

Note: Snapshot PDF is the proof copy of corrections marked in EditGenie, the layout would be different from typeset PDF and EditGenie editing view.

Author Queries & Comments:

Q1 : Abstract and keywords have been imported from metadata. Please check and confirm.

Response: Resolved

Q2 : The reference "1988" is cited in the text but is not listed in the references list. Please either delete the in-text citation or provide full reference details following journal style

Response: Resolved

Q3 : The reference "Watts 2015" is cited in the text but is not listed in the references list. Please either delete the in-text citation or provide full reference details following journal style.

Response: Resolved

Q4 : The reference "Salganik 2018" is cited in the text but is not listed in the references list. Please either delete the in-text citation or provide full reference details following journal style.

Response: Resolved

Q5 : The disclosure statement has been inserted. Please correct if this is inaccurate.

Response: Resolved

Q6 : The CrossRef database (www.crossref.org/) has been used to validate the references. Mismatches between the original manuscript and CrossRef are tracked in red font. Please provide a revision if the change is incorrect. Do not comment on correct changes.

Response: Resolved

Q7 : Please provide missing page range for reference "Freeman, 2008" references list entry.

Response: This is an online journal, so no page range.

Q8 : The reference "Watts, 2017" is listed in the references list but is not cited in the text. Please either cite the reference or remove it from the references list.

Response: Resolved

CM1 : Add Orcid ID symbol here, and number?

CM2 :

Add in references:

Meyer, E.T. & Schroeder, R. (2015). *Knowledge machines: digital transformations of the sciences and humanities*. Cambridge MA: MIT Press.

Schroeder, R.(2014). Big Data and the brave new world of social media research, *Big Data and Society*, July-December: 1-11.

CM3 :

Replace Mau 2017 in references with:

Mau, S. (2019). *The metric society: on the quantification of the social*. Cambridge: Polity Press.

CM4 :

Add in references:

Schroeder, R. (2017). Big Data and Communication Research, *Oxford Research Encyclopedia of Communication*, <http://communication.oxfordre.com/>

Schroeder, R. 2018. *Social theory after the internet: media, technology and globalization*. London: UCL Press.

CM5 : Add ORCID number here: 0000-0002-4229-1585

Big data and cumulation in the social sciences

Recto running head : INFORMATION, COMMUNICATION & SOCIETY

Verso running head : R. SCHROEDER

Ralph Schroeder^{0,0}

Oxford Internet Institute, Oxford, UK

CONTACT Ralph Schroeder ralph.schroeder@oii.ox.ac.uk¹

History : received : 2018-12-03 accepted : 2019-03-07

Copyright Line : © 2019 Informa UK Limited, trading as Taylor & Francis Group

ABSTRACT

Research using big data has become popular in the social sciences, raising many new questions². This essay focuses on the question of cumulation, and why the kind of cumulation that is characteristic of social data science is more akin to cumulation in the natural sciences. The reasons for this include how research teams are organized and how they compete to exploit certain data sets to improve upon the work of other teams. There are other factors, however, that mitigate against cumulation, including the lack of access to certain datasets and a lack of building on existing findings in the social sciences. Some of these factors pertain to fundamental philosophical issues in social science explanation, including new ideas about the workings of causal explanation. Others relate to the collaboration or absence of collaboration between different disciplines and to the difference between more applied and more academic research. The essay reviews these factors and develops an account of cumulation anchored in a realist philosophy of science and in the practices and tasks of social science research. It concludes with a call for big data research to be more integrated with already ongoing cumulative findings in the social sciences while recognizing that there are several obstacles to such an [Q1] integration.

KEYWORDS

Big data; cumulation; social data science; causation; disciplinarity

Introduction

Big data has rekindled a number of fundamental debates in the social sciences. Much attention has focused on the ethical and social implications of big data. Far less attention, on the other hand, has been paid to how big data is transforming the nature of knowledge in the social sciences. Yet here too, big data requires reexamining some basic questions, such as the role of statistics and of causal explanation in the social sciences, how data sources underpin the validity of knowledge, or more broadly how scientific knowledge cumulates. This essay focuses on these issues and argues that the increased emphasis on quantification, new directions in thinking about causality, and the increased availability of readily manipulable data are pushing social science closer to the kind of cumulation that characterizes the natural sciences. Yet there are also a number of barriers to cumulative knowledge, including access to data, the lack of task certainty in social research, and how new findings are integrated with existing knowledge. The essay explores these barriers in order to give a sociological account of this new knowledge domain. While the ethical and social implications of big data currently dominate, the essay argues that over the longer term, it is likely that the forces that promote cumulation or hinder it will play a stronger role in shaping social scientific knowledge.

The essay will proceed as follows: first, it will elaborate why the scientific nature of the social sciences is being revisited in the light of big data approaches. Next, it will provide definitions of big data (and of data) and argue that the conditions for social scientific knowledge have recently changed with access to readily manipulable data. Data sources are only one part of this change, and another goes to the heart of how cumulation takes place in science, which is the use of mathematics and of research technologies. Hence, too, new directions in thinking about causal explanations are emerging and there is a renewed emphasis on quantitative and statistical methods. The combination of these two factors – data, and how they are analyzed – has led some areas in the social sciences to develop into a rapidly moving research frontier. There have been many novel findings which have also redrawn the relations between traditional social science disciplines and newer specialisms such as computational social science. The essay concludes with a discussion of this realignment among disciplines and in knowledge, and why there are also limits to how big data approaches are transforming social science.

