breiger puts it in his essay, "Scaling Down": "what kind of a reader do we think a computer can be or should be?"

The question comes up frequently in the world of topic modeling, a methodology that has become the most widely diffused of the new text mining tools that are currently in use by social scientists and digital humanists (DiMaggio et al., 2013; Goldstone and Underwood, 2014; Liu, 2013; Mohr and Bogdanov, 2013). In this special issue, Rachel Buurma ("The Fictionality of Topic Modeling: Machine Reading Anthony Trollope's Barsetshire Series"), Paul DiMaggio ("Adapting Computational Text Analysis to Social Science (and Vice Versa)"), and Sophie Mützel ("Facing Big Data: Making Sociology Relevant") all describe their experiences with topic models. These methods make use of information about the co-occurrence of words within and across documents in order to develop clusters of terms that are hypothesized to represent common "topics" that span across a textual corpus (Blei, 2011, 2012a, 2012b; Blei et al. 2003). But topic models and other "bag of words" techniques approach text analysis from the position of a linguistically ignorant reader—the computational model knows nothing about syntax, semantics, or phonology, etc.[7] Ted Underwood, in his essay "The Literary Uses of High-Dimensional Space," calls them a "blunt hermeneutic instrument." Alan Liu (2013) describes these as 'tabula rasa' interpretations. This is generally regarded as a lamentable cost of the technique that is offset by its benefits, such as being able to quickly and automatically code and mark large textual corpora according to the primary topics of conversation.[8]

But Rachel Buurma argues that it is, paradoxically, the very denaturalized quality of a topic model's analysis that makes it especially valuable as a method of textual interpretation for literary scholars. Buurma recounts her experience using topic models to analyze the set of six Barsetshire novels by Anthony Trollope. Buurma writes, "the algorithm's very inability to know anything about the Victorian novel as a form or genre lets it point us to a very different model of the social than the kind of formal totality held out to us by the novel theory we have." The result is a method that excels at finding traces of other stylistic conventions mixed in among the rest, in seeing precisely that which we don't see. Buurma argues, "...all topics generated from literary corpuses can help take us back to earlier imaginary forms and versions—discarded drafts that authors might have written but didn't, outmoded genres that are fragmentarily recycled within new forms. Topic modeling may be most useful for humanists when we use it this way, as a kind of uncanny, shifting, temporary index to the works we know best, rather than trying to imagine it, as we too often do, only as telling us something about the stable 'contents' of large literary corpora." In short, these kinds of insights are precisely the sort of properties of texts that a conventional close reader has a hard time discerning. They are surprising discoveries that operate outside of the conventional model for humans to think about and to read novels.

In contrast, other types of computational approaches to text analysis, including many natural language processing (NLP) projects, do their best to replicate the more nuanced and context dependent understandings that human readers bring to a text. These types of projects have the ultimate goal of creating algorithmic approaches to analyzing a text that operate like a language-savvy human reader. Our own essay in this special issue, "Towards a Computational Hermeneutics," describes some of these types of programs. The essay recounts our attempts (in a collaboration with computer scientist Petko Bogdanov) to link hermeneutic traditions of textual analysis (Wagner-Pacifici, 2010) to some of the new text mining technologies. Our goal was to replicate the literary theorist Kenneth Burke's (1941, 1945) approach to rhetorical analysis in our study of a corpus of US National Security Strategy texts. We end our essay with a call for a computational hermeneutics, a scholarly project that would embrace a more poetic approach to reading textual corpora, one that appreciates the ambiguities and the contradictions of texts, one that looks for and analyzes the nuances of stylistic, semantic and rhetorical expressions.

Between these two extreme positions, there are many places along a continuum of how language-experienced or language-naive a given computational approach may be. The trade-offs associated with these various strategies have led a number of our contributors to search for a productive middle ground between mind and machine. Ted Underwood writes, "For staging a demonstration of new methods, it has been a rhetorical advantage that topic models are unsupervised—it says, in effect, 'nothing up my sleeve.' But as we move beyond the demonstration stage of text analysis, supervised predictive models may become more important". Underwood writes about his own research and his goal of hitting the sweet spot "between the two poles (where) there's a rich array of supervised methods that can use unstructured text to help us understand specific boundaries."

