Original Research Article

BIG DATA & SOCIETY

Big Data & Society
January–June 2018: 1–15

DOI: 10.1177/2053951718768831
journals.sagepub.com/home/bds
SAGE

The locus of legitimate interpretation in Big Data sciences: Lessons for computational social science from -omic biology and high-energy physics

Andrew Bartlett[1], Jamie Lewis[2], Luis Reyes-Galindo[3] and Neil Stephens[4]

Abstract

This paper argues that analyses of the ways in which Big Data has been enacted in other academic disciplines can provide us with concepts that will help understand the application of Big Data to social questions. We use examples drawn from our Science and Technology Studies (STS) analyses of -omic biology and high energy physics to demonstrate the utility of three theoretical concepts: (i) primary and secondary inscriptions, (ii) crafted and found data, and (iii) the locus of legitimate interpretation. These help us to show how the histories, organisational forms, and power dynamics of a field lead to different enactments of big data. The paper suggests that these concepts can be used to help us to understand the ways in which Big Data is being enacted in the domain of the social sciences, and to outline in general terms the ways in which this enactment might be different to that which we have observed in the 'hard' sciences. We contend that the locus of legitimate interpretation of Big Data biology and physics is tightly delineated, found within the disciplinary institutions and cultures of these disciplines. We suggest that when using Big Data to make knowledge claims about 'the social' the locus of legitimate interpretation is more diffuse, with knowledge claims that are treated as being credible made from other disciplines, or even by those outside academia entirely.

Keywords

Bioinformatics, physics, big data, social science, locus of legitimate interpretation, epistemic culture

> Q: You are sure that your statement represents scientific truth?
> A: I am.
> Q: On what basis?
> A: On the basis of the mathematics of psychohistory.
> Q: Can you prove that this mathematics is valid?
> A: Only to another mathematician.
> Isaac Asimov, *Foundation* (1951)

Introduction

Over the past decade, 'Big Data' has been positioned as the indispensable mode of 21st century research across academia (Boyd and Crawford, 2012; Kitchin, 2014a). While many of the foundational concepts and techniques of the Big Data sciences were already well-established practices across a number of scientific disciplines, only recently have they been assembled into a distinct field of research claiming legitimacy in and of itself (Beer, 2016; Kitchin, 2014a, 2014b; Ruppert, 2015; Williams et al., 2017). While social science has a

[1]Department of Sociology, University of York, York, UK
[2]School of Social Sciences, Cardiff University, Cardiff, UK
[3]Instituto de Geociências, Universidade Estadual de Campinas – UNICAMP, Campinas, Brazil
[4]Social and Political Sciences, Brunel University London, London, UK

Corresponding author:
Andrew Bartlett, Department of Sociology, University of York, York, YO10 5DD, UK.
Email: andrew.bartlett@york.ac.uk

quantitative history with 'big' datasets dating back to before Durkheim (1897 [2006]), the emergence of 'Big Data' and *computationally-intensive* social science is a contemporary phenomenon. As with much of the discourse surrounding Big Data across the board, there is a tendency to posit the application of 'Big Data' approaches to social science questions as a revolutionary innovation in the profession, both in terms of empirical reach and in theoretical advancement. Lazer and Radford (2017: 20), to cite a recent example, argue that in the span of a generation we "will witness a transformation of sociological theory through these improvements in our ability to observe dynamic social systems." Yet, as Lazer and Radford also maintain, the presence of Big Data research in the leading sociology journals is minimal, with much computational social science currently being carried out not by trained social scientists, but by computer scientists. Supporting this, a sociologist engaged in Big Data research described to us a major conference on Big Data social science as being attended by around "98% computer scientists and physicists and 2% sociologists" and that attendees "weren't engaging with the kinds of questions that [...] sociology would engage with."

While there are, undoubtedly, a significant number of social scientists developing programs of sociologically-informed Big Data work with potential for advancing social science, computationally-derived claims about 'the social' can easily become divorced from, or more worryingly contest the legitimacy of traditional social science that have developed over decades: the theoretical, epistemological, and ontological sensibilities, as well as its ethical and political commitments. In this regard, Big Data social science has 'revolutionary' potential regardless of the content (or success) of its knowledge claims. Or, as McFarland et al. (2016: 32) put it, despite the potential for innovation, it is a legitimate concern that "we may be more likely to witness engineering colonize sociology and the social sciences than vice versa." A decade earlier, Savage and Burrows (2007) also pointed out that the techniques of social research have been incorporated into the circuits of 'knowing capitalism' (Thrift, 2005) as much 'social research' takes place outside, and in (deliberate) ignorance of, academic social science.

When Lazer and Radford (2017: 25) find that what is "[p]erhaps most exciting about Big Data is the opportunity to build a science of society", one is left to wonder what a vast number of social scientists are supposed to have attempted for the past two centuries, if not such a science. Of course, the idea that the social sciences are 'soft', scientifically 'weak' or lack internal disciplinary integrity when compared to the 'hard' and more 'scientifically legitimate' natural sciences, is not new (see, for example, Cole, 1983; Holmwood, 2010; Pinar et al., 2008; Storer, 1967), and these hierarchies are the context in which disciplinary prejudices might shape the future of Big Data applications to social science questions; minimising the role of thinking about the social vis-à-vis elevating that of computational expertise through 'scientific superiority' discourse. Commentators have even gone so far as to provocatively declare on the possibility for a 'methodological genocide', in which "violence [is] being committed under the guise of 'Big Data' at a *methodological* level that is not being discussed."[1] Yet, as we suggest in this paper, the conditions in which the *locus of legitimate interpretation* for computationally-intensive Big Data social science is being manufactured is sociologically quite different to the cases of Big Data biology and physics. This paper uses our in-depth studies of these 'hard' scientific fields to critically probe the question of who is currently poised to be the legitimate interpreter of Big Data social science and the implications this may have for social science research in the future. We take our empirically robust analyses of biology and physics to make suggestions, provocations even, that point towards analyses of 'Big Data' social science that make similar use of Science and Technology Studies (STS)-derived concepts.

Like many of our colleagues, we welcome new methods of researching the social, new ways of addressing social science questions, but offer our use of the concept of the locus of legitimate interpretation in the paper as a contribution to discussions about the ways in which sociological (and other social science) sensibilities can be retained in Big Data research, as a bulwark against domination through technical discourse by badly-practised 'social scientism'.

Who are the *legitimate* interpreters of Big Data social science?

