Policy & Internet, Vol. 7, No. 4, 2015

Causation, Correlation, and Big Data in Social Science Research

Josh Cowls and Ralph Schroeder

The emergence of big data offers not only a potential boon for social scientific inquiry, but also raises distinct epistemological issues for this new area of research. Drawing on interviews conducted with researchers at the forefront of big data research, we offer insight into questions of causal versus correlational research, the use of inductive methods, and the utility of theory in the big data age. While our interviewees acknowledge challenges posed by the emergence of big data approaches, they reassert the importance of fundamental tenets of social science research such as establishing causality and drawing on existing theory. They also discussed more pragmatic issues, such as collaboration between researchers from different fields, and the utility of mixed methods. We conclude by putting the themes emerging from our interviews into the broader context of the role of data in social scientific inquiry, and draw lessons about the future role of big data in research.

KEY WORDS: big data, social science research, causation, correlation, theory, epistemology

Introduction

The emerging use of "big data" approaches—we define "big data" in the discussion section below—has led to much debate about the nature of social science research. Big data approaches, while still in their infancy, are already posing a number of methodological and epistemological challenges to common understandings of social scientific knowledge. Much of the debate has revolved around the question of whether, with big data, correlational research is supplanting causal research, with the implication that social science is losing its traditional explanatory focus and being led instead by research questions that are driven by readily available data. This article will probe this question in depth, drawing on interviews with researchers and putting these into the context of a broader discussion about the nature of causality and validity of social scientific knowledge in the era of big data. The article concludes with a discussion of how these issues are likely to impact social science knowledge, and a list of recommendations for scholars and institutions engaged in this promising but nascent research front. Others have already offered warnings about the validity

1944-2866 © 2015 Policy Studies Organization
Published by Wiley Periodicals, Inc., 350 Main Street, Malden, MA 02148, USA, and 9600 Garsington Road, Oxford, OX4 2DQ.

and limits of big data approaches (boyd & Crawford, 2012; Savage & Burrows 2007, 2009). In this article we go further by examining big data from the perspective of the practices of researchers in the field in order to gauge whether these initial concerns are applicable to all research in the same way and how researchers cope with the challenges involved.

This article focuses specifically on three interrelated challenges posed by the emergence of big data in the last several years. First, a fundamental question that has recently been brought back into focus is whether the aim of social science is to explain societal phenomena causally, or if it is enough to detect patterns without explicitly seeking explanations for them. With big data approaches, it has been argued that correlations are "fast and cheap" to calculate (Mayer-Schönberger & Cukier, 2013, 66) as compared with the more laborious process of establishing causal relations. Second, a debate that is closely related to this tension between causal and correlational approaches to social science research concerns the alleged "end of theory" wrought by the arrival of big data approaches. In a provocative 2008 article, *Wired Magazine* editor Chris Anderson argued that "Petabytes [of data] allow us to say, 'correlation is enough'" (Anderson, 2008). Anderson (2008) noted that the sheer abundance of data of use to social science research, and the rapidly improving tools for capturing, storing, and processing it, mean that researchers can now "let statistical algorithms find the patterns where science cannot," without prior theories or hypotheses about what might result. This development gives rise to the third challenge we identify here: The increased prominence for inductive over deductive reasoning, wherein data drives research questions, and simply mining data for patterns without the need for starting with questions may lead to new discoveries.

The consequences of these trends for social science and social scientists could be substantial. Yet with the exception of a few specific examples, these challenges remain speculative. Anderson's (2008) much cited "The End of Theory" essay provides mainly anecdotal illustrations. Nonetheless, in the several years since these questions were first posed, big data research has become much more prominent in the social sciences, spawning degree programs, a social science journal devoted to this field (*Big Data and Society*), and many conferences and articles (Golder & Macy, 2014). In light of these developments, it is worth asking whether the rumored decline of causation, theory, and the deductive approach is borne out, and what consequences this might have for social science.

This article assesses these challenges by presenting evidence from interviews conducted with academics and practitioners at the forefront of big data research. These interviewees were asked about their current research and future plans, as well as a host of questions about their methods, data sources, and findings. We begin with how researchers identify the challenges and disruptions that big data introduces to social science research, move on to how they respond to these challenges, and continue with how they regard the benefits of this type of research as it becomes incorporated into social science. We follow this presentation of interview data with wider reflections about the nature of big data research, and conclude with some general recommendations for how the challenges posed by big data might be met, as well as the potential it offers for social scientific inquiry.

Before embarking on the presentation of the interview data, it is worth outlining our conclusions: We will argue that the claims regarding the shift toward correlation and the "end of theory" are exaggerated. The truth in these claims, however, arises from the fact that, with big data, new sources of data about societal processes have recently become available in computational, readily manipulable form. These new sources have resulted in researchers rushing in to take advantage of them, often (though not always) without regard to considering first what the significance of the findings may be, or how these findings can be situated in existing theoretical ideas about the objects (e.g., social media) under investigation. Related to this is that many (again, not all) of the objects under investigation are newly available for research, and so they do not fall into existing traditions of research or theories (again, social media, sensor data, or prices scraped from the Web are examples). Hence a degree of uncertainty of where the research is leading can be expected in a new area of research that has not yet had time to consolidate. However, it is still possible to say why research in this area is advancing social science knowledge—primarily because of the new abundance of data sources—and moreover, that these advances are bound to push social science into new domains. This can only happen, however, once these advances have become more embedded and consolidated within theoretical frames and understandings of these new objects of research and of their significance.

To make this point, we shall provide a definition of "big data," and discuss why there has been a recent focus on objects of study that have an unprecedented amount of computationally manipulable data points. As we shall argue, the opportunity to exploit this data holds considerable promise for driving social science knowledge. At the same time, there will (again) be uncertainty about how these findings can be integrated within existing social science explanations, and thus raise issues, for example, about whether they are merely correlational or based on finding patterns in the data. Ultimately, these issues will be resolved as big data approaches become integrated within mainstream social science and the new objects of research are theorized among existing phenomena of social science interest.

The reason for making these points at the outset is that, as we shall see, this overall picture emerges only if we piece together the responses from our interviewees, who focus on various aspects related to this larger picture. These interviews highlight how the causation versus correlation issue is dealt with by researchers in a number of practical and immediate questions about their research, how they cope with these questions, and why they are also indicative of the uncertainty that we have pointed to. These questions relate to the emergence of new sources of data and the new objects being investigated that these data "belong to" ("belong to" not in the legal sense, but in the ontological sense that they are characteristics of those objects).

In short, as we shall see, questions of correlation versus causation arise in various ways because of new practicalities of dealing with new data sources, which bring with them new research practices. These new practices only become more clear once we take a step back later in our discussion to examine what we mean by the term "data," and hence what is new about the sources of data and ways of manipulating them. It can be mentioned before we do this that one reason that this larger picture does not emerge from our interviews is that so far there have been no definitions of data, or analyses of how new data sources affect social science knowledge, a point to which we return in our conclusion.