Before tackling these topics, it will be useful to discuss labels, since big data is still a controversial term and there are a number of competing ones, including computational social science and social data science. For consistency's sake, this essay will use social data science throughout, but nothing hinges on this. 'Big data approaches in the social sciences' could be used instead if it were not so clumsy. Big data and data will nevertheless be defined shortly because one argument will be that knowledge is being

reshaped. More abundant data sources are an essential element of what has led to this reshaping. It can also be mentioned that a number of other topics have recently come into fashion, including machine learning, artificial intelligence, and the role of algorithms. Although these are not the focus of this essay, they all depend crucially on big data, and big data, it will be argued, is a new departure in the social sciences.

The new sources of data in the social sciences are primarily though not exclusively digital media, which can be taken here to include both information and communication technologies. This includes the web and search engines (information), but also social media, smartphones, and email (communication). The essay will limit itself to knowledge based on digital media, for reasons of space but also because they constitute the basis for the vast bulk of new social science insights. Another restriction in the scope of the essay is that it will be limited mainly to sociology, political science and communication or media studies, thus excluding economics, geography and psychology (Backhouse & Fontaine, 2010). This limitation is due to the fact that no single essay can cover all these areas, but also because, arguably, digital media are also more central to insights about social, political and media behavior than to behavior as studied by economics, geography and psychology.

Science and cumulation

Unlike other knowledge domains, scientific knowledge is set apart by criteria of validity, and the question of what constitutes science can be considered prior to all the others discussed here, also because the definition of data that will be presented shortly will be limited to data that can be considered scientific. Hacking has defined science as 'representing and intervening' and argued that there are six 'styles' of scientific knowledge (2009; see also Kwa, 2011). Statistical knowledge is one of these styles and, as we shall see, statistics invariably plays a central role in social data science knowledge. Yet social data science is not always only statistical; sometimes a key method is 'experimental' (another of Hacking's six 'styles') and other combinations of 'styles' are possible. Yet one reason that statistics are so prominent, as mentioned already, is because of the availability of an abundance of readily manipulable digital data.

Data, insofar as it is part of scientific knowledge, can be defined as having three characteristics: it is a property of the objects being examined and thus separate from the observer; obtaining data comes before interpreting it; and data provide the most divisible or atomizable useful unit of analysis (Schroeder, 2014; Meyer and Schroeder, 2015, p.118² removed for review). These characteristics fit with a realist, objectivist, and pragmatist account of science and of technology. 'Big' data can consequently be defined as data that marks a step change in scale and scope in relation to a given object or phenomenon. In other words, there has been a change in kind rather than an incremental change in the material (data) available about phenomena that are relevant to social scientific enquiry. New – digital – sources data have recently become available that were not available before. Even if, as will be discussed below, there are many questions about the validity or scientificity of these data, for (social) scientific knowledge, the central issue is the extent to which the data enable us to penetrate social reality more deeply and comprehensively.

This brings us to cumulation in (social) scientific knowledge, and a useful entry point here is Whitley (2000), who argued that the social sciences can be characterized as having low degree of task certainty and low mutual dependence. In disciplines like physics, he argued, it is the other way around: high task certainty and high mutual dependence on each other's work. This is an oversimplification of Whitley's ideas, but for our purposes his argument points to the fact that research in social science often does not build on previous work and neither, often, does it have clear goals. However, in social data science this might be different since, arguably, there are tasks which are enabled by new data sources that should be undertaken with a high degree of – if not certainty, then at least ease. Further, in social data science there is, at least sometimes, mutual dependence in the sense, for example, that predictions build on and try to improve upon each other, as with predicting election results (Jungherr, 2016). Another indication of mutual dependence is that datasets are sometimes made available for replication or re-use.

We shall come back to these points, but apart from Whitley's criteria, there are additional reasons why social data science may be cumulative. To appreciate this, we can step back for a moment to take a broader view of science and how it cumulates which has been put forward by Collins, who characterizes modern science as 'high-consensus rapid-discovery science' (1998, pp. 532–538). A crucial ingredient for the modern take-off of science, in his view, is mathematics, and it is worth quoting him on this point:

the distinctiveness of the network of mathematical practitioners is that they focus their attention on the pure, contentless form of human communicative operations: on the gestures of marking items as equivalent and of ordering them in series, and on the higher-order operations which reflexively investigate the combinations of such operations. (1998, [92]p. 873)

In other words, instead of focusing on the organization of science, as Whitley does, Collins focuses on its epistemological practices, and identifies how scientists in practice, by having a commonly agreed upon method – mathematical symbols – can make their way forward. Collins further argues that research technologies, by which he means 'the lineage of techniques for manipulating formal symbols representing classes of communicative operations' (1998, p. 874), underpin this kind of cumulation. Again, it is the practices of scientists, the shared tools, that make for advances. Obviously, this understanding of cumulation

chimes well with social data science, where mathematics (algorithms) and research technologies (computing tools) play a central role.