Binary #3: Induction/Deduction

Another contested theme is how social scientific and humanistic analysis ought to be conducted. Here we found a frequent binary opposition balancing the virtues of a more inductive against a more deductive model of science, although the lines between the two are often blurred.

According to one line of thinking, old school social science (based on little data) is necessarily more deductive by virtue of the fact that it begins from a set of explicit ideas that must then be funneled into a limited number of variables that are then linked together by measures of association. Success comes in the form of statistically significant associations suggesting the findings can be used to deduce the true state of the relation between these properties in the population.[9] Monica Lee and John Levi Martin are highly critical of this traditional vision of social science. In their article, "Surfeit and Surface," they write, "Modern social statistics starts with the creative (if somewhat insane) idea that all members of a population are actually replications of a single underlying ideal type, the average man—just with a little bit of random error here and there that we can cleverly 'correct for' with a wave of the statistical wand." The alternative "new school" (e.g. Big Data) approach leans more heavily toward induction. As Lee and Martin then contend, the richness of Big Data creates the possibility for a very different kind of investigative strategy: "the population can be disaggregated and flexibly explored to answer a number of different questions instead of mean-averaged out to answer one poorly posed and unchanging question." They call for a more open-ended approach to social science, an approach that has more of an appreciation of how social fields operate as complex systems of relational processes (Martin, 2011), and in that spirit they urge "shaking off [of] the last commitments to causal explanation and shifting towards cartography—the construction of question-independent, though theoretically organized, reductions of information to make possible the answering of many questions."

Amir Goldberg is on the same page in his essay "In Defense of Forensic Social Science". He writes, "Ironically it appears that theoretically informed hypothesis testing can lead to the narrowing of one's sociological imagination." In his essay Goldberg seeks to chart a course beyond what he calls the "categorical myopia" of traditional survey research by proposing an alternative approach that he describes as a "forensic social science: the careful compilation of evidence from unstructured digital traces as a means to generate new theories." Neither deductive nor purely inductive, Goldberg thinks of this as an abductive approach to social science that moves back and forth between theory and inductive exploration (Timmermans and Tavory, 2014). Goldberg writes, with the advent of Big Data, "we no longer need to come to the crime scene [so to speak] with an idea about the identity of the killer." Analysts are freed from thinking about the most cost-effective data that one needs to collect, and can turn instead to "figuring out how to structure a mountain of data into meaningful categories of knowledge."

Ted Underwood reflects upon this type of statistical methodology also. He writes, "more fundamentally I'm talking about what Leo Breiman called a new 'culture' of statistical modeling—a culture that doesn't assume we need to craft a model by deciding in advance which variables matter for a given problem (Breiman, 2001). Instead, it's now possible to start a modeling process by admitting that we don't know which variables matter. We don't really know, for instance, whether murders and mansions were the key elements distinguishing literary genres. But we can still attempt to model genre, by gathering thousands of variables and asking a learning algorithm to identify the variables that do reliably distinguish examples of different genres.''

As this comment begins to suggest, the rethinking of standard approaches to social science and the humanities is not just about how to take best advantage of a new super-abundance of information. Equally important for our contributors are the encounters they have had with entirely different intellectual gestalts, the sort that lie buried in the practices and implicit theories of the computational sciences which begin from a very different understanding about how to advance the frontiers of knowledge. A frequent trope in these conversations concerns the logic of *machine learning* as a different model for science. In his essay, Paul DiMaggio walks us through the working assumptions of computer scientists who come from a machine learning tradition. In contrast to social scientists who ''customarily obsess over causality and rely on formal tests of statistical significance, computer scientists using supervised models focus on results.'' DiMaggio explains, ''It is not that they care less about getting models right; rather they understand 'getting it right' in a different (and I am beginning to suspect more useful) way than do most social scientists, focusing on model plausibility, utility, and descriptive, as opposed to causal, validation.'' DiMaggio writes, ''Ultimately, however, the computer-science perspective is liberating, as it forces us to recognize real interpretive uncertainty and seek out appropriate and substantively relevant forms of validation fitted to specific research goals.''

Embedded within these sets of analytical binaries that dynamically structure the papers are several deeply ontological themes that our Big Data practitioners reflect upon—what is a thing? what is an agent? what is time? what is context? what is cause? We introduce these themes here.

Theme #1: What is a thing?