Background

Questions about the nature and aims of social science knowledge have generated a voluminous literature (e.g., Goertz & Mahoney, 2012; Rule, 1997). The existing literature also deals with the related question of how to design different types of social research and the role of different theoretical approaches in this process (Ragin, 1994). In this article we are specifically concerned with the question of causality, which has of course also been much debated in the social sciences. But here, it is also possible to distinguish between science-in-the-making, where there are rival views, and science-already-made, where there is consensus. For the consensus view, we take the account of causality from a standard statistical textbook (Agresti & Finlay, 1997); statistics being the primary method used in big data research. They state that a causal relationship exists between two variables when three criteria are met: first, when there is an association between the variables; second, when there is an appropriate time order; and third, when other variables have been eliminated (Agresti & Finlay, 1997, p. 357). Much could be said to elaborate on this definition, but a key point in relation to the third criterion is that the elimination of other variables can only take place under experimental conditions. As we shall see, experimental conditions also occasionally obtain in big data studies. However, there is a different way to establish causality on this third criterion, which is simply to see if the causal stimulus works. In this respect, for big data approaches which proceed on the basis that correlations were found between variables x and y, and the y's will be manipulated in accordance with x, there is no need to eliminate other variables: if the manipulation works often enough or to a sufficient extent, the purpose to which knowledge is to be put will have been achieved. This point leads to a distinction between applied research as against agreed-upon valid knowledge, and we shall come back to it in our discussion: at this stage it is merely worth mentioning that in terms of previous literature, what is agreed-upon valid knowledge is only established with time, rather than within the existing literature (again, we shall return to this point).

There are many examples of previous studies that have discussed big data, but few have raised the issue of causation versus correlation directly, or more than in passing. Golder and Macy (2014) provide an overview of social science studies using big data, reviewing the challenges and opportunities of new sources of digital data. Einav and Levin (2014) discuss these new data sources specifically for economics, raising challenges mainly in relation to access to proprietary data and methods for managing and analyzing these sources. Borgman (2015) offers a comprehensive analysis of data uses in research, though without defining data. Closest to our discussion is the article by Ruths and Pfeffer (2014), although the methodological and epistemological issues they focus on are concerned with how accurately big data sources represent human behavior and how well methods are matched to the data sources being analyzed. Mayer-Schönberger and Cukier (2013) also discuss causation versus correlation, but mainly in terms of the policy implications of this debate; we withhold our discussion of this until the final section of the article. However, it can be mentioned already that the challenges and regulatory issues associated with how big data research is being used continue to generate much discussion (Lane, Stodden, Bender, & Nissenbaum, 2014; Pasquale, 2015).

Method

The interviews reported in this article ($n = 26$) were conducted as part of a larger project investigating the emergence of big data research in the social sciences,[1] which has so far held several workshops, conducted more than 125 interviews, and produced a number of studies of various issues surrounding this new type of research. Interviews were conducted with practitioners at the forefront of big data research. Using a semistructured interview approach, we posed both technical questions (in regard to data sources, tools, methods, and so on) and pursued more philosophical, reflexive issues explicitly—the responses to many of which are abstracted in the analysis that follows. Interview participants were sampled purposively rather than systematically: Given the nascent nature of big data approaches, interviewing pioneers, or "early adopters" within this exploratory fashion best served the aims of the project. Nonetheless, in scoping prospective participants, we were careful to sample a diverse set of perspectives.

All interviewees touched on issues related to the validity of big data research, but the interview data we present here is from those 26 interviewees who most directly addressed causation and correlation from among the wider set. Among the interviewees cited, all held positions in academic departments except one, Bernardo Huberman, who works in a commercial research lab (another interviewee, Webber, has spent most of his career in industry, but was a visiting researcher at a university at the time of interview). However, there is a great deal of variation among these interviewees, in terms of disciplinary affiliation by institution: Five were in media and communications, five in computer science, four were in business schools, two in sociology, four in information science, and one each in social psychology, geography, economics, political science, Internet studies, and physics. It is worth mentioning that the current disciplinary identification of our interviewees was in many cases not the same discipline that they had been trained in. Also, many of our interviewees identified with several disciplines (or new ones without established disciplinary labels, such as computational social science). It can be added that all the respondents consented to be quoted and named.

Findings

Identifying the Challenges of Big Data for Social Science Research

Traditionally, social science researchers often had to expend considerable effort in gathering data. Now, with more readily available sources of data in large-scale, digital form (e.g., from social media), they can use computational tools to exploit these. Has this shifted the focus of social science research to correlational exploratory research, which can be conducted quickly and cheaply, at the expense of causal, explanatory research? Sara Degli Esposti, a researcher at the Open University Business School, thinks it has:

> Big Data is all about correlation; it's not about causation, which means that you don't need to have a theory beforehand. You just start looking for correlation (…) so you don't have any idea about the structure of the data, you just find a funny correlation.

A concrete example of such an approach is that taken by James Pennebaker of the Department of Psychology, University of Texas, who discovered that the use of different types of words in American university application essays is associated with how they fare in further education. As Pennebaker notes, this was achieved without predefining any particular set of words:

> Once you have that big data set you can stand back and just look and see how a group of ten or twenty small, insignificant words such as *I*, *the*, *have*, *to*, and *and*, can have this big effect. And then when we went and started reading the essays that had a lot of those words versus don't we realised that, wow, they were, in fact, really different.

It is noteworthy that the association detected by Pennebaker does not imply causality in a simplistic sense; after all, he examined the progress of students who had been admitted to university, whether because of or despite their application statement. The statements therefore did not necessarily determine or cause their later success or failure. But the data set did highlight an important correlational measure in a manner that could not be established with less data or less computational power.

Jillian Wallis, a social science researcher at UCLA's Center for Embedded Network Sensing, discussed how the different characteristics of data across different fields impacts upon the ability to explain phenomena:

> And then you look at [the] research and you're like, "Yeah, but there's all this stuff going on that explains why they're doing these things, and you're not actually explaining any of that… It may be totally different in other sciences, like places where you have a low number of variables … [but in the] social sciences, we all have ridiculously huge levels of

> complexity, and so that's where the Big Data stuff doesn't work so well for us, I don't think, and, more or less, you lose the explanatory power.

Big data research places new and different demands on the researcher, including the knowledge of how to capture, store, and manipulate digital data; seeking correlations in massive databases, for example, requires technical expertise on the part of the researcher. Such skills are not typically part of traditional social science syllabi. As such, the first phase of big data social science research has been driven by those most able to conduct it: typically computer scientists and a small number of social scientists who already had or have quickly gained the requisite technical skills.

Yet disciplinary differences can also lead to uncertainty about how this research has been and should be conducted. Those who have the right set of skills to capture and manipulate data, such as computer scientists, may not be so well versed in the theories, concepts, and practices that have traditionally characterized social science research. There are epistemological differences in opinion here between computer scientists and social scientists. Ron Deibert is the Director of the Citizen Lab at the University of Toronto, overseeing an interdisciplinary team researching global security and human rights using computational methods. He describes some of these differences in his team as follows:

> So my level of understanding drops off at a certain point because I'm not a trained technical person, and that's frustrating as a director of the organization ... On their part I think the technical-minded people have a certain... it's hard to describe actually. Putting it not very generously there's almost a know-it-all attitude that people who are trained in the social sciences don't have, because I think they're more accustomed to 'There are many sides to an argument' whereas people who come out of engineering it's like 'There's a right way and there's a wrong way.' ... I'm definitely partial to the more relativist 'Okay let's look at this from all sides.'