While there is an emerging body of critical, social theory-informed Big Data social science (Cockayne, 2016; Kitchin, 2014a; Kitchin and McArdle, 2016; Niederer and Taudin Chabot, 2015; Symons and Alvarado, 2016), we propose a critical engagement from a different vantage point: Science and Technology Studies (STS). STS has historically dealt with its objects of study – socio-technological systems and research fields – with a highly critical eye, and thus offers a number of tools for probing Big Data computational social science (hereon BDCSS) futures. Additionally, a vast number of STS studies exist analysing the role of scientific fields where Big Data computation has been the dominant experimental paradigm. We present two cases in which the authors have extensive sociological expertise: (i) Big Data biology in the '-omic' sciences, and (ii) in Big Data high-energy physics (HEP). Both disciplines have, for decades,

performed Big Data production and analysis, but they engage computational practices, data analysis and data interpretation in different ways. Furthermore, both assign epistemic legitimacy and power differently to computational versus 'traditional' practices in their fields. The -omic sciences and HEP thus offer a primary, comparative, empirically rich point of departure for understanding the way in which scientific disciplines structure legitimacy of interpretation around computational practices.

Our primary contribution to the debate will be to show that in contrast to sociology, where 'analysing society' is often framed by outsiders as if it were a ubiquitous expertise,[2] biologists and physicists face few, if any, challenges to their monopoly on making legitimate knowledge claims about those aspects of the world encompassed by their discipline. Despite increasingly relying on computational practices that have demanded the introduction of new forms of expertise, the computational aspect in biology and physics is often subjugated as a *tool*, a *service* even, to be used by those with disciplinary grounding in the sensibilities of their discipline. However, this arrangement is not unconnected to the way in which the Big Data of biology and physics is made – crafted within, and for use by, these disciplines – and we therefore describe the critical distinction between such 'crafted' and 'found' data in this paper. As our extensive empirical work on biology and physics shows, the transformation of a discipline into one that produces and uses Big Data need not entail a revolutionary transformation in the locus of legitimate interpretation. However, as our observations of some Big Data claim-making about social science questions also show, Big Data *does* have the potential to erode even further the primacy of knowledge claims about the social made by those with groundings in disciplinary social science, and sociology in particular.

We situate our paper in dialogue with Beer's (2016) work on how we should study the history (and for us, equally the sociology) of Big Data. For Beer, ''we need to place Big Data within the genealogy of social data of various types . . . [and]. . . approach this history by treating Big Data as both material phenomenon and also a concept'' (p. 1). While completely agreeing with this approach, we advocate the value of extending beyond social data to include other comparative articulations of Big Data practices. In this paper, we show how Big Data in the two domains that we have studied in depth (biology and physics) is associated with a change in the arrangements of work that produce and analyse these data, the legitimation of knowledge claims, and, to some degree, the rhetoric of the underlying epistemology, which we use as a basis for our observations of difference in BDCSS. By considering Big Data as an *enactment*, pointing to the importance of how sociomaterial sets of practices *achieve* and *accomplish* Big Data as a meaningful phenomenon, we place Big Data in its social and cultural context. Of course, our notion of enactment does not deny the affordances and impacts of Big Data, nor its increasing popularity as a novel mode of research and powerful research tool. However, for the scientific fields we use to illustrate this argument, we locate 'Big Data' within a sociomaterial account that recognises not only the scientific and technical promises of Big Data, but also the *performative capacity* of the promises, the practices, and the specificities around Big Data. As Beer (2016) argues, Big Data as a concept ''defines, enacts and ushers in'' (p. 9) the materiality it describes.

Fieldwork

Our argument is informed by a set of empirical projects and participatory activities that ran over a 15-year period. These include extended ethnographic contact in multiple sites of scientific work, sets of qualitative interviews, a survey of UK academic bioinformaticians, and the participant comprehension of working in and on Big Data programmes in both biology and physics. Bartlett has conducted postgraduate and postdoctoral research on 'big' biological projects, including the Human Genome Project (HGP) (Bartlett, 2008) and large-scale psychiatric genetics (Arribas-Ayllon et al., 2010). With Lewis he has also conducted research on the development of academic bioinformatics in the United Kingdom (Bartlett et al., 2017; Lewis and Bartlett, 2013; Lewis et al., 2016). In addition to qualitative data, this research involved a survey examining the attitudes and opinions of those working in UK academic bioinformatics (Bartlett et al., 2016). Lewis has also conducted interview and observational research with scientists working in bioinformatics and proteomics including those working at the European Bioinformatics Institute (Lewis, 2008) and has participatory experience of working in research centres engaged in both Big Data biology and Big Data social science. Reyes-Galindo has carried out extended research in the sociology of physics, in particular the mediating role of computational and data analytic cultures in both 'big' and 'small' science settings. This included fieldwork with a computational physics group at the *Conseil Européen pour la Recherche Nucléaire* (CERN) and other physics institutes around the world, and has produced work on the role of algorithms in defining physics communities (Reyes-Galindo, 2016). Stephens has conducted a four-year ethnographic study of the development of biological Big Data tools in the novel context of cell culturing (Stephens et al., Forthcoming; Stephens and Lewis, 2017). Collectively these form a substantial and robust body of studies from which our STS-based analysis of Big Data in practice is drawn.

In what follows, we articulate four key theoretical concepts used in our analysis. We apply them to biology and physics in specific detail, and reflect on their relevance for Big Data applications to social science in a more general way. In doing so, we demonstrate how the histories, institutional forms and power dynamics of a discipline play a part in producing different forms of Big Data enactment; with concrete empirically-grounded examples in the cases of biology and physics, and suggestions of possible futures and research agendas in the case of the social sciences learnt from those examples.

Four key concepts: Enactment, primary/secondary inscriptions, crafted/found data, and the locus of legitimate interpretation

The classic notion of *enactment* in STS captures how scientific work operates to bring into being the knowledge-world it seeks to explore (Pickering, 1995). Studies of sociology (Law and Urry, 2004), economics (Callon, 1998; MacKenzie, 2006), public understanding of science (Michael, 2016), biology (Borup et al., 2006; Brown and Michael, 2003) and physics (Barad, 2007; Galison, 1997; Pickering, 1995) have demonstrated that research visions and methodologies are performative in the making and re-making of scientific disciplines and their knowledge. Here, we show how in biology and physics the notion of Big Data, its manipulation, and the institutional forms that support it, are brought into being through extensive physical, intellectual and symbolic labour and material configuration. Furthermore, these enactments bring with them particular sociomaterial relations, power dynamics and implications for what form Big Data science takes and its efficacy as a research tool.