When we try to get computers to react to the social world, we immediately run up against some of the most basic and simple ontological dilemmas. Jana Diesner's essay is insightful on this score. She describes some of the practical problems of text analysis that involve having to decide what or who is an ''entity'' and how this creates enormous dilemmas in processing something as seemingly simple as the presence or absence of middle initials in the listing of a name. Diesner describes her efforts to build *disambiguation devices* such as ''Relation Extraction'' techniques that can assist analysts who are attempting, for example, to ''identify the structure of covert and hard-to-access networks, networks for which only archival records exist, e.g. bankrupt companies or groups that have ceased to exist, and networks that primarily interact through digitized communication.''

Timothy Hannigan also writes about the challenges of word disambiguation, or the problem of ''how to decide between different meanings of the words'' in his textual corpus. Hannigan and his colleagues analysed more than 30,000 newspaper articles with coverage of the ''2009 British 'Members of Parliament Expenses Scandal'''. The goal was to extract those bits of news stories that applied to each of the separate Members of Parliament. As Hannigan explains in his essay, ''Close Encounters of the Conceptual Kind: Disambiguating Social Structure from Text'': ''The computational linguistics analysis yielded a set of millions of noun phrases.'' And so they had to devise an efficient procedure for having the computer identify Members of Parliament and sorting out the text that was associated with each one. Hannigan points to a host of broader theoretical questions that are brought to the surface by these kinds of practical dilemmas. ''In resolving words to match concepts representing units of social structure, is it the particular meaning (sense) of a word that matters, or the linguistic role it plays?'' These and other matters are theoretically charged, and as Hannigan explains, ''The issue is not about more powerful machines, or statistical tools. Rather, it is about employing a set of competencies to enable a qualitatively different type of text analysis than what is done currently with Content Analysis.''

These strategies turn out to have both practical and theoretical undertones. If our analytical goal is to tease out social structure through networks of entities and their relations, we have to begin somewhat earlier by asking whether we will focus on individual words or on word phrases. Here, as with so much of this work, we have to begin by asking ourselves: just what is our basic ''ontological unit''? DiMaggio notes that ''(o)ne promising class of methods involves using natural language processing tools to separate terms that should be treated differently (e.g. distinguishing homonyms through part-of-speech identification) or uniting terms that should be treated as the same (e.g. using named entity recognition to identify organizations or actors that may be referred to in slightly different ways).''

But this is not just a problem for text recognition; this is the kind of ontological conundrum that is coming into focus in a whole variety of new ways. Peter Bearman describes a similar problem for historians in his article, ''Big Data and Historical Social Science'': the question of what is an era. Traditionally a matter of judgment and debate, increasingly the availability of digital archives has created new opportunities for empirical study. Bearman writes, ''how should we case historical event sequences; that is, how should we induce periods? Newly available textual corpora spanning long periods of time provide a powerful setting for identifying turning points.'' Rachel Buurma gives a different kind of example when she asks: what is a novel? She notes, ''Critics have a particularly difficult time accounting for small groups of related novels like the Barsetshire series, in part because theories of the novel almost always assume the single novel as the unit of analysis.'' She prefers to read the Barsetshire series as ''semi-detached'' and their ''social relations as more partial and unfinished.''

Kevin Lewis raises a different, but related issue—how do we distinguish between data that reflects events rather than individuals as the unit of analysis? Lewis writes that ''many digital datasets have events (e.g. transactions or communications) rather than individuals as the primary unit of analysis, some people may 'turn up' in our data more or less frequently than others—and it is unclear what these differences actually capture and whether (if it all) it is a distinction we care about''. Ryan Shaw gives us an historical perspective on this problem. He recounts the story of William Kent, who in 1978 described the complexities of data storage. Kent wrote that ''deciding how to store data involves answering a number of interrelated questions. What is one thing?'' How many things are there? What kinds of things are there? How real are they? How long do they last? ''Kent emphasized that there are no right answers to such questions: different people in different contexts with different goals will choose different answers as they construct their data models. Data models are practical tools; like maps, they are ''correct'' to the extent that they get you where you want to go.''