Deibert is a social scientist, and it may be that social scientists are more cautious about producing definitive findings whereas computer scientists are less so, though this may also reflect that computer scientists are less familiar with continually contested theories than social scientists. Another example here comes from Bernardo Huberman, who echoes Deibert's views. Huberman was originally a physicist and is director of the Information Dynamics Laboratory at HP Labs. He has written extensively about big data (Huberman, 2012):

> The problem is that people are just measuring almost anecdotal data, namely, 'I saw this and I measure.' Yes, the statistics is perfect, your methodology is great... your T value, your Q value, whatever it is, is fine, they know all about that. The question is why should I care? ... The question that I'm asking is what are the important questions here? ... [many] of the top researchers ... come from computer science, they come from the physical sciences, and they haven't marinated enough in the social sciences to really know what is important to ask.

Huberman is thus an example of someone with a computer science background who realizes that computer scientists and other natural scientists may not have the questions "important" to social scientists from the outset that social scientists themselves may have, even if they can produce robust findings. A recurring theme among our interviewees (echoing the debates sparked by Anderson [2008] and Mayer-Schönberger and Cukier [2013]) was that merely digging into data without beforehand having a sense of the significance of what can be expected to be found is unlikely to advance knowledge.

A related issue is the incentive structure in different disciplines. Kevin Lewis, Assistant Professor at the Department of Sociology at the University of California, San Diego, described these motivations affecting research choices as follows:

> The fact is a lot of the people who are best equipped to work with these data—in other words the programmers—just aren't asking interesting questions with it. And you read these articles in conference proceedings or other analyses that folks have conducted and it's a pity because these people have the tools to do what I want to do but they're just really not doing sociologically wealthy stuff with it, which I guess is understandable when you don't have to in their field.

Lewis's sentiments here inform our discussion around the role of theory in the era of big data research: His characterization of interesting and "sociologically wealthy stuff" alludes to the value of investigating underlying sociological concepts and theories through the use of these new, potentially revealing sources of data. What we see here are different expectations of what research leads to: Novel results, versus findings that can be examined in terms of what they yield for improving upon existing social science theories or understandings.

By way of contrast, consider an influential article from Kell and Oliver (2004, p. 99), which discusses how the promise of "computational methods of data analysis" can complement research in "fields [which] are data-rich but hypothesis-poor." They restrict their discussion to this issue as it relates to the natural sciences, but it is useful to consider the differing impact of big data approaches on sociology, a field that is far less "hypothesis-poor." Notably, the computer scientists we interviewed—such as Robert West, a PhD student in computer science at Stanford University—echoed Lewis in acknowledging the tension between theory and data in sociological research, but from the perspective of his own discipline:

> I wasn't so influenced by sociological theories, because I just don't have a big background in that. So it's more post hoc, that people tend to fill in, I think, the sociological things. It's not truly interdisciplinary yet, I don't think. People still mostly come from, like the data angle, and we have this data set, we want to make interesting findings. And then maybe afterwards, you try to fit it in with sociology.

West articulates an inductive perspective here, although it is clear that this is informal, haphazard, and not part of a broader epistemological rejection of the hypothetico-deductive approach, but rather simply the result of a system of training or background, and incentives encouraging novelty over depth. This is not to suggest, however, that computer scientists are the only researchers with this mindset. Rich Ling, a social scientist and professor at Nanyang Technological University in Singapore, described the new relationship between findings and data in similar terms:

> In a lot of respects I almost feel like we're, sort of, digging around in the data and finding interesting things and then pursuing those. It's almost like we're having a focus group with the data, you know, where you can, sort of, root around and, 'Here's an interesting line of thought. Let's follow that one for a while.' And after you've found it you attach it to some sort of a theoretical perspective instead of the more traditional social science thing where you, you know, set out the theory and do the literature review and test the hypothesis and all that sort of stuff.

Here Ling, a social scientist, recognizes the "traditional" approach of his field but acknowledges that his own research using big data initially involves an inversion of this process, even if he recognizes that ultimately the findings need to contribute to—or be embedded—in theories. Ling's perspective thus demonstrates that the new "inductive" approach, which Lewis characterized as belonging to computer science, has permeated the direct experience of a social scientist. Yet it hardly needs stating that just because an existing paradigm like theory-driven research is "traditional" does not make it inherently more valuable than newer approaches. Indeed Richard Webber, a Visiting Professor of Geography at King's College London (though with a background in private sector marketing research), reflected on the benefits of a big data approach:

> I've long thought that the process of hypothesis generation has been assumed by many academics as a process that just happens, whereas I do believe that if you have access to big data, that is a wonderful way of bringing to light relationships between behaviors which you might otherwise not have seen any particular reason to exist.

Just such an inductive process has been adopted by Lyle Ungar and Andrew Schwartz, computational linguistics researchers based at the University of Pennsylvania. Ungar and Schwartz's team analyze the use of language in postings on social networks like Facebook to yield psychological insights. They use language clustering techniques to see which words tend to cluster together:

> Schwartz: And so that allows us to examine the language that's being used at a topical level, but still keeps it what we call 'open vocabulary,' which means we don't restrict our analyses to a pre-chosen set of words; 'data driven' as it's called more generally—we let the data tell the story

> of what is significant. So we're not just restricted to a priori hypotheses of what might correlate.

Freed from the constraint of having hypotheses at the outset, Ungar and Schwartz can uncover associations that may not have been anticipated or theorized. Their work exemplifies the research approach that has been elucidated in this section, which has benefits for researchers (as in this last example) but which would also be seen as problematic by social scientists and those (like Huberman) who agree with its goals. As this section has shown, the trend toward correlational, atheoretical research at the expense of the causal, hypothetico-deductive study, is undergirded by the incentives involved in research and by disciplinary backgrounds that lack traditional social science training. Such disciplinary differences are thus contributing to the adoption of a variety of approaches; including approaches questioned by social scientists, or at least that they question in terms of what is made of findings. In the following section we turn to ways in which researchers respond to these challenges.

Responding to the Challenge of Big Data

While our interviewees were certainly aware of the challenges outlined in the previous section, many of them rejected the idea that these represent insurmountable obstacles to social scientific inquiry. Many researchers we spoke to flatly rejected Chris Anderson's (2008) "end of theory" idea, for example David Jensen, associate professor at the School of Computer Science, University of Massachusetts, and Jillian Wallis:

> Jensen: I have a very strong opinion on the end of theory. ... I just laugh out loud when people say it's the end of theory because this is such a basic misunderstanding of how knowledge is developed and how you need to approach observations in science or anything else.
>
> Wallis: None of the people that I work with have any idea of abandoning theory. And I think everybody I work with just scoffs at that. Unfortunately, this is something that has been more embraced by the public at large, those who don't understand how science actually works and the need for theory.