Thus scientific advance needs to be reconceptualized to include technology, as it does for Collins, hence his use of 'technoscience'. But on the organizational side, apart from the forces identified by Whitley, there is another factor which is how research technologies change the way that research is organized: social data science is often undertaken by teams rather than individuals. These teams, if they have access to the necessary tools and data, operate at the research frontier and compete with a few other groups (Gläser, 2006) to achieve the most advanced research results in a manner that is more like computer science or other scientific disciplines that are typically organized in labs or groups compared to traditional 'lone wolf' social science research. The technoscientific and organizational sides of research thus need to be brought together. Data science research that analyzes data about large populations typically requires technologies that are organized by teams, and they often also require collaboration with large-scale organizations, such as national statistics or surveys but also digital media companies.

Finally, cumulation can also be seen in a different light, from the point of view of the phenomena being investigated. And in this respect, social data science entails moving onto the new terrain of digital media, mastering this new terrain, and moving onto ever newer terrain. Now it may seem that this idea does not apply to social science as it does in the natural sciences – decoding the human genome, say, or detecting the Higgs Boson. Social science, in contrast, often returns to the same terrain, such as explaining religion. However, digital media provide a new terrain for mastery, and again, the data available from these phenomena are readily manipulable: The terrain is new insofar as an increasing amount of social life takes place in a digitally mediated way, so that there are new phenomena for social science research to investigate. This could be expressed differently by saying that there has been an expansion in the ontology of social life; or at least in the ontology of the types of data about social life. This additional material provides new terrain for social science and enables 'rapid discovery' in Collins' sense and allows researchers to build on and improve upon each others' work – though whether there is 'high-consensus' is a separate question to be discussed shortly. Yet there are obvious limits to what this knowledge will add to social science – apart from the limits to technical advances or advances in methods – because the extent to which digital media shed light on the social world, as well as the extent to which social life is digitally mediated, are also limited.

An idealized account of cumulation has been provided here which helps to explain how social data science is distinctive within the current social science landscape. Yet this account leads to a larger question which has already been touched on: how will this cumulation affect the social sciences? First, it may affect some areas or subfields more than others; the analysis of social networks being a prime example (see Rule, 1997, pp. 120–147; and Freeman, 2008, who also details how between- physics and biology have borrowed from the social sciences in network analysis). Second, does only this kind of computationally driven knowledge advance social science? Clearly not, since many of the most important insights at the research frontier of social science may not be quantitative or based on digital data. Some previous attempts to summarize the scientific advances of social research to date have done so without much recourse to quantitative or digital media-based research (including generalizations about the cumulativeness of social science itself (Collins & Sanderson, 2016, pp. 10–11)). Yet as long as the premise that social science should aim to be as scientific as possible is accepted, and the premise that, wherever possible, empirical material, including data, should be sought, then the idea that social data science plays an outsize role in social science as a cumulative pursuit follows.

Two brief examples of cumulation can be given: One example is analyzing poverty. Here two studies can be taken from a wide range as illustrations. The first is the analysis of wealth and poverty, where Blumenstock, Cadamuro, and On (2015) used mobile phone records in Rwanda. Using data about how people move around with their phones, they were able to develop a powerful model of wealth and poverty as well as characteristics such as motorcycle ownership and electricity use. Moreover, they could test their model against the 'ground truth' of a government survey – and performed well against this much more costly means (in terms of effort and expenditure) of surveying the population. But although the code and data in this study were made available for replication, obtaining mobile phone records obviously requires the good will of a commercial mobile phone operator. Jean et al. (2016) used publicly available datasets of satellite imagery instead, for five African countries. For Rwanda, they were able to improve upon Blumenstock et al.'s predictive model by analyzing the roofing material of housing. More studies using mobile phones and satellite images have followed these, and other examples and are bound to continue to do so.

A second example is blockbuster movie prediction. This is quite an important area for Hollywood movie studios due to the enormous marketing budgets. Asur and Huberman (2010) showed that films that are most talked about in advance on Twitter will perform best at the box office by using a Twitter dataset of movie mentions comprised of 2.89 million tweets from 1.89 million users referring to 24 films released over a period of three months. They were able to demonstrate a strong correlation between Twitter mentions of a film and its box office performance, though Twitter data, as mentioned, is difficult to replicate. Mestyán, Yasseri, and Kertész (2013) used Wikipedia data instead, both the edits to a movie entry and the views of Wikipedia movie entries. Again, they were able to improve upon Asur and Huberman's model by predicting the first-weekend box office revenue of a set of 312 films released in the United States. And Wikipedia data, unlike Twitter data, is freely available and so it can be built upon. Many other studies have subsequently tried to outdo these two studies, with varying degrees of success.

Statistics, causality, and prediction

Against this backdrop, we can return to the main argument: Statistical knowledge, as already discussed, is only one of several of Hacking's styles of scientific knowledge. But, as has been documented on a number of occasions (foremost by Gigerenzer et al., 1990), this type of knowledge has become increasingly widely used in society, and so, over the course of history, have the purposes to which this knowledge has been put. Thus, statistics has been shaped by society and also shaped it. Hacking has pointed out that, in the nineteenth century, statistics brought new objects – in social science, populations – into being (1992). At the same time, these new objects and the data that belong to them are not purely 'socially constructed'; they have an independent or in this sense 'objective' existence or are 'out there' in view of the realist epistemology proposed earlier. What also came into being, particularly in the nineteenth century, were new infrastructures for data collection – censuses, surveys, and the like. More recently, these technological infrastructures have changed inasmuch as digital media provide abundant data about new objects, again, mainly about the user behavior of the populations of these media. As Porter (2008) has argued, the drive for quantification in the social sciences has waxed and waned in tandem with different demands from society. Thus recently, apart from increasing uses of statistics by governments and marketing and the like (Mau, 2019³), the imperative to use statistical methods in social research has also been reinvigorated because of digital data.