To do this, we draw upon the notion of *inscriptions* (Latour and Woolgar, 1986), and especially *primary* and *secondary* inscriptions (Lewis and Bartlett, 2013). Laboratory instruments in this metaphor are seen as ''inscription devices'' that ''transform a material substance into a figure or a diagram'' (Latour and Woolgar, 1986: 51), or 'nature' into 'knowledge' – understood as socially legitimised, portable and stabilised 'facts'. In previous work examining Big Data biology, we distinguished between Latourian *primary inscriptions* that transform the material into the symbolic, and *secondary inscriptions* that are the result of separate, distinct transformations of the symbolic into a second set (Lewis and Bartlett, 2013). An example from biology of producing a primary inscription is the physical, material work of drawing blood samples, extracting the DNA, and genotyping the DNA on a gene chip, producing hundreds of thousands of data points. An example of producing a secondary inscription is taking these existing primary inscriptions and conducting Genome-Wide Association Studies (GWAS), using a dataset containing the genomic and phenotypic data of thousands of individuals, in order to discover associations between genetic variants and phenotypic traits.[3] These processes of secondary inscription can be conducted without ever entering a traditional laboratory, and produce a distinct form of representation with different standards.

We extend this analysis from biology to physics and discuss the implications of this way of thinking for Big Data social science. In doing so, we develop a distinction between *crafted* and *found* data. 'Crafted' data are inscriptions produced within the scientific community which will use these inscriptions in order to make knowledge claims. In other words, data that have been *produced with the disciplinary sensibilities of scientists in mind*; specifically oriented, for example, towards answering questions that are meaningful within the discipline. 'Found' data, on the other hand, are inscriptions that 'exist' *independently of the intent and design of the scientific community doing the analysis* (for example, administrative or transactional data that reflects the priorities and purposes of its producers).[4] Found data are 'out there', already existing as inscriptions, independent of any prospective or imagined *disciplinary* use and control. This paper suggests that the differences between the way in which Big Data is enacted in biology and physics, and in its application to social science questions, is related to the fact that the natural sciences craft their own inscriptions, while those applying data to social science questions often draw on existing, 'found' inscriptions from Twitter, Google or Amazon for example.

The notion of the *locus of legitimate interpretation* originates in the work of Collins and Evans (2007: 120). It describes the (social) 'location', in terms of communities and expertise, from which legitimate knowledge claims and judgements of those knowledge claims can be made. To illustrate, we can think briefly about where we would find the locus of legitimate interpretation in two starkly different cases. Collins and Evans argue that the locus of legitimate interpretation for art extends beyond the community of art-producers. Audiences and art critics are treated as legitimate assessors of the quality of the work. By comparison, in the sciences, the locus of legitimate interpretation usually lies well inside the community of producers, as only those with specialist expertise are deemed sufficiently equipped to make valid judgements. When deploying this analytical framework it is important to keep distinct its *descriptive* application – documenting empirically *who* holds legitimacy within a specific context – and its *normative* application – arguing that particular groups *ought* to be deemed legitimate interpreters. In

this paper we do both; by drawing on our empirical sociologies of biology and physics to describe the locus of legitimate interpretation in these cases, and by raising questions about where and how the locus of legitimate is to be found in questions of BDCSS.

Describing this in terms of inscriptions and crafted/found data, we argue that in -omic biology and HEP physics, both primary and secondary inscriptions are *crafted*, analysed, and interpreted within the established scientific communities. By contrast, an emerging paradigm in Big Data applications to social questions is to create secondary inscriptions from data *found* outside the discipline. Further, it is possible to make 'social' knowledge claims using Big Data, which are taken to be legitimate, from outside of the disciplinary cultures of social science. In other words, the 'location' of the making, analysis, and interpretation of big social data, and the judgement on these knowledge claims allows us to say that the locus of legitimate interpretation is much more widely distributed in the social sciences than is the case in biology and physics.

Enacting Big Data in biology

Big Data underwent its foundational enactment as a scientific and political force in biology during the period of 1990–2003 with the HGP. The HGP remains the biggest biological collaboration in history (Hilgartner, 2013; Tripp and Grueber, 2011). It provided the technological and informational platform, as well as the inspiration and model, for the post-HGP data-driven sciences (Collins et al., 2003), such as genomics, proteomics, transcriptomics, and metabolomics, collectively known as the '-omic' sciences. In the post-HGP era, techniques for producing significant amounts of data have meant that both the production and analysis of big biological data have become available to smaller and smaller research groups (Check Hayden, 2014; Grada and Weinbrecht, 2013). The establishment of biology as, at least in part, an informational, computational science, has been accompanied by claims that this new way of doing biology is data-driven (Leonelli, 2016; Stevens, 2013), and even (perhaps erroneously) 'hypothesis-free' (Cooke Bailey et al., 2014).[5]

Biologists, bioinformatics and bioinformaticians

While bioinformatics has a history almost as long as the history of computing itself (Garcia-Sancho, 2012; November, 2012; Strasser, 2010; Suarez-Diaz, 2010), the HGP was the catalyst for its rapid growth and disciplinary infrastructure of conferences, journals, grants and undergraduate and postgraduate courses (Lewis et al., 2016; Stevens, 2013). The move to enacting Big Data brought with it a requirement for a formalisation of mathematical and computer literacy through -omic-oriented bioinformatics (Lewis and Bartlett, 2013; Lewis et al., 2016). This collective enterprise encompasses a broad set of actors including data curators, data analysts, and computer engineers that seek to align computational analysis with large data sets (Harvey and McMeekin, 2002). However, bioinformatics – like any other field – is itself a distinct and recognisable community and set of practices (Bartlett et al., 2017). In the theoretical language developed here, bioinformatics is the work of producing secondary inscriptions through the application of computational techniques to the primary inscriptions made in the laboratory. That is, the data are *crafted* by the biologist and their contact with the material, biological world – be that a cell line, a living organism, or a survey of people – and is then further transformed by a process of *secondary inscription* by a bioinformatician.

The status of bioinformatics is contested within the reward and recognition structures of biology. This is clear in the contrast between Stein's (2008) celebrative account of the total integration of computational methods into biology in the statement "[b]iologists are all bioinformaticians now" (p. 151) with Chang's (2015: 151) pessimistic claims that "there are not enough bioinformaticians" and that "[b]iological data will continue to pile up unless those who analyse them are recognized as creative collaborators in need of career paths" (see also Bartlett et al., 2016, 2017). In our own work, we refer to contestation over the institutional position of computational analysis in biology as the 'middling' of bioinformatics (Lewis et al., 2016), bridging the gap between computer science and biology but as yet not forming its own, coherent, disciplinary space, nor occupying those of its 'parental' disciplines.[6] Although *conceptually* central to the doing of post-genomic science, bioinformatics is *institutionally* peripheral, and is often positioned by biologists as a service to biology (Bartlett et al., 2017; Lewis and Bartlett, 2013), blending into the background (Baren-Nawrocka, 2013). In many cases, despite the rhetoric, we have found that data analysis and computation is not a particularly highly valued or rewarded activity within biology (Lewis et al., 2016).