As Wallis points out, Anderson's "end of theory" argument has remained well-known, even if it has found little acceptance among social scientists. This is ironic given that even Anderson himself acknowledged, soon after the article was published, that he had engaged in overstatement, and backtracked from claims that the scientific method was being rendered obsolete (Gladstone, 2008). Thus by its very bluntness, Anderson's article has become something of a straw man: The argument can be rejected outright without much engagement. Yet the previous section has presented real evidence of a shift toward atheoretical, correlational approaches—even if these were not conceptualized in the dramatic terms Anderson used. There is thus a need to go beyond the false dichotomy represented by the "end of theory" and its absolute rejection. In this section we explore how our interviewees engaged with this issue in a more nuanced manner, taking seriously the challenges outlined before.

Interviewees also restated the value of causal explanation as a fundamental aim and motivation for social science research. This was articulated by David Jensen:

> And this, of course, is a central concern of social science: we don't just want to find statistical associations, we actually want to uncover the underlying causal processes by which social systems work (. . .) The data themselves don't tell you about cause and effect, there's actually a often very complex inferential process you have to go through in order to extract from the data the things that you really care about.

This point begins to hint at the difficulties we shall discuss later (they have already been mentioned); namely, that the data in big data research "belongs to" objects of research that are not yet well-theorized, so that while interesting patterns are beginning to be uncovered, in some cases even causal patterns, it may be some time before the complexities that Jensen mentions can be overcome to arrive at the significance of findings about these objects (the "things that you really care about"). Correlational research also poses further challenges, such as that the findings may not be revealing. Mike Cafarella, a professor of computer science at the University of Michigan, pointed out that there is often a danger of spurious correlations in big data sets with billions of relationships, potentially affecting the validity of these correlations:

> if you look at the data long enough you'll find predictive signals that are in fact completely spurious (. . .) for a multiyear period, the US stock market was highly correlated with the level of butter production in Bangladesh . . . if you look at hundreds and hundreds of these indicators, whether it's the level of Bangladesh butter production or the number of cars in New York City or whatever it is, eventually you'll find something that just by pure chance matches what you're looking for.

This danger was echoed by Patrick McSharry, the Head of Catastrophic Risk Financing at the University of Oxford's Smith School, whose research focuses on probabilistic forecasting for policy making:

> if you look enough, if you've got something that you're interested in predicting and you've got lots of possible variables and you just look at correlations, if you look at 100 time series you will find five of them that are significant at the 5% level, by definition, but that doesn't mean they're really going to be significant tomorrow or in the next year.

As we see again, the focus here is on the resulting danger of spurious correlation arising from such large amounts of data. Yet even when a correlational finding is valid and consistent over time, there is a broader issue of applicability of a finding to academic research, in terms of its relevance and substantive significance, as Miguel Centeno, professor of sociology at Princeton University, argued:

> you can find out eight billion things that don't matter (. . .) there's so many of those that you can just get drowned in it, I mean, it becomes a little bit like the Borges story of the map that's as large as the territory that it's mapping so it becomes unreadable. (. . .) I think we still need that human being; you still need that imagination to be asking (. . .) a good question.

Thus a good question is as important as a correct answer in determining the value of a finding to the pursuit of greater understanding. Obviously, this is nothing new for researchers, but with the use of big data, the "distance" or distinction between questions and answers is often brought into greater relief: Big data researchers now deal with data sets which are not only massive but also have unprecedented heterogeneity—for example, a corpus of tweets—of which myriad questions, spanning a multitude of disciplines, could be asked of the data. Compared with most "small data" studies, the choice of research question when using big data is often less self-evident in relation to the abundance of data available.

Far from answers "presenting themselves," then, in actuality big data research requires sophisticated understanding of the topic on the part of the researcher, so that sensible research questions can be generated. How might this be emphasized in practice? For other respondents, the supposed rise of correlational–atheoretical approaches could be countered in disciplinary terms, particularly in regard to computer science-driven research:

> Sara Degli Esposti: If you put a computer scientist in front of a database, what kind of answers will you get? You will get a huge number of spurious correlations.

A social science background, which emphasizes the role of the human researcher, can provide greater reflexivity, providing greater value to big data research. This point was made by a number of researchers from various disciplines, including Nate Hilger, an assistant professor in economics at Brown University, Duncan Simester, a professor at the Sloan School of Management at MIT, Sean Goggins, assistant professor at the University of Missouri's Informatics Institute, as well as Bernardo Huberman:

> Hilger: I think it's not like if you get your hands on big data, you're all set. It still takes immense creativity and like persistence and statistical

> sophistication. It's the same as before, it's not like if you get big data, you can do good social science research.

> Simester: we're paid for creativity and insight and that's really what distinguishes successful academics from others, and having this good data certainly makes our life easier, but it's, sort of, necessary and not sufficient... you've got to work out what to do with it.

> Goggins: the way that you're going to aggregate the data should be motivated by theory, method, and research questions.

> Huberman: What we need, that will make it all better, is to ask the right questions. We need people with a social science background that have good ideas of what to ask.

These points, again, foreshadow the point that we will make in the conclusion, that what is required is not only the theoretical insights of individuals, but also that the social science community with big data as an emerging research front needs to coalesce around theoretical approaches or around the significance of findings. Brent Hecht, who has degrees in both geography and computer science, and whose research frequently involves collaboration with researchers from other disciplines, is assistant professor at the University of Minnesota. He offered additional insight that hints at this point ("reinventing the wheel"), and why successful big data analysis requires awareness and incorporation of theory:

> I see a lot of people in computer science reinventing the wheel over and over and over again on some very basic geographic knowledge. For example, the number of articles that have found that people who are close together interact more than people who are further apart which has been known at various levels ... You can probably go back and say it has been known since the 1700's ... And to see them getting the credit for [discovering something 'new'] is frustrating to me as a geographer while, at the same time, I'm perhaps guilty of that in other disciplines as a computer scientist.

Put the other way around, it is essential to incorporate theory in research design, according to Darren Gergle, an associate professor at the School of Communication at Northwestern University:

> I actually think that theory, whether it's coming from the linguistics perspective or the geography perspective, gives us insight into large-scale data that we can begin to do things that we weren't thinking about before (...) It gives us insight into where to look and how to look and how to compare. And it also highlights what to watch out for

> when you're working with these large datasets where you don't necessarily know where the data are coming from. (...) It's [a case] of having the theoretical understanding of the space and knowing what a good question would be or using that to drive the particular analysis a lot more. That's the preferred approach that I've taken.

As Gergle intimated, theories provide a check on the validity of data and an understanding of the context in which a research question is asked and answered. Josh Introne, an assistant professor at Michigan State University and a computer scientist studying new media, took this reasoning further, suggesting how theories help counteract the inherent biases of the researcher, which can arise from the sheer scale of a big data set:

> It is very hard to totally remove your bias from analysis, so I think that theories are really important because they will help you put on new lenses, you know, new ways of looking at things. We could have all the data in the world and all the computational horsepower we want, and in order to simplify that data and make it understandable you have to reduce the dimensionality of what you're looking at.

A different way to make this point is that a big data research approach must be made commensurate with what other researchers examining similar theories (or "dimensions") in an unbiased way are doing. This overview of the ways in which academics at the forefront of big data research dispute claims about the diminishing utility of theory and the abandonment of causality allows us to take the next step: as noted, Anderson's "end of theory" notion is easily rejected out of hand, but the responses included here have explained exactly why theory and causal research continue to be important. Some researchers warned of the dangers inherent in correlational research, while others emphasized the role of theory in shaping valuable research, putting findings in context and preserving objectivity. In what follows, we will seek to synthesize how the challenges outlined and responses offered also suggest ways in which the benefits of big data can best be incorporated.