As an example, Shaw considers Twitter data. ''Treated as the subject of a scientific inquiry, 100 million tweets are a series of observations generated by the same implicit and unchanging mechanism, the nature of which is to be discerned via statistical generalization from that series. Treated as the subject of a historical inquiry, 100 million tweets are an assembly of individual utterances, the circumstantial relations among which must be discerned through a process of mutual criticism and interpretation.''

Theme #2: What is an agent?

As Daniel McFarland reminds us in his essay, a particularly tricky problem in both social science and computer science is knowing when an entity is an agent. Is it an individual or a group, a human or a robot, a one-time user or a frequent user? McFarland describes his strategies for working with Big Data that focus in on discerning ''recency, frequency, and value'' as a way to disentangle this puzzle of identifying agents.

Patrick Park and Michael Macy discuss the types of distortions that occur in assumptions about agency in the analysis of Twitter data. Their essay focuses on how active users create systematic biases in the way Twitter information is created and studied. The basic problem is that active users are overrepresented, whereas analysts often take all representations of language to be equally weighted (i.e. analysts look at tweets without recognizing that many tweets are initiated by a small number of producers, some of which may be bots.) Three areas are investigated in this tight essay: geolocation inference, the problem of group accounts, and different types of tie (social vs. coworker vs. acquaintance). The authors note that analysts of Big Data often assume a behavioral model, but these authors show that the reality is different. For example, frequent tweeters are more likely to be seen as having high mobility, and ''therefore'' more likely to be classified as a professional rather than an acquaintance.

The problem of agency also figures in Lee and Martin's essay. They write that one of the great virtues of Big Data is that ''we can analyze multiple trends in the same total population, moving from one average man that poorly represents one big population to multiple average men that represent segments of the total population more accurately—not analyze a given population, but discover populations inductively.'' The focus then shifts away from studying variations from normal toward an appreciation of the complexities of uniqueness. Here, agency is about identifying the actual specificity of agentic life, not hollow average men. In mainstream social statistics ''we liquefied everyone into a homogenous soup despite how much they varied individually. But when you look carefully enough, we are all deviants. 'It's not weird to be weird—in fact, it's absolutely normal,' as Harrison White wrote.''

Theme #3: What is time?

Many of these essays show the ways that our understanding of and our ability to measure and theorize temporal processes has been changed with the move to Big Data. For example, in his article, ''Lost in a Random Forest: Using Big Data to Study Rare Events'', Christopher Bail's use of Big Data to study rare events brings up a variety of issues about time and temporality in social science. Bail writes, ''Even if scholars were more clairvoyant conventional methodologies cannot capture both the speed and scale at which cultural change occurs. Public opinion surveys, in-depth interviews, and ethnographic observation obscure the social processes that allow a civil society organization's messages to cascade across the potentially boundless social networks that comprise the public".

In his essay, "Structure from Interaction Events," Wouter De Nooy points out that this is not merely a question of obtaining more complete temporal measurement, it is also fundamentally a question about how the analyst conceptualizes time: "Big Data on social actors mainly record events, e.g. interactions between human beings that happen at a point in time. In contrast, social network analysts tend to think in terms of social relations that exist over a timespan. The challenge, then, is to rethink our conceptions and models of social relations and social structure. I conceptualize social structure and social relations as forces."

Bearman tells us about a number of ways in which Big Data affords us an opportunity to use the flow of temporality in new and enlightening ways. Describing the research of Jose Atria on the "Old Bailey" criminal court archival records, Bearman suggests that these kinds of data can help us understand: "How did the different processes involved in the emergence of modern society interact with one another? Which changes came first or at a more dramatic pace, and which followed either in timing of onset or speed of change?"

Theme #4: What is context?

Next there is a perhaps even more general problem of just what counts as context in any given analysis. Multiple understandings of context surface in these papers. Breiger's essay raises the critical issue of what and where is the "situation" for the Big Data model. In his essay Breiger describes a number of ways in which there are often big gaps between the social world and the forms of Big Data that purport to measure it. Breiger writes that it is a "question of the extent to which big-data research applies to human behavior at the human scale of church suppers and department politics in which we spend much of our lives."

Diesner explains how critical for the use of Big Data is the need for proper contextualization. Interpreting Big Data results requires "a deep understanding of content domains, their features, and trustworthy findings." Park and Macy talk about "geolocation inference" of social media devices or context as "network neighbor[hoods]".