It should be noted here that much computing expertise is brought into biology from outside the discipline. Bioinformatics is often blackboxed as far as many biologists are concerned – with analysis being conducted by collaborating bioinformaticians, often at a distance (even if they are within the same institution) – or through the use of standardised bioinformatics tools (Lewis and Bartlett, 2013). Some in the field see this as a positive, while others recognise the problems

of this 'collaborative or collective interdisciplinarity' (see Calvert, 2010; Lewis and Bartlett, 2013). For example, while the computational work is performed by bioinformaticians, the burden of analysis is shared with biologists, who through the disciplinary and institutional systems of prestige, retain a dominant position with regard to the locus of legitimate interpretation. Importantly, biologists have institutional 'ownership' of the data of Big Data biology.

There are instances in which 'big biology' produces Big Data – the case of the HGP, for example. Here, bioinformatics can be conducted in large-scale settings, such as the National Centre for Biotechnology Information (NCBI) in the USA, the DNA Data Bank of Japan (DDBJ) and the European Bioinformatics Institute (EBI). But Big Data also produces big science. We observe this not only in the epistemic demands for Big Data to tackle genetically and phenotypically complex disorders, but also in that the resulting Big Data (and the techniques and technologies developed in these projects) enables (scientifically, institutionally, and 'politically') further big science projects. This is clear in the way in which the accomplishment and legacy of the HGP has helped to shape many other satellite centres of post-HGP -omic science. The scale of bioinformatics projects therefore can also be much smaller, sometimes involving only a handful of researchers. Such smaller science Big Data work often draws on computational and statistical expertise from a centralised group within their host institution. Many universities establish a central bioinformatics hub for its researchers to work with when they see fit. Yet these smaller science settings may still rely on components of big science, by drawing upon the training, data, applications, and collaborative skills of institutions such as the EBI. Furthermore, the work of these smaller science Big Data projects can be aggregated using standardised protocols for data collection and recording (Wallis et al., 2013). This is already underway in the '-omic sciences' (Harvey and McMeekin, 2010; Leonelli, 2012, 2013), and is spreading to other areas of biological research such as cell culturing (Khan et al., 2014). Reflecting this, big biological data science is promised as being geography-free, as collation disentangles it from the peculiarities and particularities of localised settings, with global infrastructures allowing seemingly 'frictionless' international flow.

The epistemic culture (Knorr-Cetina, 1999) of biology has shifted in the post-HGP era. Some argue this is a move from 'hypothesis-driven' research into an era of 'hypothesis-free' biology (Cooke Bailey et al., 2014), although, perhaps more accurately, the move to data driven biology is a change from deductive to inductive reasoning (Leonelli, 2012). This change has already been institutionalised and embedded within the distinctive nomenclature of the '-omic' sciences. An important point must be reiterated with regard to this new, inductive, mode; Big Data biology '*crafts*' its data. The vast databases of -omic data that are said to 'drive' much contemporary biology have been crafted in a laboratory, by technicians and scientists trained in biological ways of thinking and according to the disciplinary sensibilities of biologists. Even when computational biologists come to use these large data sets at one step removed, having played no part in the production of the data, the data that they use are still crafted within the epistemic culture of biology.

The locus of legitimate interpretation for Big Data biology is located firmly within the epistemic, disciplinary culture of biology: data are produced within the discipline, in laboratories, by biologists, or by computer scientists with biological sensibilities in mind. That is, although computational and statistical expertise has been drawn into the discipline, bringing with it a new style of statistical reasoning (Leonelli, 2012; Lewis et al., 2016), it has been done so in a way that positions it subordinate to the disciplinary concerns of biology (Lewis and Bartlett, 2013). We now turn our attention to Big Data in physics.

Enacting Big Data in physics

Historical and ethnomethodological studies identify three families of practice in physics: theory, phenomenology, and experiment. Phenomenology encompasses the cumulus of disciplines that does the bulk of the 'translation' between theory and experiment (Galison, 1997; Merz and Knorr-Cetina, 1997; Reyes-Galindo, 2011). Specifically, Reyes-Galindo (2014) describes physics as being structured around two opposite poles of practices: theoretical and experimental, but with many intermediate micro-cultures mediating the transmission of knowledge across them. In all these micro-cultures, computation has for a long time been an important element, particularly in HEP experiments, which are nowadays recognised as forerunners of Big Data science (Murray, 2014). Yet it is only recently, with the rise of 'Big Data' rhetoric in the media and in commercial and academic discourse, that physicists have begun to market their traditional practices as 'Big Data'. Indeed, the first occurrences of the term 'Big Data' in the physics arXiv preprint server – the single most important resource for vanguard physics – are as recent as 2013 (Anderson et al., 2013), while the earliest mentions of 'Big Data' related to physics in the scientific press refer not to a discourse on the promises or possibilities of Big Data, but to the problems of sustainable and reliable computational infrastructure that Big Data sets imply (Lynch, 2008).

The locus of legitimate interpretation in HEP

Nowhere in physics has the rapid accumulation of vast amounts of data been more visible than in HEP, as experiments have increasingly demanded more data-points to reach the confidence levels required by physicists to claim that a 'finding' is in fact a 'discovery' (5 standard deviations are the norm in HEP). The paradigmatic case of Big Data in physics is CERN and the Large Hadron Collider (LHC) experiments (Knorr-Cetina, 1999; Kriege, 1996), though other projects such as the Sloan Digital Sky Survey and other sky-mapping experiments in astronomy and astrophysics also produce terabytes of analysable data (Zhang and Zhao, 2015). For example, one of the core CERN experiments, Compact Muon Solenoid (CMS), produces around 1 petabyte (100 gigabytes) of 'raw' data per second, and there are similar figures for the other experiments. The quantities of data produced are only expected to increase, although CERN currently only stores in the order of 35 petabytes a year as the overwhelming majority is filtered out. Yet a modern physics experiment does not necessarily require the size and complexity of CERN to reach the data-acquisition numbers of CERN. The smaller-scale, Mexico-based High-Altitude Water Cherenkov Observatory (HAWC) international collaboration generates about 1 terabyte of data per day, just under the same order of magnitude as the data produced at CERN, but – unlike CERN – records all data for possible later analyses (Gitler and Klapp, 2016).