Incorporating the Benefits of Big Data

In the interview responses presented so far, big data research has been discussed primarily in a pragmatic light. Veering away from hype, interviewees have discussed in practical terms how big data does and does not contribute to the advance in social science knowledge. Certainly, this is in part an outcome of our sampling frame—our participants are actively engaged in big data research so we should not be surprised about this practical focus—but it is also an indication that researchers at the frontline of research are increasingly able to conceptualize the potentials and pitfalls of big data more from experience than speculatively. However, this pragmatism is offset by lingering divisions and tensions, particularly around disciplinarity. Many of our participants alluded to traditional academic boundaries or differences in approach—for example, social science versus computer science—and the somewhat stereotypical characterizations that follow from these. Yet as we have already noted, the new demands that big data places on researchers includes skills that are not generally part of existing social science training and teaching.

As such, when we asked our participants for their ideas about how the emergent tensions and oppositions—between correlational and causal research, and inductive and deductive paradigms—could be reconciled, many of their responses related to reconciling divisions and tensions between the different disciplines that are required to work together at the leading edge of big data research. Above all a number of interviewees recommended greater collaboration and communication between computer scientists and social scientists in order to bridge disciplinary divides. Scott Hale is a data scientist at the Oxford Internet Institute, University of Oxford. He has a background in computer science, and works closely with a political scientist and a physicist:

> the best research will often emerge in collaboration between computer scientists who will have access to the tools and the background to further develop and apply those, and with social scientists who will have good pressing social questions that we can get insight into with the data that is now available.

David Jensen echoes this view:

> This is actually one of the areas I think where computer science and social science have a lot to inform each other about. … So this is an interesting area where I think the computer science and the social science come together and produce something that's actually very new.

Suzy Moat, a computational social scientist, is assistant professor at Warwick Business School. She says that divisions between disciplines impede more constructive engagement:

> I really feel myself sitting in the middle between this natural sciences or computer science approach and the social science approach. I just think the only thing that really matters is that people keep an open mind and understand that just because—and I mean from both sides of this—that just because people are doing it differently doesn't mean that they're doing it wrong.

Similarly, Darren Gergle argues that it is necessary to seek various sources of expertise:

> I'm trained a lot in the area of psycho-linguistics (…) and so I can kind of apply some of that and at least have enough knowledge to know what I

> don't know and where to go to other people for the right context from a theoretical perspective.

The responses in the previous section have already emphasized that theoretical grounding—or "marinating," to use Bernardo Huberman's phrase—in the social sciences was as important to good big data research as having the technical skills required to analyze data. These responses take this argument further, stressing that social scientists and computer scientists should collaborate on an equal and reciprocal basis, even if they may initially bring different skills and questions or concerns to the table.

Apart from the pragmatic benefits of collaboration and communication between disciplines, many interviewees also argued that it is necessary to reconcile the seemingly contradictory strands of correlational, exploratory research on the one side and causal, explanatory studies on the other on a more conceptual level. Thus a number of respondents recommended the adoption of mixed methods approaches, combining big data analysis with other research approaches, to maximize the validity and credibility of findings. Frauke Zeller, assistant professor at Ryerson University, outlined this strategy using an example from her own work with online communities.

> At the end of the day, you always have to try to do some mixed methods approaches ... That's what I did in my Habilitation research [part of a German doctorate degree]—I developed a methodological design that allows you to start off with quantitative analysis of big data sets, for instance, discourse analysis is qualitative, and using that to double-check and reaffirm the result from the quantitative data. And if you find some new aspects, then you can go back. So it's an iterative process that you can do.

Similarly, Bente Kalsnes, a PhD student at the University of Oslo, advocated such an approach in the context of her cross-national research into the impact of social media on election campaigns:

> For me, I think if I only look at the numbers I don't get the whole picture... If we look at, for example, Twitter data, you can see some tendencies, but if you want to answer the right question then I think it's necessary to do more qualitative studies ... So I'm doing interviews with political parties, I'm also doing interviews with journalists, in order to talk about how they are using social media as journalistic tools.

From a methodological point of view, the term "mixed methods" or integrating qualitative approaches seems apt. It seems equally apt to say that these ways forward are also ways of integrating theoretical concerns or questions about the significance of the data sources into the research. A different way of making this point is that correlation can also be used as a starting point to investigate deeper causes, as Richard Webber notes:

> So you start off with the patterns and then what you should be doing is saying 'Well, here's some possible explanations,' and then when you've found some relationships which really deserve more detailed investigation then you would undertake a more detailed qualitative assessment as to whether this explanation was valid or not.

However, in many cases it is not so simple, and the relation between different methods must be more dynamically iterative. Alex Leavitt, a PhD student at the Annenberg School, University of Southern California, has experience with both computational and ethnographic methods, and this informs the design of his research:

> It's just this constant iteration back and forth between the data and the research questions and there are areas where in a sense they actually kind of look like traditional ethnographic methods where you're going out into the field with some kind of like vague research question in mind, finding data and comparing people, coming back to the research question, shifting a little bit here and there depending on what you find, going back out into the field, doing the same thing over and over and it constantly iterates and reshapes over time.

Scott Hale also sees practical benefits of combining quantitative and qualitative methods:

> Having that good question ultimately often is an iterative process of past experience with large-scale data or experience with some of the qualitative work that leads to a question and you go, oh, right, there's something we couldn't answer with that method, but can we use the large-scale data and ask the question there.

What we can see among our interviewees is a growing awareness of the need to allow findings to be informed by questions on an iterative basis. We also asked our interviewees about the future direction of big data research more generally, and they pointed to how the current state of the big data research front is likely to change over time, as methodological sophistication and conceptual understanding advances, more effective interaction between different methods emerges, and big data approaches become incorporated into mainstream social science research. Rich Ling suggests that this would return theory to a more central role:

> We're relatively early in the access to these types of data with the tools that we can use to examine them and so we don't have that many cases of it and it's, sort of, an exploratory phase. Eventually, there are, sort of,

> perhaps ill-thought-out hypotheses that people are using, but still people are following their hunches and that will coalesce as time goes on into more theory-driven types of analysis.

Jonathan Zhu, a professor of media and communication and founding director of the Web Mining Lab at the City University of Hong Kong, echoes Ling when he suggests big data research is still in a preliminary phase:

> Right now people think that we are coming back to exploratory research to try to identify patterns, try to identify, to discover things we didn't know, or we have no hypothesis to guide us (...) But that doesn't mean you can do this forever. I would guess that after a year or two, usually, people become tired of looking aimlessly for new things from Twitter (...) What will happen then? People will come back and say, okay, let's try to make sense out of it, let's try to see what we know from theory.

In short, big data research is coming to be seen as complementing more traditional forms of social science inquiry. Collaboration and communication between different disciplines and methodologies, as well as more flexible and iterative approaches, are likely to bridge the divides registered in the earlier sections of this article. Moreover, the interview responses emphasize the temporary nature of these divides: These may naturally recede over time as exploratory findings are exhausted or at least depleted and theories are leaned upon more heavily.