The way that statistical knowledge is typically used in social science is to talk about dependent (effects) and independent (causes) variables. The two are linked by means of a 'significant' relationship. The characteristic that sets statistical knowledge apart from other types of knowledge is that this relationship can be summarized in a number (the p -value) or condensed into mathematical formulas or other forms of abstract notation such as visualizations. (Again, it is worth mentioning as an aside here that another name for mathematical or abstract formulae or rules in computing is 'algorithms'). At this point, it will be useful to give an example of how big data has called forth new directions in thinking about causality: Pearl (Pearl & McKenzie, 2018) has argued that visualizing causal pathways is more important than big data per se in advancing science. The details of his argument about big data can be left to one side here, but Pearl's ideas fit well with the ideas about the role of mathematics in science that have been discussed in the previous section: the new approach that he is arguing for consists of formalizing causality using visual (or abstract symbolic) notation. On Pearl's view, once causal pathways (or the absence thereof) have been visualized, systematic examinations of relationships become possible (including counterfactuals, a particular approach championed by Pearl). These pathways, in turn, allow quantities or valences to be assigned to them which indicate the strengths of the causal relationships subject to statistical analysis. Visualization can thus be seen a form of mathematization, in Pearl's case a visual notation (recall Collins' 'pure, contentless form of human communicative operations') that captures relationships between objects and how they can be quantified.

Pearl is a computer scientist and philosopher. There are many other attempts to establish when causal claims are epistemologically valid that have come from within the philosophy of social science, including about counterfactuals (for example, Morgan & Winship, 2015). Pearl offers a guide to implementing his ideas, but his and others' philosophical ideas about causality are rarely connected directly to the many varied ways in which statistical and causal analysis are actually carried out in practice. The distance between philosophy of (social) science and technoscientific practice has often been noted in the sociology of science and technology (Fuchs, 1992). So, despite new directions in thinking about causality and statistics being stimulated by new computational techniques and the availability of data, the implementation of these ideas is likely to take time to percolate into widespread adoption of practices. Put differently, there is no generally agreed upon or accepted idea of causality in social science, only different schools of thought or approaches. Finally, one feature of causality in social science research is that the causal claims that are put forward are often cautious and tentative. Yet this feature should be expected for science-in-the-making as opposed to science-already-made (Cole, 1992).

This brings us to the link between ideas about causality and cumulation: whether causal claims are accepted or not should depend, for advancing scientific knowledge, on whether they build on or improve upon other causal claims on one side, and on the other side depends on whether there is consensus within the discipline or field about this acceptance or otherwise. Causality can thus be seen as an overall aim, variously interpreted and striven for – even if there is no general consensus about how to adjudicate causal claims. Again, as we have seen, there have been attempts to summarize the scientific – causal – generalizations in the social sciences that have been arrived at to date (Collins & Sanderson, 2016 & Rule, 1997, as mentioned, are examples). If these generalizations can be built upon, it is possible to see advances as stepwise additions to knowledge in relation to (causally) explaining the relationships between certain objects. Yet there is another principle of scientific explanation; parsimoniousness: other things equal, the complexity of explanations should be minimized or the number of independent variables (causes) and dependent variables (or effects) reduced to the fewest number. 'Reductionism' in social science, however, has mainly become a term of abuse that indicates that it is unbecoming of the complexity of the social world. The same goes for 'positivism' or 'empiricism'. Yet if causal explanations are sought, this effort inescapably reduces the social world to a few law-like relationships, whether this premise is made explicit or not.

To summarize, new directions in using statistics and thinking about causality have recently reinvigorated social science. Even if

little by way of consensus has emerged in philosophical debates, there are also methods textbooks which summarize the state-of-the-art for big data statistical methods (Morgan & Winship, 2015) and techniques (Salganik, 2017) for social data science. Methods textbooks contain much of science-already-made (Cole, 1992), though with new techniques and data sources, there are bound to be areas constantly in need of updating for science-in-the-making. One further point can be added here, which is Collins' argument that statistics is not just a method but also a theory (Collins, 1984). The reason for this, Collins says, is that a probabilistic universe is taken as a given and the strength of relationships is assessed against this backdrop of a universe of possible relationships. He also points that, howsoever the relationships that are established are expressed in terms of statistics, they must in the end also be expressed in words. Put differently, statistical techniques do not lend validity on their own in social science but must be integrated into existing bodies of knowledge in ordinary (or at least non-mathematical) language. Finally, some would claim that only experiments provide a way of truly achieving causal knowledge, testing an intervention in one group often with a control group without an intervention (on this debate, see most recently Deaton & Cartwright, 2018). Yet experiments are only one way of representing and intervening in the world, as Hacking argues, and 'experiments' are only one of several styles of science.