In big physics settings such as CERN, which involves work in a number of highly specialised areas, the multiple loci of legitimate interpretation are found in the collectively-vetted effort. Locating these is somewhat more complex than the situations captured by Collins and Evans' (2007) portrayal of inter-expertise dialogue and meta-expertise interdisciplinary management, as the multiple expertises in CERN often overlap and become too complicated for a single actor to fully comprehend. Despite the detectors being physically grouped at two sister sites near Geneva and there being a known set of core group leaders, CERN is a globally-distributed knowledge-producing network in which the acquisition, handling and processing, and interpretation of the data is carried out by several independent communities within 'the experiment'. The data are cleaned even as it is being acquired, recorded and then interpreted and re-interpreted in several steps in which interpretative legitimacy is 'lent' to the expert groups that intervene in each step. Once all the steps come together, a final stable consensus is reached after a long gruelling process of micro-data crafting, for example, when a 'discovery paper' is published in collaborative authorship. The importance of each parallel interpretational mesh-point and the locality of each step is made most obvious by the number of authors in contemporary discovery 'megapapers' – the Physics Letter B paper announcing the discovery of the Higgs boson was signed as 'CMS collaboration' and 'ATLAS collaboration' and jointly included more than 6,000 authors. As a senior computational physicist remarked in interview:

> ''Nowadays even building the detectors has become an industrial enterprise. In the past, a group or a small set of groups was responsible for designing, building, operating the whole detector, the calorimeter, the vertex detector, particle identification detector [...] Nowadays, each detector is an enterprise of many institutes, many people, so you don't even get the overview of the whole detector.''

Though distributed among the collaboration teams, the locus of legitimate interpretation remains within 'the experiment' as a whole; that is, within the CERN community. The distributed interpretation makes it impossible for a single member of the collaboration to draw away into a personal interpretation of the entire experimental setup, as Delfanti (2016) has described in his discussion of deliberative democracy methods for producing authorship in HEP. Once the 'results' have been stabilised within each interpretative step in a multiplicity of primary inscriptions, they are then collectively cohered into project-overarching secondary inscriptions which are then put into collected tables of 'definitive' data, such as the massive Review of Particle Physics (RPP) published by the Particle Data Group at the Lawrence Berkley National Laboratory.

Bibliometric investigations (Basaglia et al., 2008) and fieldwork at CERN by Reyes-Galindo suggest that, as described by Lewis and Bartlett (2013) in bioinformatics, computational physics is generally regarded as a less prestigious activity than other areas of research, despite its importance for the generation of experimental outcomes. That programming and computation in scientific settings – and specifically in physics – is, broadly speaking, regarded as ''production rather than research'' (Slayton, 2013: 38) and is known to be a feature not just of scientific computing but of the whole field of programming (Ensmenger, 2012). It is therefore unsurprising that the critical component of data analysis at CERN, infrastructure computing, is generally looked down upon as 'technical' and 'service' work and is perceived as an activity at some remove from the prestige of 'real' research.

Low-level data reconstruction, though seen as being closer to 'research' and requiring significantly more specialised knowledge of physics, is still seen as being less

prestigious than other research practices such as hardware development. This work – save for a few individuals who are considered the leading experts of their fields – is the domain of the graduate students and postdoctoral researchers who do most of the grunt work. Nevertheless, informants in Reyes-Galindo's empirical research described the way in which reconstruction is divided into many subspecialties. Each of these requires intensive specialisation that is often experiment-specific. Between these specialities hierarchies have formed, though in general reconstruction work is of lesser status to that of conducting 'original' research. Speaking of the status of computational work, a senior computational physicist at CERN reflected that:

> ''you're not a technician because you still have a degree in physics, often a PhD anyway, but nevertheless you are not doing the bulk of the attraction of the field . . . [computational work] is not itself, you know, the most attractive topic to talk about.''

In other words, computational physics is not only subordinate to other modes of physical thinking, it is a dispreferred way of working and thinking when compared to those which attracted physicists to the field in the first place. The production of primary inscriptions at CERN (the local experimental groups, e.g. CMS, ATLAS) and the *interpretation* (not production) of secondary inscriptions (overarching high-level analysis) are the most esteemed and desirable 'scientific' work, valued far above the 'technical' domain of primary inscription analysis (both general and experiment-specific IT services). Data are crafted according to the sensibilities of physicists, and much of the computational and statistical expertise is found within the discipline. Yet all these practices are arranged within the experimental organization. Through the coordination of all these local practices, modern HEP crafts its own data and keeps the locus of legitimation interpretation firmly inside the scientific community. For all their differences in epistemic culture, -omic biology and HEP demonstrate important similarities in the way in which they have incorporated 'Big Data'.

Thoughts on the social sciences, computational social science, and the Locus of Legitimate Interpretation

So far, in this paper, we have analysed the ways in which data are crafted within big biology and big physics, and which communities and bodies of expertise are deemed to be the legitimate interpreters of that data. In this section, we make some observations about the way in which the application of Big Data to social science questions can be enacted in a fundamentally different way to the examples provided by the 'harder' sciences. In the social sciences, Big Data can exist independent of the labours of social scientists, described in this paper as 'found'. This is often posited as one of the *epistemic strengths* of Big Data social science, despite the assumptions that must be made about data found outside the discipline, regarding, for example, the comprehensiveness and representativeness of online populations, etc. (Lazer and Radford, 2017). This is a fundamental epistemic difference between the social and the natural sciences with regard to the relationship between the 'scientist' and her 'data'. For the most part, only physicists and biologists are legitimate interpreters of Big Data produced in physics and biology; the locus of legitimate interpretation is firmly within the disciplinary community. However, the ability to make a knowledge claim about the social that is treated as credible is afforded to a much wider spread of people. As we discussed in the opening sections, the locus of legitimate interpretation in Big Data applications to social science questions is much more diffuse. Thus, the *organisational* and *epistemic* model of Big Data science that we find in the natural sciences does not find a direct reflection in Big Data social science.

In this section, we provide some clear examples of the way in which Big Data applications to social science questions can be performed outside of established social science communities. It is important to stress that these examples are not used to suggest that they are *representative* of BDCSS as a whole. While we have conducted extensive sociological research on physics and biology, as yet we have no solid research program to this end. Rather, these examples are intended to serve as illustrations of the way in which Big Data *can* be performed and scientifically positioned as 'legitimate' social science. This is crucial because, as we show in our final discussions, recent studies in the sociology of physics have shown that analysis of physics 'data' performed outside of the traditionally-constituted locus of legitimate interpretation is overwhelmingly rejected as 'crackpot science' (Collins et al., 2017). This is true even if the knowledge claims that are being produced are the technically-savvy products of people with significant expertise in physics or related disciplines. Knowledge claims produced by outsiders are almost never considered legitimate by the physics community, and are often portrayed as the antithesis of 'good' physics. The boundaries of physics set by physicists match very closely the boundaries of good physics as seen by funders, policy makers, science journalists, etc. The examples in this section show that, in the case of 'social' questions, the locus of legitimate interpretation is 'diffused', extending outside the established, disciplinary social sciences.