Discussion and Conclusions

This article has drawn on interview data from 26 researchers to investigate how the emergence of a new, disruptive area of research has provoked fresh debate over traditional epistemological issues, some of which—such as causation—stretch back hundreds or thousands of years. At this point, we can put the themes that have emerged from our interviews into a broader perspective. First, there are many questions that continue to be debated in the philosophy of social science concerning the difference between causal and correlational analyses and the role of theory. Interestingly, our interviewees did not focus on detailed technical aspects of these debates, but rather discussed broader issues such as the scientific nature of the social sciences and the role of theory on the one hand, and on the other the ways of coping with the issues raised in the various practicalities of their research—such as combining methods, ways of fostering interdisciplinary collaboration, and the like. Our interviewees also commented on the sources of the data, but primarily reflected on the sources they used or were familiar with. They further discussed data sources in practical terms of access to commercial data or difficulties cleaning the data, and also the bigger picture that there are now new and more readily available sources of data.

The focus of relating questions about big data to their own research is to be expected: It is difficult for researchers to step outside the confines of their own research and reflect on larger shifts in the research landscape of the social sciences. But taken together, the evidence gleaned from our interviews also offers broader insights into the current state and future direction of the big data research front as it pertains to questions of theory and causation. This bigger picture, of how the sources of data contribute in more and less powerful ways to the current frontiers of knowledge, we shall argue, is essential.

This issue of the powerfulness of knowledge, and the question of where big data fits into the landscape of current social science research as a whole, requires a definition of big data. Here, "big data" can be defined as a knowledge advance that represents a step change in the scale and scope of knowledge about a given phenomenon (Schroeder, 2014, 2015). Note that this definition does not rely on "size" per se, but on size in relation to the object being investigated, and how research advances beyond (which is one way of thinking about "powerfulness") previous research about this type of object. In other words, we can think about whether big data derived from, say Twitter (assuming we have the whole Twitter data set, see Kwak, Lee, Park, & Moon, 2010) compares with data that we have from letter writing or television watching: This allows us to see that a step change has taken place inasmuch as we have much more data about the object being studied that allows us to advance our understanding of—in this case—various kinds of media.

This has implications for how advance in social science can be gauged, and presumes a realist and pragmatist epistemology (Hacking, 1981) because the definition requires that there is an object out there—realism—about which more useful or powerful knowledge has been gained—pragmatism (the following discussion is based on Schroeder [2015], where there is also a more detailed definition of "data"). Hacking (1981) also defines scientific knowledge in terms of "representing and intervening" in the world. Further, as mentioned earlier, the "big data" discussed here "belongs to" (in the ontological rather than legal sense) the object of study, and exists prior to analysis: As Hacking (1981, p. 48) puts it, the view that "all data are of their nature interpreted" is misleading: "data are made, but as a good first approximation, the making and taking come before interpreting." He adds, "it is true that we reject or discard putative data because they do not fit an interpretation, but that does not prove that all data are interpreted" (Hacking, 1992, p. 48). He also distinguishes data from other parts of the scientific process, such as the calibration of instruments. The research pursued by our interviewees fits these definitions of data and big data since it often consists of social media data (or other digital data) that pertains to objects (the social media platforms or other technologies used) that have more interactions or individual pieces of evidence about objects (such as text) such that they constitute a step change beyond existing research about these or similar phenomena.

Yet this definition also highlights a limitation that is overlooked by those who promote the revolutionary impact of big data on the social sciences (Giles, 2012) and on society generally (Mayer-Schönberger & Cukier, 2013): If big data belongs to the objects of research (as per the "realist" definition of data, drawing on Hacking), then those objects (such as new media) which feature such data are also limited: There are only as many such social media objects as people who use them, and the same goes for other objects which have digital social traces. Once the usefulness of analyzing them is exhausted—hypothetically, if all possible social scientifically interesting relationships on Facebook or Twitter were to have been researched—then that would entail diminishing returns in the advance of social scientific knowledge.

In our review of previous literature, we gave an example of a consensus definition of causality (association between variables, time order, and elimination of alternative explanations), and mentioned the key point that an experimental elimination of alternative explanations is only sometimes possible. Yet if causal (or other) explanations work in advancing knowledge, they will be integrated into agreed-upon social science knowledge in due course. At this point we can see the difference between scientific as against applied knowledge in action: Scientific knowledge advances based on improving upon existing knowledge in terms of representing the world more powerfully but without intervening in it except within the confines of the condition for producing knowledge (such as laboratories, or testing if a particular factor or stimulus has a particular effect). Applied knowledge, on the other hand, puts knowledge to use to intervene in the world and to manipulate it (such as changing the behavior of populations) without regard to whether this advances existing scientific knowledge but only with regard to whether the searched-for phenomenon has been observed—or the intended effect achieved in practice.

This way of bringing causality down to the level of how it is operationalized in both scientific and nonscientific practice also points to the difficulty of mapping current efforts: On the one hand, there are many studies using big data approaches with powerful results which go beyond existing studies—simply because objects of study which have unprecedented amounts and varieties of data are available. At the same time, there are many warnings about the dangers of how big data research is becoming used in applied settings, as when they do not conform to scientific standards and hence also cannot be challenged on the scientific grounds of transparency or reproducibility. This applied research therefore calls for new regulatory approaches because these studies seek to predict or manipulate behaviors on the basis of more powerful techniques than existing ones. In this situation of science-in-the-making, what cannot be seen by individual observers (including the authors of this article) is the impact of a broad shift toward a more quantitative social science based on these new objects and of the shift toward more powerful quantitative manipulations based on these objects in the world-at-large: Both are ongoing and create new uncertainties, and the consequences for general knowledge advance (how the findings will become integrated in what is known about new digital media and other objects that make digital data available more generally) and how quantification affects society generally is not yet known. What is certain is that the limit of this knowledge is set by the objects on which they are based and the tools by which they can be exploited, even as these objects are still being added to and the tools are being improved.

All this can be put differently: It is possible for new knowledge to be generated where digital data objects are available that represent a part of the world (or a population) of interest. This limits the power of causal knowledge because in the case of applied knowledge, where the context of prediction or manipulation tends to be fleeting, whereas scientific knowledge aims to create knowledge that applies to populations of interest more generally and indefinitely (either by studying broad populations, as in much quantitative research, or deriving wider insights from the study of narrow populations, as in much qualitative research). But if applied knowledge does not need to meet the scientific standard of causality, within science uncertainty is created not so much about causation in a technical sense but about the role of causal knowledge in the social sciences more generally (and thus also uncertainty about the role of theory and induction). This uncertainty will only be resolved as this new frontier of knowledge becomes integrated in a new consensus, and only insofar as consensus characterizes the research front in social science, about causally (and theoretically) valid and more advanced knowledge.

This broader perspective allows us to see what is and what is not at stake in the "end of theory" and "causation versus correlation" debates: While these debates are about methods and validity, a more useful question might be: Which phenomena or objects yield data that allow researchers to achieve what kinds of results? While our interviewees focused on the role of theory and the continued need to ask questions, also iteratively, a different way to see the problem is simply that while the early stage of big data research has often been occupied with extracting patterns from the data that has arisen from (and belongs to) new data sources, there has not yet been sufficient time to theorize the new objects of investigation or to iterate questions about the significance of the findings.