For Wouter De Nooy context is critical. De Nooy seeks to construct a social science grounded in the lived social experiences of situations generated out of person to person interactions. He argues that this is where we find our best ground for theorizing and measuring social network processes. De Nooy writes, "it took me quite some time to realize that overall social network structure is merely a by-product of how social actors respond to their local network context". By this account, larger social structures are nothing more than the aggregated effects of local interactions between individuals who connect with one another in actual network locales dynamically across time. For De Nooy the great virtue of Big Data is that it provides "unprecedented access to contextualized and longitudinal action at the micro level that allow us to model relational expectations."

Kevin Lewis provides a telling example of the importance of context. Lewis points out that how friendship operates in the real world is the inverse of how it operates in the Internet world. In the former you have to work to sustain friendships. In the online world you have to work to terminate friendships. In his essay, Breiger suggests that remedying this gap between (for example) Internet friendship and real friendship would benefit by being recognized within the wider problem of "scaling down."

Theme #5: What is a cause?

Finally there is the fundamental question of causality itself. Certainly the shift from deductive toward inductive and abductive models of science has upended some of the conventions we had for causality (Martin, 2011). But also, the shift in scale to Big Data can change the nature of what causality actually means. Christopher Bail reports on a new style of social science modeling which builds upon the ability to measure the social world in entirely new ways. Bail describes his research which is based on an "app-based technology that collects hundreds of variables that describe the interaction between a very large group of civil society organizations and millions of social media users....Unlike conventional research methods, these data describe the spread of cultural messages across entire populations of people down to the millisecond."

Bail points out that when working at this scale there are likely to be multiple causal pathways toward an outcome. Bail writes: "(t)here are numerous reasons to expect causal complexity within the process of how organizations' social media messages go viral. To begin, there are many different types of organizations—each with a unique message. There are also many different types of potential audiences for their message. Finally, there are a variety of broader social conditions (e.g. current events or other 'opportunity structures') that may enhance the likelihood of virality. Within each of these broad categories there could be multiple indicators that combine—in different ways—to produce an outcome such as virality or cultural change more broadly." In other words, perhaps when working with Big Data, the very meaning of causality itself needs to be updated.

The essays

We have grouped the 18 essays into four topics. The first group, "Impacts of Big Data on Academic Fields," contains four papers that take up the question of how Big Data has impacted a specific sub-discipline. Peter Bearman describes the effects of Big Data on the field of historical sociology. He especially highlights the importance of new digital archives, "These include—and this is only an idiosyncratic sample drawn mainly from Britain,—the complete text record of the Old Bailey, the extant records of the Atlantic slave trade, and the British East India Company." And, "from the United States—State of the Union speeches, the Congressional Record, transcripts of Supreme Court decisions." As Bearman points out, these kinds of records are especially important for studying history precisely because they were produced by the very same institutions that made that history. According to Bearman, we are witnessing "the beginning of the archival revolution."

Paul DiMaggio writes about the impacts of Big Data on the field of text analysis. DiMaggio also sees Big Data as having a transformative impact on his field and specifically on healing the longstanding divide between the natural and the human sciences. Writes DiMaggio, "I can report that the era of the 'two cultures' (Snow, 1959) is over. Instead of epistemological chasms, I have found modest differences in orientation, of which I shall mention three, reflecting computer scientists' and social scientists' respective intellectual traditions." Sophie Mützel describes her experience as a sociologist encountering the tools of Big Data, an account in which she reflects upon her own trajectory through the world of Big Data analysis. While appreciating the power that new text mining tools have provided, she also insists on the importance of grounding these tools in sociological problems. She writes, "because of its insights and techniques to study meaning and how the social is structured, sociology makes itself very relevant to data science projects mining large data sets." Finally, Ted Underwood describes the transformation that Big Data has brought to the field of literary studies. Like DiMaggio, Underwood describes an important erasure of longstanding disciplinary divides. Underwood writes, "The boundary we used to take for granted between the humanities and quantitative social sciences no longer has a rationale rooted in the nature of our material."

Our second group of papers, "Cautionary Tales for the use of Big Data," includes six essays that draw attention to the distortions that Big Data sometimes smuggle into empirical studies of *the social*. Included here are Diesner's comments on the need to pre-process data, Adams and Brückner's studies of the politics of Wikipedia, Park and Macy's essay on problems with Twitter data, McFarland's research on sampling problems and the essays by Kevin Lewin and Ryan Shaw on the systematic biases of Big Data.