Data sources, data biases

Against this background, we can turn to how new sources of data enable cumulation. The data in social data science (or indeed in social science statistical analysis generally) can be thought of as consisting of numbers assigned to units where the rows are the social phenomena and the columns their attributes or vice versa. As we have seen, these are related to each other as causes (independent variables) and effects (variables) and a number is assigned for the strength or significance that links the two. Again, the data that provide the evidence for these relationships are now abundant, but what type of objects are these data and the phenomena they belong to? Unlike the 'populations' that came to provide the main basis for social statistics especially during the nineteenth century, they are not always people and their characteristics – they could also be texts, locations, and the like. Nevertheless, they are still always 'populations' of things, and again, the vast bulk are derived from digital media.

Unlike the traditional populations of social statistics that were often 'national', a distinctive feature of data in social data science derived from digital media is that these data often do not coincide with nations but rather, for example, with languages (Wikipedia, VKontakte for Russians) or with certain media functions such as the number of people with whom messages are exchanged. Put differently, the unit of investigation of the analysis is often given in terms of the medium – for example, the whole of Twitter, or Facebook users – which is in some (but not all, for example, Wikipedia) cases commercial. This characteristic is relevant to cumulation, since the fact that digital media are confined to certain populations (of users) and/or time (how long they have been used) and how uses change (Salganik, 2017, pp. 33–35, calls this 'drift') and above all for what type of mediated social interactions they are used – which, again, is a growing but limited domain of social change – is a constraint on cumulation.

It is sometimes claimed (Bowker & Star, 2000) that data are inherently political or social, that they are 'constructed' for certain political or social purposes (this has already been mentioned in connection with how ♦♦ populations' came into being). Yet this is true at best in a restricted sense. Take, for example, the idea that statistics creates new data objects such as 'unemployment'. This seems to go against the argument made earlier that data belongs to the object and that taking data comes before making it. Yet an unemployed person is not purely 'constructed'; obviously they are also real objects in the social world that do not have regular paid employment (or people who are unemployed based on another definition). And unemployed people have existed beyond particular social contexts in many different periods and times; though this is not to say, of course, that unemployment statistics cannot be used in a politically motivated way (as with domestic labor, which is often overlooked).

From the point of view of cumulation, more important than this alleged context-dependency is that the questions that are examined (and variables identified) in social data science are often shaped by the data sources. Here we can think of Twitter hashtags, Facebook 'likes' or shared links, or keywords used in Google searches. Put differently, in seeking social data science explanations, the data are often treated as 'given' in the sense that what is available is taken as the point of departure. It may also be possible, as already mentioned, to create new (non-observational) data in certain domains, as for example with experiments (though again, these sometimes relate to 'given' features of digital media). And 'may' was used in the previous sentence because there are areas where experiments are impossible or difficult for practical and/or ethical reasons. In any event, this relation of data to what it represents is a key question for social data science.

These considerations allow us to pinpoint the main strengths and limitations of digital data: The strengths are that they provide datasets that do not need to be gathered but are 'found'. The weaknesses include that digital data are divorced from social contexts or social structures, for example, from what shapes the media behaviors of different groups (though it can be noted that this information may not be provided by surveys or other media sources either). Instead, the data reveal patterns of activity or of social interaction from digital media rather than answers provided in a survey or behaviors from an experiment: the regularities in these cases come from the data.

This given-ness of data is an obvious limitation. At the same time, as mentioned, social life is becoming ever more mediated by

digital technologies. Apart from this growing significance, the reliance on digital media data increasingly skews social science research towards certain phenomena for which data are readily available (or away from those where it is not available). This issue also applies to the natural sciences, though natural sciences also tend to direct attention to objects (or terrain, see above) that can be quantified. In the social world, in contrast, there is often a return to well-researched objects or terrain, but also a turn to new objects or terrain like the new behaviors with digital devices. Again, this is not new: We can think of other areas where research has been driven by available data, such as the reliance of political science on political manifestos or speeches or newspaper text – even before these were available in digital formats. However, again, digital data sources are now more readily available and in manipulable form and at larger scales.

To fully assess how digital data is transforming cumulation in the social sciences would require a systematic survey of the new data sources that have become available – plus the tools for analyzing them and which social phenomena they pertain to. Such an account is not possible for knowledge that is still in-the-making; some historical distance or hindsight would be required. It is also not necessary since it is well-known which digital media have been most thoroughly studied in academic social science (Twitter, Facebook, Wikipedia, search behavior, and others), though less is known about this from the private and public sector side. We also know that digital media data have led to a rapid growth in the number of publications that have yielded new insights and research directions, including some that have led to controversy (see, for example, the journal 'Big Data & Society'). There have also been some reviews of studies using digital data (for example, Golder & Macy, 2014), and apart from individual data sources, there are data repositories (or data infrastructures) where several different types of sources are aggregated (for example, <https://dataverse.org/>). Datasets used to be gathered for specific purposes; now some are taken as given. All this can be regarded as a growing new terrain. At the same time, social science asks questions about specific aspects of social life; and what is missing despite the abundance of data and the insights derived from it are theories about how digital media data sheds light on the social world, as well as an understanding of how the newly won insights extend existing social science knowledge.

Prospects

As already mentioned, questions concerning the ethical and social implications of social data science have been dealt with here only insofar as they relate to the role of knowledge in society. But changes in the role of knowledge have such implications, whereby it is thought that deterministic (or causal or predictive) knowledge constrains individual freedom because action is shaped by external forces. This allegedly undermines human beings as free agents who reflect on and shape their social world as well as being shaped by it. But there are two parts to this argument: the first is 'reflexivity' (Giddens, 1991, pp. 36–44); the idea that when, for example, people become aware of what is known about them, that very knowledge may change their behavior in relation to it, thus making findings invalid. This criticism only needs to be spelled out to see that it is baseless: in some instances, people may change their behavior in the light of becoming knowledgeable. But in this case social science can, of course, establish whether and to what extent they do so, and so also gauge how findings may be affected by this knowledgeability.