We start with a *Nature* special feature article (Giles, 2012), which described the status of 'computational social sciences' research. This article discussed several examples, such as the research carried out by Liben-Nowell and Kleinberg (2007) which supported existing social science claims about social networks, as well as that challenged established social science views (Ugander et al., 2012). Critically, all the studies mentioned in the feature were carried out not by social scientists but by computer scientists. Kleinman is quoted describing how he "... realized that computer science is not just about technology", but rather "[i]t is also a human topic" (Giles, 2012: 448). Kleinberg also adds how he thinks of himself "as a computer scientist who is interested in social questions" (Giles, 2012: 450). Nowhere in the feature is the absence of social science knowledge and expertise portrayed negatively, except possibly in terms of way in which these researchers are not tied into addressing questions that are interesting to those working in established fields of social research. It is, according to this view in *Nature*, scientifically *legitimate* for a computer scientist to conduct research into 'social phenomena' despite having, in the best cases, low-level working knowledge of social phenomena, traditional social science methods and social science theory.

To make the above asymmetry clearer, we next turn to another *Nature* feature by the same author that discussed the *opposite* case, that of social scientist claiming to be a legitimate interlocutor *about* (not in) a natural science field. Giles (2006: 8) describes how sociologist Harry Collins had to prove, through an incredibly difficult 'imitation game' test judged by a panel of gravitational wave physicists, that through *thirty years' experience* interacting directly with the gravitational waves community he had acquired sufficient 'interactional expertise' to meaningfully and *legitimately* speak the language of gravitational waves (Collins et al., 2006). As Giles comments, Collins' point about legitimacy was one of the most strongly contested positions of the 1990s 'Science Wars' in which some natural scientists were angered by the fact that "sociologists studying science did not understand the disciplines involved, in part because they did not practice them" (Giles, 2006: 8). In fact, the asymmetry is even more extreme when we realise that the Science Wars criticism of social science legitimacy was not about social scientists practising natural science (which no sociologist of science would claim to do), but indeed only on talking *about* the natural sciences.

While some in computational social science stress the revolutionary aspects of their work, others pursue the research agenda without exclamation on its novelty, rending it normal and uncontroversial. Such work includes people-centric sensing and social sensing to track physical sensors in mobile devices to "learn about (possibly hidden) social structures" (Campbell et al., 2008: 20) and "infer social relationships" (Krishnamurthy and Poor, 2014: 3). In the latter, interaction through social media posts are analysed to produce models that "facilitate understanding the dynamics of information flow in social networks and, therefore, the design of algorithms that can exploit these dynamics to estimate the underlying state of nature" (Krishnamurthy and Poor, 2014: 3). As the authors explain, the "motivation for th[eir] paper stems from understanding how individuals interact in a social network and how simple local behavior can result in complex global behavior." They defend their methodology by pointing out that "[t]he underlying tools used in this paper are widely used by the electrical engineering research community in the areas of signal processing, control, information theory, and network communications" (Krishnamurthy and Poor, 2014: 19). Similar analytical forms have been applied to studying emergency events (Xu et al., 2016), online rumour detection (Liu and Xu, 2016), and appreciation of cultural heritage (Pilato and Maniscalco, 2015). Often connected in some way to the *Institute of Electrical and Electronics Engineers*, these publications operate in the space where computer science, social media, and social analysis overlap, yet they are conducted largely in isolation of the traditional knowledge and expertise bases of social science.

The differences between the enactment and positioning of Big Data in the social and the natural science are also clear in the work of another computer scientist featured in the Nature article – Alex Pentland – and his fascinating (and revealingly titled) book Social Physics (Pentland, 2014). Here, 'found' big social data is described as "the millions of digital bread crumbs people leave behind via smartphones, GPS devices, and the Internet" (p. x). To put it in the terms offered by Latour and Woolgar (1986), the primary inscriptions that constitute big social data are 'written' independent of academic big social data practices. *Social Physics,* and other recent pop-social science books (such as Stephens-Davidowitz's *Everybody Lies*), promise a *new* and *revolutionary* social science, in which society is understood in terms of relationships between and within data written by our interaction with, among other things, the digital economy, and in which knowledge claims about society are made by experts in data analysis, with the sensibilities of social science largely irrelevant in the face of the new data-rich world. In this, we hear echoes of the rhetoric of 'hypothesis free' science that Big Data has brought to biology (Cooke Bailey et al., 2014). This view was similarly put forward in an academic review-cum-manifesto for computational social science in which, it is concluded that

through computational social science "sociology in particular, and the social sciences in general, would undergo a dramatic paradigm shift, arising from the incorporation of the scientific method of physical sciences" (Conte et al., 2012). Indeed, Pentland's vision (exciting though it is) for a 'social science' of studying information exchange without knowledge of the content or meaning *is indeed* a radical (and revolutionary) departure from the intellectual mission of much of 20th century social science – that which aims for the kind of *comprehension* of human socialities that can be gained by slower, craft-orientated methods such as ethnography.[7]

Will such a 'new discipline' be "an example of statistics-led research with no theoretical underpinning"? This is how Professor Susan McVie, professor of quantitative criminology at the University of Edinburgh, responded to the publicity surrounding a recent paper uploaded to the most important 'hard' science e-print server, the arXiv (BBC, 2016; Wu and Zhang, 2016). This paper claimed that, using supervised machine learning, the authors – who work in an Electrical Engineering department[8] – had developed a system for distinguishing criminals from non-criminals (or as the authors label them, 'normal people'), with criminals successfully identified 89% of the time. McVie is quoted by the BBC as stressing the various biases involved in producing a criminal conviction – the 'found data' used by Wu and Zhang – pointing out that "[t]his article is not looking at people's behaviour, it is looking at criminal conviction". Using the vocabulary proposed here, McVie is not only highlighting the weakness of naïvely found data, but is demanding that the locus of *legitimate interpretation* of 'Big Data criminology' remains within criminology, a community able to draw on decades of collective knowledge of dealing with crime statistics, as well as understanding the biases and cultural differences in criminal justice systems, etc.[9] Surprisingly, McVie's view did not find support from prestigious voices in 'harder' fields of science and technology. Quite the contrary. The *MIT Technology Review*, for example, though acknowledging the study as "controversial", supported Wu and Zhang by noting that their work was consistent with a previous 2011 psychological experiment from Cornell University (Emerging Technology from the arXiv, 2016). We stress that we are not, *per se*, against the new possibilities afforded by computational social science, but rather worried by computational exercises such as Wu and Zhang's study that rely on the rhetorical weight of Big Data to convey epistemological strength on its own. It is telling of the state of things that even critical (yet optimist) views on the impact of computational social sciences on traditional social sciences call for social scientists to "embrace Big Data" (González-Bailón, 2013), while computational experts dealing with social phenomena are rarely called to conversely embrace traditional sociological tradition or thought in their research.