Anderson (2008) says correlations can be found without theories: Perhaps, but such "chancing upon" interesting relations between causes and effects or correlations still entails putting these chanced-upon findings into the context of what we know, which inevitably consists of both a representation of the known world—theory, again, in a broad sense, in relation to findings to date—and the reality (the data and the objects to which they belong) which is being represented. On a practical level, it is always necessary to begin research, or to examine results in the course of research, with an expectation about what one might find in terms of causes (or otherwise) that improve upon or beyond what is known. The need to have expectations is sometimes disguised by the fact that one's expectations are based on previous research and the need to think about how one can advance beyond it. In any event, this involves thinking about the causes or factors involved, or about theory or hypotheses. Data cannot be examined without such thinking, by simply trawling through data, since that process, too, entails that one stops trawling—when one has found something interesting in the light of previous research, or when useful results are obtained in the light of the current state of knowledge and how it is possible to go beyond this state. (It can be added that the need for expectations about the usefulness of findings does not invalidate Hacking's [1992] idea that data is "found" independent of being shaped by these ways to frame it.) Moreover, no matter if the analysis is causal or correlational, in either case there will be an attempt to link cause and effect in a broad sense.

Anderson's (2008) argument about the end of theory is often lumped together with causation versus correlation. But again, the "end of theory" notion conflates the practicalities of how research is done (the fact that data is mined, without having a theory or hypothesis at the outset) with questions about whether research contributes to valid and cumulative social science knowledge. For the latter, it is inconceivable that research could contribute to social science knowledge unless it improved upon existing knowledge that had been led by or fed into certain questions or theories, and our interviewees in different ways articulate this position. Mining data may be good enough to find patterns and, by implication, correlations in the data may be good enough to show these patterns; but it is still necessary to think about how these fit into causal and theoretical explanations.

In view of this broader picture about big data, which relates to how the sources of data have changed in the course of time, there are indeed issues about validity: How do social media relate to the populations that use them? Is the population of Twitter users, for example, representative of the population-at-large? Do problems—such as multiple people using the same social media account, or individuals having multiple ones—get "washed out" by large numbers? Does it matter if we infer from someone receiving a message or passing on a message that they have read its contents? These and similar problems, which are frequently found in recent big data research, have repercussions for establishing causation and building theory in the sense of the kinds of claims we make—in these cases problems in relation to the analysis of social media. But these problems relate to the sources of the data rather than to the thinking (theory) or the models (causal or otherwise) that go into analyzing them.

Mayer-Schönberger and Cukier (2013, p. 12) think that big data marks a departure from the standard model of deductive research, which they consider "an artefact of a period of information scarcity." Big data, they argue, lends itself to correlational findings rather than causal explanations: As they put it, cheap and fast correlational research means that "big data turbocharges non-causal analyses," such that "causality . . . is being knocked off its pedestal as the primary fountain of meaning" (Mayer-Schönberger & Cukier, 2013, p. 67). Further, they say that experiments used for "demonstrating causality" require more effort (Mayer-Schönberger & Cukier, 2013, p. 66) and they associate correlational analysis with statistics, which they contrast with analyses of causality, which are typically undertaken in carefully controlled experimental conditions.[2]

To be sure, in the context of the business uses of big data (the main focus of Mayer-Schönberger and Cukier's [2013] book), seeking correlations, for example, in order to predict purchasing behavior or influence in social media has become commonplace. For business and other practical uses, this may be sufficient, though it would not necessarily be sufficient to advance social science knowledge.

Perhaps the boundaries between commercial and academic uses of big data are becoming blurred in practice (boyd & Crawford, 2012; Savage & Burrows, 2007, 2009) inasmuch as social science researchers may use commercial sources of data, but this is a separate issue and does not necessarily have implications for the validity of social science analysis, unless the commercial nature of the data affects validity (or replicability in allowing other researchers access to the same data source).

The uses of big data by commercial as against social science researchers brings us back to disciplinarity: Big data research is inescapably tied to the kinds of data that are available. These come from sources like social media, sensors, and business-related sources like purchasing records. If we leave the commercial nature of these data sources to one side for the moment, it is obvious that computer scientists and social scientists have seized the opportunity to analyze new sources like social media. As we have seen, while computer scientists may tackle these sources because of the novel computational and technical challenges they pose, social scientists are more likely to begin research against the background of having certain theories in mind. But this is partly a matter of training and background rather than being intrinsic to how disciplines typically approach problems: Some computer scientists, for example, have investigated social theories, for example the attempt by Backstrom, Boldi, Rosa, Ugander, and Vigna (2012) to replicate Milgram's "six degrees of separation" experiment; others do not. Again, however, even theory-less research is bound to become embedded in knowledge that is informed by theory.

The idea that correlations should not be mistaken for causation comes from thinking whereby patterns in the data should not be mistaken as valid (causal) explanations. Yet there are standards of validity for both causal and correlational explanations. The question that should be asked instead is: How does big data social science research—research that relies on sources with many data points (more than available previously or elsewhere, which fits our definition of "big")—advance knowledge? The philosophical (or philosophy of science, or indeed, philosophy of social science) issue that correlation is not enough to establish causality needs to be put into the context of different types of research—in the applied world of business and government, and in the world of social science research—and what these are aiming at, and how they advance knowledge (or not) by means of taking advantage of the digital data that have become available from different sources, together with the computational techniques to analyze them.

To be sure, there are new business and policymaking opportunities in predicting behavior by finding correlations in data. These also give rise to the concerns articulated by Mayer-Schönberger and Cukier (2013) about how policies based on this type of knowledge may undermine autonomy. This is a valid concern: Social scientific knowledge (as with business knowledge) also sometimes attempts to predict human behavior. Yet there are distinct issues here: Only when policies are designed to preempt autonomy do the concerns expressed by Mayer-Schönberger and Cukier (2013) arise. These concerns apply if, say, insurance policies are denied on the basis of certain driving or health behaviors, or if someone is wrongly put under surveillance by government because the person falls into the category established by certain correlations. Yet this is an issue about how knowledge is applied: The concern does not apply to scientific knowledge per se, and it will not pertain to social science big data knowledge, unless it is used in the service of implementing policies (or changing behaviors).

Apart from predictive knowledge that is applied for practical—commercial or policy-related—reasons, there is the perspective of human subjects of the research: People do not like to be thought of as engaging in behavior that is subject to causal or correlational patterns (as evidenced by the outcry over the recent Facebook experiment; Kramer, Guillory, & Hancock, 2014; Schroeder 2015). At the same time, human beings of course also rely in everyday life on people and things behaving in causal and correlational—or predictable—ways. This point also relates to causation versus correlation. With regard to big data, it is argued that "mere" correlation does not entail causation in the sense that correlation is thought to be weaker than causation. Put differently, just because a correlation between two sets of data has been observed, this is not enough to say that a causal law or causal explanation can be derived from this. This complaint is partly a way to defend against determinism (if there is only correlation, behavior is not determined). However, this way of thinking is misleading: A business or government policymaker may not be looking for a causal explanation of behavior. Rather, it may be enough to be able to steer people toward certain desired forms of behavior: If customers buy this, then they can also be induced to buy that; or if the police can focus their efforts on where incidents are likely to happen, they can prevent more crime. To be sure, these are problematic approaches (though they can also be refined: If the correlation proves weak, a more sophisticated or powerful one can be attempted). Yet the problems relate to the application of knowledge, not to the validity of different types of knowledge.