The second part of the argument is that people shape their social world. To be sure, the idea of free will or that autonomous decision-making is the source of our moral worth is the foundation for ethics, Kantian or otherwise (for example, Sen, 2009). But the idea that people make ethical choices on the basis of valuing their own and others' autonomy and free will is compatible with scientific explanations of behavior. Kant stressed that the realm of ethics and the realm of science are two separate worlds (Gellner, 1974, pp. 168–191). The paradox whereby a free observer allegedly cannot be subject to a deterministic universe has been convincingly deconstructed (for these points about 'observers', see Fuchs, 2001: esp. 20–29). And social scientific or causal knowledge is compatible with the everyday self-understandings people have of themselves as free and autonomous agents; again, this knowledge and these self-understandings are separable. It is true, however, that in a culture that prizes freedom and individuality, peoples' worldview is predisposed against determinism because they do not like to think that their actions are shaped by outside forces – technological, social, or otherwise.

Another shortcoming of social data science that is often highlighted is unrepresented people (Hargittai, 2015). But while not everyone leaves digital traces, this issue can be reframed in terms of whether enough is known about the difference between those who are represented by digital data and those who are left out. If there is sufficient knowledge about this disconnect and how it skews findings, then this issue can be overcome. The same point applies, incidentally, to concerns not about people's digital traces but about digital traces like texts which skew evidence to written records. And while this essay has focused on peoples' use of digital devices, there are also social data science explanations which do not rely peoples' digital traces as sources; for example, how cars give an indication of voting patterns (Geburu et al., 2017), or for the types of housing structure or luminosity gives an indication of wealth and poverty (Steele et al., 2017). Finally, non-digital data were of course vulnerable to similar criticisms as are now applied to digital data, such as how phenomena are not captured by non-digital data, as when political activities that produce no recorded texts or speeches are overlooked.

Another criticism that has already been mentioned is that some argue that rigorous causal explanations are only possible with experiments; subjecting one group to a treatment and having another without treatment or with a different treatment. Yet this is

too restrictive; among other problems, it would exclude knowledge where experiments are not possible in the case of digital media data for reasons of access, ethics, and the like. It is also an overly narrow conception of science since, again, 'experiment' is only one of several 'styles' of scientific knowledge. Again, this issue can be decomposed into two parts, the question of scientific validity, and whether experiments are the only basis on which to undertake interventions in the social world (but see Deaton & Cartwright, 2018). Here it can also be mentioned that it is not necessary to know causes in order to undertake social science policy interventions or to know how to act; one example is that I don't need to know what caused rain when the forecast is for rain in order to make it sensible to take an umbrella (Kleinberg, Ludwig, Mullainathan, & Obermeyer, 2015). Finally, there is the perennial criticism against merely quantitative approaches. Yet this criticism can be left to one side; there is no reason why quantitative knowledge should exclude knowledge based on qualitative data or non-quantitative techniques, which can of course also contribute to advancing social science.

For the prospects for cumulation, an important question is the relation between academic social science knowledge and applied (public and private sector) knowledge. There has always been an overlap in the Venn diagram between the two, and this also applies to social data science knowledge. For social data science, there is often an overlap where academic researchers work with digital media companies in order to share data, or where the insights of social data science knowledge can be directly related to marketing strategies, or where knowledge is applied to nudging and other ways of shaping behavior (Watts 2017^[Q3]). Yet in policymaking, there are as yet few direct applications of social science knowledge, apart from security-related policy implementations in areas such as policing and the military (Poel, Meyer, & Schroeder, 2018).

Applied social research is well-established, but it is also diffuse and the contributions to science limited since the bulk of this research is focused on practical problems such as getting people to pay more attention to advertising or to buy more things. This means not only that the scope is limited but also that the effects are hard to gauge in a way that can be replicable or generalized. Furthermore, applied knowledge often seeks predictive power for the sake of intervention even if prediction and intervention may not be closely coupled. Meanwhile marketing and public opinion research have become mundane forms of applied knowledge, even if there are also calls for greater regulation of this knowledge where it is seen, for example, to manipulate shoppers (Turow, 2017). Scientific knowledge is judged by validity, yet its diffuse effects are bound to enhance capabilities for good and ill. And as already mentioned, statistical knowledge has been tied to different social goals over the course of history.

A final consideration that affects the cumulation of social data science knowledge is disciplinarity. As mentioned, disciplines outside traditional social sciences such as computer science now increasingly make contributions to the field. This means that many different directions are being pursued that build less on each other's findings and methods if those in different disciplines are not aware of the state-of-the-art. At the same time, researchers from computer science or natural science disciplines may, like social scientists with a scientific bent, take for granted that the use of scientific method (or of causal or statistical or predictive knowledge) will of its own accord improve social scientific knowledge. Computer scientists also typically and implicitly take cumulation in their own field for granted; by applying scientific methods (more often improving technology) to discrete areas, there is steady improvement. Moreover, computer scientists and other natural scientists sometimes import models about the natural world, such as complexity theory or evolutionary theory, into social science analysis. What is overlooked is that these theories have far less traction in established social science knowledge. Among less scientifically inclined social scientists, on the other hand, there is little expectation of cumulation because mutual dependence and task certainty have not been strong features of social science.