Wu and Zhang did not engage with existing criminological research yet their claims to a contribution to criminology were treated seriously. Their example, egregious though it is, shows how the naïveté of a 'hypothesis-free social analysis' can mutate into a pathological form in which knowledge claims are produced which turn back decades of careful empirical, conceptual, and ethical work. Other recent research, such as facial recognition of 'sexual orientation' through deep-learning algorithms (Wang and Kosinsky, 2018) also work against the grain of informed reflections on stigmatised populations and reveals the intricate problems linked to disciplinary-uninformed interpretations of 'social' Big Data. There is therefore a problem beyond mere epistemological or methodological quibbling in using 'found' data without sociological insight. Pentland's *Social Physics* is subtitled 'lessons from a new science', and this is perhaps exactly the point. While biology and physics are, to a greater or lesser degree, enacting 'Big Data' by absorbing a new way of looking at the objects of their disciplinary gaze into the body of their disciplines, the locus of legitimate interpretation of claims about 'the social' is so broad that Big Data social science can be enacted outside traditional social science disciplinary boundaries, even when it is conducted inside academic institutions, and afforded public legitimacy without much say by social scientists. While we hope, like optimistic social scientists such as Smith (2014), that in a 'Big Data social science' sociologists will be required to interpret (and critique) the outputs, we worry that the cultural legitimacy of such demands appears to be weaker than might be needed in order to make this so.

Discussion

The enactment of Big Data can tell us something about the differences between disciplines and between epistemic cultures, especially by concentrating on such notions as 'crafted' and 'found' data and the 'locus of legitimate interpretation'. We have presented two disciplinary case studies on Big Data epistemic cultures, illustrated their relationships to crafted and found data, and shown how each disciplines' locus of legitimate interpretation is structured and connected to the cultures of primary/secondary data producers and interpreters in each field. Physics, with a long tradition of dealing with Big Data, 'produces' its own computer scientists, and 'Big Data' physics is, mostly, conducted within the disciplinary space of 'physics'. In other words, the way in which Big Data has been enacted is

in response to, and in sensitivity to, the disciplinary needs and priorities of physics.

As described, Big Data biology, a more recent development than Big Data physics, has had to recruit expertise from outside the discipline. Even though there are claims made that Big Data biology is a revolutionary new form of hypothesis-free science, the locus of legitimate interpretation still remains firmly within biology. Expertise in data analysis alone is not deemed sufficient to make legitimate biological knowledge claims. Biologists, as the creators of the primary inscriptions and the holders of cultural and institutional power, are the legitimate interpreters of Big Data biology, with the computer scientists/bioinformaticians who produce the 'secondary inscriptions' being dependent on, and deferring to, biologists. Bioinformatics may be an offshoot of biology, but it is tied inextricably to the disciplinary culture and institutions of biology. Both of these natural sciences enact Big Data science in a significantly different way to that in which it is being enacted in the social sciences.

Unlike the biologists and physicists, social scientists in many cases do not to have disciplinary control over the production of the Big Data that they will use – it is not crafted but is, instead, found. As such, social scientists can make no claims of exclusivity or control over this data; anyone with the computational skills can conduct an analysis of social media, and as Metzler et al. (2016) suggest, it is rare that social scientists have the required computational skills. So, as with biology, in order to enact Big Data science, social science must recruit computational and statistical expertise. However, given that social scientists are not (always) the crafters of big social data, their sensibilities are not written into these inscriptions. Further, claims that Big Data allows an atheoretical, hypothesis-free analysis of the social gain traction due to the low esteem in which much social science is held. The consequence of this is that the locus of legitimate interpretation is not firmly fixed within the communities trained in the social sciences. Anyone can make an acceptably credible knowledge claim, whether by virtue of controlling (access to) the primary inscriptions – as is the case with proprietary data – or on the basis of bring to bear the tools and perspectives of 'harder' disciplines.

Kate Metzler (2016) quotes Clive Humby, the man responsible for Tesco's Clubcard scheme, as saying as long as a decade ago that 'data is the new oil'. In this account, it is not just *a* resource, but *the* resource of the 21st century, and those who control the data will have tremendous economic, social and political power (Boyd and Crawford, 2012). Even as the 'found' character of much big social data renders some questions tractable and others unaskable, data grants power to those asking social questions working with and within the organisations – often private – which hold the data (Beer, 2016). Metzler et al.'s (2016) survey of Big Data research in the social sciences found that, out of 3077 respondents involved in Big Data research, just over half (1690) had most recently used administrative data, 927 used social media data, and 697 used commercial/propriety data. In current BDCSS practice, both the locus of legitimate interpretation and the ownership and control of data can lie outside the boundaries of social science as social scientists wrestle with others over control of empirical materials and the right to analyse it.

This contrasts sharply with physics. Bartlett and Reyes-Galindo have carried out extensive empirical analysis regarding the legitimacy of physics claims made by scientists who are not professional, practising physicists (Collins et al., 2017). This physics 'boundary work' (Gieryn, 1983) has shed light on the sociological structures of so-called 'fringe' or unorthodox physics and, importantly, to the relationships between producers of physics' primary inscriptions and outsiders to the physics community. What is observed is that, in physics, the *legitimacy* of primary and secondary inscription production is highly closed in itself: those that produce primary inscriptions belong to the same social group (or network) as those that produce secondary inscriptions, and the legitimacy of interpreting the results *legitimately* is based on belonging to these social networks, not on personal characteristics or specific skills.

There *are* cases in which 'outsiders' to these closed networks attempt to create alternative readings of established physics, such as when mathematically-informed engineers (and particularly electrical engineers) re-evaluate recognised theoretical claims or experimental results. The 'exclusion boundary work' that follows is the same across all the fields of physics explored. Outsiders are ignored when they are not scientists, isolated when they are practising scientists, and in the more extreme cases ridiculed and declared 'cranks' or 'crackpots' by the scientific community (Reyes-Galindo, 2016). Compare this to the response to a criminology paper produced by electrical engineers; 'social physics' and hypothesis-free Big Data social science have developed into legitimate areas that are autonomous *and* authoritative despite their revolutionary intent. The locus of legitimate interpretation in physics presents strong social closure, while in social science it is considerably more open.

Thus, while in physics outsiders who attempt to overturn established knowledge claims or methods are de-legitimised *because* of their status as an outsider, a significant part of (for example) 'social physics' legitimacy-talk hinges on the *strengths* of being an outsider.