Finally, causal and correlational analyses should not be mistaken for quantification, though big data is necessarily quantitative. Quantification is characteristic of big data or computational social science because the objects that are tackled are objects that lend themselves to statistical analysis. Big data analyses invariably provide statistical results, results that establish regular patterns in quantitative data. In order to advance social science knowledge, however, the regular patterns elicited from these numbers need to be turned into words (Collins, 1984), since that is the form that social scientific knowledge inevitably takes when it becomes part of scholarly communication about how these regularities or generalizations advance upon other regularities or generalizations in the existing literature.

Research in the social sciences (and elsewhere) has recently experienced a take-off in the availability of research objects that provide readily available data such as social media, and this can be contrasted with situations where it has been necessary to collect data in a resource-intensive way. Social science research is partly driven by the availability of—in this case digital—data sources. Of course there is no reason why this should necessarily lead to shifts toward more causal and correlational approaches to research. More generally, how much these or other sources are relied upon is a matter of how research is advancing and where there are gaps. In practice, however, there is a shift by social science into these new directions because of the practical availability of these data sources. Hence it is undoubtedly important to think again about the nature of explanation in the social sciences and how it is affected by this shift. However, this should be done in a holistic way—considering where data come from and how they are used—and not just by pointing to the issue of causation versus correlation or the end of theory. Instead, again, it needs to be established how big data analyses advance social scientific knowledge, how they fail to do so, and what is needed to overcome this shortcoming in future research. Building on the views of our interviewees about the continued need for theory in the practicalities of their research, we have argued that this shortcoming can be addressed by reflecting on how to embed the new data sources into larger social science theories, as well as by gauging the social significance of these new objects.

Josh Cowls, M.Sc., Oxford Internet Institute, University of Oxford, Oxford, UK [josh.cowls@oii.ox.ac.uk].
Ralph Schroeder, Ph.D., Oxford Internet Institute, University of Oxford, Oxford, UK [ralph.schroeder@oii.ox.ac.uk].

Notes

1. The project "Accessing and Using Big Data to Advance Social Science Knowledge," funded by the Sloan Foundation. http://www.oii.ox.ac.uk/research/projects/?id=98.
2. It can be mentioned in passing that it is not the case that big data do not allow experiments, as Mayer-Schönberger and Cukier (2013, p. 66) suggest. Counterexamples are studies of social effects on voting (Bond et al., 2012) and emotion (Kramer et al., 2014) on Facebook, both of which created a "natural experiment" among Facebook users. Thus "causation" in experimentally controlled conditions is not ruled out, pace Mayer-Schönberger and Cukier.

References

Agresti, A., and B. Finlay. 1997. *Statistical Methods for the Social Sciences*. Upper Saddle River, NJ: Prentice Hall.

Anderson, C. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *Wired Magazine* 16 (7). http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory/.

Backstrom, L., P. Boldi, M. Rosa, J. Ugander, and S. Vigna. 2012. "Four Degrees of Separation." In *Proceedings of the 3rd Annual ACM Web Science Conference (WebSci '12)*. New York, NY: ACM, 33–42.

Bond, R., C.J. Fariss, J.J. Jones, A.D.I. Kramer, C. Marlow, J.E. Settle, and J.H. Fowler. 2012. "A 61-Million-Person Experiment in Social Influence and Political Mobilization." *Nature* 489: 295–98.

Borgman, C. 2015. *Big Data, Little Data, No Data*. Cambridge, MA: The MIT Press.

boyd, D. and K. Crawford. 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological and Scholarly Phenomenon." *Information, Communication and Society* 15 (5): 662–79.

Collins, R. 1984. "Statistics Versus Words." *Sociological Theory* 2: 329–62.

Einav, L., and J. Levin. 2014. "Economics in the Age of Big Data." *Science* 346 (6210): 1243089.

Giles, J. 2012. "Making the Links: From E-mails to Social Networks, the Digital Traces Left Life in the Modern World are Transforming Social Science." *Nature* 488: 448–50.

Gladstone, B. 2008. "Search and Destroy" [interviewer; radio broadcast episode]. July 18. In *On the Media*. New York City: WNYC.

Goertz, G., and J. Mahoney. 2012. *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences*. Princeton, NJ: Princeton University Press.

Golder, S., and M. Macy. 2014. "Digital Footprints: Opportunities and Challenges for Online Social Research." *Annual Review of Sociology* 40 (6): 1–6.

Hacking, I. 1983. *Representing and Intervening*. Cambridge: Cambridge University Press.

Hacking, I. 1992. "The Self-Vindication of the Laboratory Sciences." In *Science as Practice and Culture*, ed. A. Pickering. Chicago: University of Chicago Press, 29–64.

Huberman, B.A. 2012. "Sociology of Science: Big Data Deserves a Bigger Audience." *Nature* 482: 308.

Kell, D.B., and S.G. Oliver. 2004. "Here Is the Evidence, Now What Is the Hypothesis? The Complementary Roles of Inductive and Hypothesis-Driven Science in the Post-Genomic Era." *Bioessays* 26 (1): 99–105.

Kramer, A., J. Guillory, and J. Hancock. 2014. "Experimental Evidence of Massive-Scale Emotional Contagion Through Social Networks." *Proceedings of the National Academy of Sciences* 111 (24): 8788–90.

Kwak, H., C. Lee, H. Park, and S. Moon. 2010. "What Is Twitter, a Social Network or a News Media?" In *Proceedings of the 19th International Conference on World Wide Web*. New York: ACM, 591–600.

Lane, J., V. Stodden, S. Bender, and H. Nissenbaum, eds. 2014. *Privacy, Big Data, and the Public Good*. Cambridge: Cambridge University Press.

Mayer-Schönberger, V., and K. Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London: John Murray.

Pasquale, F. 2005. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge MA: Harvard University Press.

Ragin, C. 1994. *Constructing Social Research*. Thousand Oaks: Pine Forge Press.

Rule, J. 1997. *Theory and Progress in Social Science*. Cambridge: Cambridge University Press.

Ruths, D., and J. Pfeffer. 2014. "Social Media for Large Studies of Behaviour." *Science* 346 (6213): 1063–64.

Savage, M., and R. Burrows. 2007. "The Coming Crisis of Empirical Sociology." *Sociology* 41 (5): 885–99.

Savage, M., and R. Burrows. 2009. "Some Further Reflections on the Coming Crisis of Empirical Sociology." *Sociology* 43 (4): 762–72.

Schroeder, R. 2014. "Big Data: Towards a More Scientific Social Science and Humanities?" In *Society and the Internet*, eds. M. Graham and W.H. Dutton. Oxford: Oxford University Press, 164–76.

Schroeder, R. 2015. "Big Data and the Brave New World of Social Media Research." *Big Data and Society* July-December: 1–11.