Cumulation provides a means of charting the landscape of knowledge or research and its challenges and opportunities; it is not something that researchers are generally aware of except in relation to particular topics or techniques on the research frontier. However, it is also important to keep in mind that the point of cumulation in science, whatever the discipline, is the ideal of a deeper and more extensive – more useful – penetration of reality (in social science, of 'social' reality), of being able to represent and (in some cases) intervene more powerfully in social change. What is specific to cumulation in social data science is that certain tasks – for example, analyzing aspects of digital media – are taken as 'given'. Merely one indication of this powerfulness is that those with social data science skills can move easily between academia and the private and public (and third) sectors, often with salaries considerably higher than in academia. This is rather rare in the social sciences generally outside of some areas of economics and political science and will affect how much expertise is available within and outside of academia.

To summarize, social science research pertaining to digital media is now being driven by approaches that promote cumulation. Yet one limitation, as we have seen, is that these are typically based on interactions taking place on digital media that are quite distinctive; in other words, from which only certain generalizable insights can be drawn. Moreover, these insights need to be located in an understanding about where the role of digital media fits into overall social change, and what the limits of this role are. As for other barriers, these consist not so much of insufficient access to data, as when connections to companies or purchasing commercial data is needed. Nor is it that replication is not possible because there is a 'black box' in how the data were generated; for some types of data, this information is available. Nor is the main barrier the relation between what online data says about offline life; digital media data sometimes provide powerful insights into offline social phenomena. The main

problem for cumulation is rather, first, that there is as yet little by way of a thorough account of the role of digital media in social life, and second, that both for digital media and the offline phenomena they shed light on, there is weak mutual dependence and task certainty – and hence weak cumulation – apart from individual findings in different areas. Examples of the first problem include that there has been a rapid shift to mobile phones and to social media and search engine use where the role of digital media is still poorly understood. Just one example of the second problem is that prediction, such as for election results using Twitter, still falls short (Jungherr, 2016).

The ethical and social controversies surrounding the analysis of digital media are thus only one barrier to the cumulation of academic research. There have been a number of controversies (Salganik, 2018[Q4], pp. 281–354) that have made social scientists and others wary of using certain types of data and of manipulating people. These issues are now well known even if solving them is still on a distant horizon. Yet manipulation, which is the main concern arising from cumulation, is not so much a concern of social scientists since, unlike political and commercial actors, they are at one remove from using this knowledge even if they seek it. Public concerns may put a brake on this type of knowledge outside academia, but academic social science is rarely directly a handmaiden of applied social science. Social science is bound to continue to pursue cumulative and reliable knowledge while applied knowledge will pursue knowledge that improves sales or policy.

This essay has shown that certain elements of social data science – mathematization, technologies, data, new objects – make for cumulation and a rapidly moving research front. Yet apart from formalization and technology, there has so far been little by way of cumulation in terms of integrating substantive findings in existing bodies of knowledge (Schroeder, 2017; Schroeder, 2018, pp.131-39⁴ removed for review), even if extensive new territory is being rapidly discovered and from different directions. The advances being made could be measured, for example by counting how many publications are being published on the various digital media, yet this would only yield a partial understanding of cumulation. A complete understanding of cumulation could only be achieved via a summary of how thoroughly findings about digital media have penetrated social reality, which can only be captured by hindsight. It is possible, however, to summarize the factors contributing to and preventing cumulation, as follows:

Favoring cumulation:

- Greater task certainty and mutual dependence
- Agreed upon symbolic tools and re-purposable and extensible research technologies
- Abundant sources of readily manipulable data about mediated social relations
- Greater striving for scientificity and applicability

Constraining cumulation:

- Access to data mainly about mediated social relations and lack of access to and transparency about the nature of data, often due to the commercial nature of the datasets
- Lack of agreement about integration into theory and existing substantive findings
- Absence of agreement about what constitutes causal or statistical or predictive explanation
- Limited disciplinary organizational identity and harnessing knowledge to discrete and immediate applications

In view of how disciplines and research are organized, with many directions explored simultaneously, it is unlikely that there will be a tighter coupling or integration between social data science and existing social science knowledge. If there were, perhaps this would lead to a truly major transformation of social science. The kind of cumulation that is taking place in social data science is nevertheless taking social science in new directions and shifting it to concentrate on certain approaches and objects. This cumulation is shifting the science of society, but it is almost entirely the science of mediated society.

Notes on contributor

Ralph Schroeder is Professor in Social Science of the Internet at the Oxford Internet Institute. He is also the director of its MSC programme in Social Science of the Internet. Before coming to Oxford University, he was Professor in the School of Technology Management and Economics at Chalmers University in Gothenburg (Sweden). His publications include 'Social Theory after the Internet: Media, Technology and Globalization' (UCL Press, 2018) 'Knowledge Machines: Digital Transformations of the Sciences and Humanities' (MIT Press, 2015, co-authored with Eric T. Meyer), 'An Age of Limits: Social Theory for the Twenty-First Century' (Palgrave Macmillan, 2013), 'Being There Together: Social Interaction in Virtual Environments' (Oxford University Press, 2010) and 'Rethinking Science, Technology and Social Change' (Stanford University Press, 2007). His current research interests include digital media and right-wing populism, and the social implications of big data [email: ralph.schroeder@oii.ox.ac.uk].