The third group, "Constructing a Computational Cultural Science," includes six papers that point to the new possibilities that the world of Big Data is creating for traditional projects in the social and human sciences. Included here is Goldberg's essay on forensic social science, Bail's research on the study of rare events, De Nooy's work on using Big Data to change the study of social networks, Hannigan's essay on the transformations in management science, Buurma's essay on literary studies and our own essay on constructing a computational hermeneutics.

The last topic we call "A Big Picture View of Big Data" and we have included two essays that are both notable for the scope and depth of their reflections on these matters. Monica Lee and John Levi Martin's essay "Surfeit and Surface" offers a broad view on the problem of Big Data and the study of the social. Their conclusion? Big Data is a remedy for much of what ails the social and human sciences.

The concluding essay, "Scaling Down," is by Ronald Breiger. In his essay Breiger describes a number of ways in which there are often big gaps between the social world and the forms of Big Data that purport to measure it. Breiger also questions the behavioral models (or the lack thereof) that are passed unnoticed in many applications of Big Data. Ultimately Breiger counsels us to follow a kind of abductive logic of his own design, encouraging us to seek "a duality of scaling up and down" that has the goal of making concrete and precise the articulations that link micro-interactional processes and the sorts of institutional, structural or field level processes that also shape the social world.

Conclusion

> What is a thing? The question is quite old. What remains ever new about it is merely that it must be asked again and again. (Heidegger, 1967)

The social sciences and the humanities look as though they are on their way toward significant transformations as they become ever more fully engaged with the computational sciences in the inevitable pursuit of the best use of Big Data. Perhaps, as Paul DiMaggio proposes in his essay, the era of the two cultures truly is coming to an end. What is certainly true is that as these two worlds of science meet, so too is there a clear clash of cultures and confusion over some of the most basic ontologies of scientific knowledge and practice—what is a thing, an agent, what is time, context, causality?

How does one best advance the frontiers of knowledge? As these essays have suggested, this is a liminal zone, one in which many of these fundamental assumptions are being engaged, contested, and negotiated.

These essays also suggest that, in the very best work, there is a balance between the two styles of research. When it comes to studying the social, the most productive research is that which most effectively leverages the insights of both the social and the computational sciences. In part this is because of the higher level of sophistication that leads to a higher quality research. We have seen, for example, in the various cautionary tales that our contributors have told, that there are dangers associated with the unguarded use of Big Data. They have illustrated again and again that one needs to ground this style of research in an effective sense of the phenomenology of the social. But we also see many other examples in these essays of what potentials might be unlocked when both disciplines leverage their best practices and theories toward a common goal of scientific discovery.

We are, of course, by no means the first scholars to have raised this complex weave of questions about Big Data (e.g. Anderson, 2008; Ignatow, 2015; Jockers, 2013; Kitchin, 2014; Lazer et al., 2009; Liu, 2013; Mayer-Schonberger and Cukier, 2013; Moretti, 2013; Ruppert, 2013). Indeed, we have found that Kitchin's essay (2014) provides a thoughtful and far more extensive intellectual history of most of the issues that we have covered here. In particular he highlights what has been one of the more important themes in our essay as well, and that is the relationship between induction, deduction and abduction. Kitchin recounts how Chris Anderson's (2008) provocative article ‘‘The End of Theory’’ expressed much the same sentiment of some of these essays, that traditional hypothesis testing was a less useful scientific model than a more purely inductive exploration of the complexity of Big Data. But, as Kitchin also points out, in the end, Anderson's vision is an unsupportable fantasy of a pure and unmediated induction of the sort that never happens. As an alternative, Kitchin describes the features of a research program that would seem to suit many of the projects described here, a program that he calls ‘‘Data Driven Science.’’ This is characterized by its efforts ‘‘to hold to the tenets of the scientific method, but is more open to using a hybrid combination of abductive, inductive and deductive approaches to advance the understanding of a phenomenon’’ (2014: 5). Kitchin writes, ‘‘the process is guided in the sense that existing theory is used to direct the process of knowledge discovery, rather than simply hoping to identify all relationships within a dataset and assuming that they are meaningful in some way’’ (2014: 6).