Disclosure statement

No potential conflict of interest was reported by the author[Q5].⁵

References

- Asur, S., & Huberman, B.** (2010). [Q6]Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology – Volume 01* (pp. 492–499). Washington, DC: IEEE Computer Society.
- Backhouse, R., & Fontaine, P.** (Eds.). (2010). *The history of the social sciences since 1945*. Cambridge: Cambridge University Press.
- Blumenstock, J. E., Cadamuro, G., & On, R.** (2015). Predicting poverty and wealth from mobile phone Metadata. *Science*, 350(6264), 1073–1076.
- Bowker, G., & Star, S. L.** (2000). *Sorting things out: Classification and its consequences*. Cambridge: MIT Press.
- Cole, S.** (1992). *Making science: Between nature and society*. Cambridge, MA: Harvard University Press.
- Collins, R.** (1984). Statistics versus words. *Sociological Theory*, 2, 329–362.
- Collins, R.** (1998). *The sociology of Philosophies: A Global theory of Intellectual change*. Cambridge, MA: Belknap Press of Harvard University Press.
- Collins, R., & Sanderson, S.** (2016). *Conflict sociology: A sociological classic updated*. Abingdon: Routledge.
- Deaton, A., & Cartwright, N.** (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210, 2–21.
- Freeman, L.** (2008). Going the Wrong Way on a One-Way Street: Centrality in physics and biology. *Journal of Social Structure*, 9(2). [Q7]
- Fuchs, S.** (1992). *The Professional Quest for truth: A social theory of science and knowledge*. Albany: State University of New York Press.
- Fuchs, S.** (2001). *Against essentialism: A theory of culture and society*. Cambridge, MA: Harvard University Press.
- Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E., & Fei-Fei, L.** (2017). Using deep learning and Google Street view to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences*, 114(50), 13108–13113.
- Gellner, E.** (1974). *Legitimation of belief*. Cambridge: Cambridge University Press.
- Giddens, A.** (1991). *The consequences of modernity*. Cambridge: Polity Press.
- Gigerenzer, G., et al.** (1990). *The empire of chance: How probability changed science and everyday life*. Cambridge: Cambridge University Press.
- Gläser, J.** (2006). *Wissenschaftliche Produktionsgemeinschaften: Die Soziale Ordnung der Forschung*. Frankfurt: Campus.
- Golder, S., & Macy, M.** (2014). Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology*, 40, 129–152.
- Hacking, I.** (1992). Statistical language, statistical truth, and statistical reason. In **E. McMullin** (Ed.), *The social dimensions of science* (pp. 130–157). Notre Dame: University of Notre Dame Press.
- Hacking, I.** (2009). *Scientific reason*. Taipei: National Taiwan University Press.
- Hargittai, E.** (2015). Is bigger always better? Potential biases of big data derived from social network sites. *The ANNALS of the American Academy of Political and Social Science*, 659(1), 63–76.
- Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S.** (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301), 790–794.
- Jungherr, A.** (2016). Twitter use in election campaigns: A systematic literature review. *Journal of Information Technology & Politics*, 13(1), 72–91.
- Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z.** (2015). Prediction policy problems. *American Economic Review*, 105(5), 491–495.
- Kwa, C.** (2011). *Styles of knowing: A new history of science from Ancient times to the present*. Pittsburgh: University of Pittsburgh Press.
- Mau, S.** (2017). *Das metrische Wir: Über die Quantifizierung des Sozialen*. Frankfurt: Suhrkamp.
- Mestyán, M., Yasseri, T., & Kertész, J.** (2013). Early prediction of movie box office success based on Wikipedia activity big data. *PLOS ONE*, 8(8), e71226.

- Morgan, S., & Winship, C.** (2015). *Counterfactuals and causal inference*. Cambridge: Cambridge University Press.
- Pearl, J., & McKenzie, D.** (2018). *The book of why: The new science of cause and effect*. London: Allen Lane.
- Poel, M., Meyer, E. T., & Schroeder, R.** (2018). Big data for policymaking: Great expectations, but with limited progress? *Policy & Internet*, 10, 347–367.
- Porter, T.** (2008). Statistics and statistical methods. In **T. Porter & D. Ross** (Eds.), *The modern social sciences* (pp. 238–250). Cambridge: Cambridge University Press.
- Rule, J.** (1997). *Theory and Progress in social science*. Cambridge: Cambridge University Press.
- Salganik, M.** (2017). *Bit by bit: Social research in the digital age*. Princeton, NJ: Princeton University Press.
- Sen, A.** (2009). *The idea of justice*. London: Allen Lane.
- Steele, J., Sundsøy, P., Pezzulo, C., Alegana, V., Bird, T., Blumenstock, J., ... Bengtsson, L.** (2017). Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, 14(127), 20160690.
- Turow, J.** (2017). *The aisles have eyes: How retailers track your shopping, strip your privacy, and define your power*. New Haven, CT: Yale University Press.
- Watts, D. J.** (2017). Should social science be more solution-oriented? *Nature Human Behavior*, 1(1), 0015[Q8].
- Whitley, R.** (2000). *The Intellectual and social organization of the sciences* (2nd ed.). Oxford: Oxford University Press.