It is then, as Kitchin notes, not an end to theory, but rather a creative redeployment of theory and the mechanisms of surprise. In their discussion of abduction and science, Tavory and Timmermans (2014) write, ‘‘Theorizing is not a separate form of research. Theory is part of the research act that emerges as we build and problematize the generalizations produced by others and offer generalizations of our own.’’ ‘‘Abduction is this speculative process of fitting unexpected or unusual findings into an interpretive framework’’. The 18 essays here theorize their methods, their discoveries, and their surprises with their own abductive energy and processes.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. We employ the term ‘‘ontology’’ in its broad, traditional sense as referring to (often implicit) assumptions about the nature of being or, in our case, the nature of the social. Our usage is to be contrasted with the more formal meaning the term has in computer science and information science, although there are a number of interesting connections between the two usages of the term (Mallery et al., 1986). Timothy Hannigan discusses some of these linkages more fully in his essay.
2. We mean both a formal duality (e.g. Breiger, 1974) and more abstractly a duality as a kind of articulation system that links different modes and levels of social experience and analysis (Breiger, 2000; Breiger and Melamed, 2014; Lazega et al., 2013; Mohr and White, 2008; Mohr and Neeley, 2009).
3. DiMaggio et al. (2013) demonstrate new relevance for Big Data of Bakhtin's ([1975] 1981) work on heteroglossia, ‘‘the capacity of a text to contain multiple voices and thus to speak in different ways to different audiences.’’ Mützel picks up on a similar point in her essay in this special issue. How do researchers ‘‘hear’’ these different voices and assess their uptakes by different audiences?
4. Most recently, Lee and Martin (2015) present what they courageously assert to be ‘‘formal techniques that do not involve imposition of interpretation *before* the analysis’’ (p. 1, original emphasis). These authors construct networks of terms where the ties between terms are counts of the number of paragraphs in which both terms appear (p. 29). Lee and Martin's assertion that their approach does not impose interpretation prior to analysis appears to us to ignore their own stipulations that (a) the English translations they use are ‘‘likely to achieve more reliable results’’ than would the use of the German-language

original texts, that (b) the authors use (unspecified) "a priori" reasoning to consider when the singular vs. plural forms of words have different usages, and that (c) they insure that the terms they include in their network are of "recognized philosophical import," and they apply their (unstated) procedures for doing so prior to analysis (pp. 29–30). In all these respects of small decisions with big impact, we see Diesner's essay as relevant to Lee and Martin (2015).

5. Big Data practice with regard to tests of statistical significance, as well as our depiction of it here, are misleading in that "conventional" tests of significance can be applied only to data representative of a population, and cannot be applied to non-representative "found" data. Of course the significance tests can be run even if the assumptions are not met, and routinely they are.
6. Shaw deconstructs things as simple but as consequential as "autocompletion", viz. "Of course tools are not always used as intended, but an understanding of how a system was intended to be used, and how those intentions were hypostatized in the system's design, can help one interpret users' conduct: to what extent are they working 'with' or 'against' the system?"
7. Of course, as Sophie Mützel reminds us in her essay, "topic modeling only shifts substantive interpretation to a later position in the analytical process—it does not replace it." See also Mohr and Bogdanov (2013).
8. DiMaggio points out that although unsupervised models are proving useful, they are also notoriously difficult to assess or validate, "...unlike supervised models, for which straightforward means of validation using held-out samples are available, assessing the quality of solutions from unsupervised models is more challenging, with criteria that are more varied and less definitive." Beyond this, the field these days now includes a wide variety of variations of topic modeling technologies including many variants that make different trade-offs with regard to using supervised vs. unsupervised procedures.
9. Ted Underwood (in his essay for this special issue) describes it this way: "The modeling methods that prevailed for most of that century were best suited to structured datasets with relatively few variables, and that isn't the form our subject usually takes. Sociologists could use linear regression to model social mobility, but it wasn't clear how we could use that method on unstructured text."

References

Anderson C (2008) The end of theory: The data deluge makes the scientific method obsolete. *Wired*, 23 June.

Bakhtin MM ([1975] 1981) *The Dialogic Imagination: Four Essays*, trans. Emerson C and Holquist M. Austin: University of Texas Press.

Blei DM, Ng AY and Jordan MI (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research* 3: 993–1022.

Blei DM (2011) *Introduction to Probabilistic Topic Models*. Princeton, NJ: Princeton University.

Blei DM (2012a) Topic modeling and digital humanities. *Journal of Digital Humanities* 2(1): 8–11.

Blei DM (2012b) Probabilistic topic models. *Communications of the ACM* 55(4): 77–84.

Breiger RL (1974) The duality of persons and groups. *Social Forces* 53: 181–190.

Breiger RL (2000) A tool kit for practice theory. *Poetics* 27(2–3): 91–115.

Breiger RL and Melamed D (2014) The duality of organizations and their attributes: Turning regression modeling 'inside out'. *Research in the Sociology of Organizations* 40: 263–275.

Breiman L (2001) Statistical modeling: The two cultures. *Statistical Science* 16(3): 199–231.

Burke K (1941) *The Philosophy of Literary Form*. Berkeley, CA: University of California Press.

Burke K (1945) *A Grammar of Motives*. Berkeley, CA: University of California Press.

DiMaggio P, Nag M and Blei D (2013) Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. Government Arts funding. *Poetics* 41(6): 570–606.

Goldstone A and Underwood T (2014) The quiet transformations of literary studies: What thirteen thousand scholars could tell us. *New Literary History* 45(3): 359–384.

Heidegger M (1967) *What is a Thing?* Chicago, IL: Henry Regnery Company.

Ignatow G (2015) Theoretical foundations for digital text analysis. *Journal for the Theory of Social Behaviour* 1468–5914. Available at: http://dx.doi.org/10.1111/jtsb.12086

Jockers ML (2013) *Macroanalysis: Digital Methods and Literary History*. Urbana, IL: University of Illinois Press.

Kitchin R (2014) Big Data, new epistemologies and paradigm shifts. *Big Data & Society*. DOI: 10.1177/2053951714528481

Lazega E, Jourda M-T and Mounier L (2013) Catching up with big fish in the big pond? Multi-level network analysis through linked design. *Social Networks* 30(2): 159–176.

Lazer D, Pentland A, Adamic L, et al. (2009) Computational social science. *Science* 323(5915): 721–723.

Lee M and Martin JL (2015) Coding, counting and cultural cartography. *American Journal of Cultural Sociology* 3(1): 1–33.

Liu A (2013) The meaning of the digital humanities. *PMLA* 128: 409–423.

Mallery JC, Hurwitz R and Duffy G (1986) Hermeneutics: From textual explication to computer understanding? MIT Artificial Intelligence Laboratory Memo No. 871, Cambridge, MA.

Martin JL (2011) *The Explanation of Social Action*. New York, NY: Oxford University Press.

Mayer-Schonberger V and Cukier K (2013) *Big Data: A Revolution that Will Change How we Live, Work and Think*. London: John Murray.

Mohr JW and Bogdanov P (2013) Introduction—Topic models: What they are and why they matter. *Poetics* 41(6): 545–569.

Mohr JW and White HC (2008) How to model an institution. *Theory and Society* 37: 485–512.

Mohr JW and Neeley B (2009) Modeling Foucault: Dualities of power in institutional fields. *Research in the Sociology of Organizations* 27: 203–256.

Moretti F (2013) *Distant Reading*. London: Verso.

Ruef M (1999) Social ontology and the dynamics of organizational forms: Creating market actors in the healthcare field, 1966–1994. *Social Forces* 77(4): 1403–1432.

Ruppert E (2013) Rethinking empirical social sciences. *Dialogues in Human Geography* 3(3): 268–273.

Snow CP (1959) *The Two Cultures*. Cambridge: Cambridge University Press.

Tavory I and Timmermans S (2014) *Abductive Analysis: Theorizing Qualitative Research*. Chicago, IL: University of Chicago Press.

Timmermans S and Tavory I (2014) Theory construction in qualitative research: From grounded theory to abductive analysis. *Sociological Theory* 30(3): 167–186.

Wagner-Pacifici R (2010) Theorizing the restlessness of events. *American Journal of Sociology* 115: 1351–1386.

This article is part of a special theme on *Colloquium: Assumptions of Sociality*. To see a full list of all articles in this special theme, please click here: http://bds.sagepub.com/content/colloquium-assumptions-sociality.