By contrast, the physics community would and does act swiftly in cases where not only individual knowledge claims but also its *professional legitimacy* is contested.

With reference to our normative (rather than descriptive) application of the concept of the 'locus of legitimate interpretation', we draw attention to what might be the consequences of the differences between the tightly bounded locus of legitimate interpretation found in biology and physics, and the much more diffuse, contested locus found in the social sciences. The Wu and Zhang episode suggests that an extreme flexibility of the locus of legitimate interpretation can lead to 'pathological' data-driven social science that can, importantly, be taken seriously. Sensitivity to the pathological dimensions of this kind of work is not necessarily found outside of the social sciences – that is, those with the technical expertise required to make expert judgements. Furthermore, one can easily think of research (involving stigmatised populations and minorities, race and gender relations, etc.) in which theoretically and ethically uninformed data-driven social science could have quite profoundly negative wider impacts, if given legitimacy. At the very least, critical accounts of Big Data sociology are required to counterbalance the data-driven hype. Social scientists should not be shy of performing their own boundary work.

While Big Data-driven social science has presented itself as immensely disruptive to existing research, and certainly introduces new tools, methods and possibilities to probe societies and cultures, our research resonates with previous discussions about the importance of examining Big Data claims with greater scrutiny and clarity (Beer, 2016). We resist the picture that the future of social science is made up exclusively of Big Data-given research, even while acknowledging that Big Data sociology can become a parallel field of research to 'traditional' sociology. However, we do argue there is a key issue for social science in terms of retaining and monitoring control over which collectives and individuals constitute the locus of legitimate interpretation in BDCSS. Unlike physics, this has been complicated for social science due to the lack of substantive base in computational mathematical methodologies, and unlike both physics and biology, because of the relative intellectual prestige of social science at the 'softer' end of the disciplines. This paper has argued that STS provides the basis for important critique of Big Data science. Following Eyal (2013), there is a question of jurisdiction as to who has control over a set of tasks and who are the legitimate interpreters of the findings. There is also a question as to what social and institutional arrangements need to be in place for that authority to be maintained, and in what situations it can be challenged. Developments in computation and access to large data sets (as well as pre-existing hierarchies) have meant that sociologists and other social scientists face challenges to be the legitimate interpreters of social data in ways that biologists and physicists do not.

Acknowledgement

The authors are grateful for the previous early-career support of the Economic and Social Research Council (ESRC) and the British Academy.

Declaration of conflicting interests

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

Funding

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

Notes

1. Carrigan M. Emma Uprichard: Big Data and 'Methodological Genocide', editorial, *Methodspace*. Accessed 9 September 2017. URL: https://www.methodspace.com/emma-uprichard-big-data-methodological-genocide/
2. Indeed, this goes beyond mere hierarchies of disciplinary prestige and we grant that there is a minimal 'commonsense' sociological knowledge intrinsic to all individuals living in any society; all socialised individuals must have some tacit capacity to understand and analyse society in order to live in society. By comparison, these individuals need no such understanding of biology to keep their blood flowing, or physics to prevent them spinning off into space.
3. It is important to note that these statistical associations are not themselves 'interpretations'. Bartlett, attending a psychiatric genetics workshop, observed a senior bioinformatician present the statistical associations discovered during their work who ended his presentation by saying that he couldn't tell you what any of this meant *biologically*, and that it was the job of the biologists present to perform the interpretation. This anecdote also points us towards where the locus of legitimate interpretation is to be found in Big Data biology.
4. The distinction between found and crafted data is well known within qualitative social science, though not always articulated in these terms. For example, in diary analysis, crafted diary data would involve asking (perhaps even training) participants to complete a diary of their experiences as part of the research (see Alaszewski et al., 2007) whereas found diary data would be analysis of diary entries the participants created independently of the research (see Coffey, 2014).
5. Anderson (2006) boldly announced that the advent of the 'Petabyte Age' rendered theory and the scientific method 'obsolete'. As with many commentators, he talked of 'Big Data' in the language of a natural event – in this case as a 'deluge'. Franks (2012), for example, steps up the level of

destruction (disruption?) and describes it as a 'tidal wave', while Steimle (2015) warns of us 'drowning in Big Data'.
6. It is important to recognise that we are making a distinction here between bioinformatics as a recognisable community and bioinformatics as a legitimate discipline that exists independently of biology.
7. There has been some excitement about a future social science that moves beyond 'outmoded' methods devised for 20th century societies, unfit for new, 21st century forms of sociality (see, for a starting point, Savage and Burrows, 2007).
8. Curiously, electrical engineers are well represented in 'fringe' physics communities (Collins et al., 2017).
9. In contrast to Wu and Zhang's theoretically light work, Williams et al. (2017) conducted an ESRC funded Big Data study using classic criminological theory to inform the collection, transformation, classification and modelling of over 200 million tweets to identify their affordances and limitations in relation to crime pattern estimation.

References

Alaszewski A, Alaszewski HP, Potter J, et al. (2007) Working after a stroke: Survivors' experiences and perceptions of barriers to and facilitators of the return to paid employment. *Disability and Rehabilitation* 29(24): 1858–1869.

Anderson C (2006) The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine* 23 June.

Anderson J, Brock R, Gershtein Y, et al. (2013) Benefits to the U.S. from physicists working at accelerators overseas. arXiv: 1312.4884.

Arribas-Ayllon M, Bartlett A and Featherstone K (2010) Complexity and accountability: The witches' brew of psychiatric genetics. *Social Studies of Science* 40(4): 499–524.

Barad K (2007) *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Durham, NC: Duke University Press.

Baren-Nawrocka TV (2013) The bioinformatics of genetic origins: How identities become embedded in the tools and practices of bioinformatics. *Life Sciences, Society and Policy* 9(7). DOI: 10.1186/2195-7819-9-7.

Bartlett A (2008) *Accomplishing sequencing the human genome*. Unpublished doctoral dissertation, Cardiff University.

Bartlett A, Lewis J and Williams ML (2016) Generations of interdisciplinarity in bioinformatics. *New Genetics and Society* 35(2): 186–209.

Bartlett A, Penders B and Lewis J (2017) Bioinformatics: Indispensable, yet hidden in plain sight? *BMC bioinformatics* 18(1): 311.

Basaglia T, Bell ZW, Dressendorfer PV, et al. (2008) Writing software or writing scientific articles? *IEEE Transactions on Nuclear Science* 52(2): 671–678.

BBC (2016) Convict-spotting algorithm criticised. Available at: http://www.bbc.co.uk/news/technology-38092196 (accessed 22 December 2016).

Beer D (2016) How should we do the history of Big Data? *Big Data & Society* 3(1). DOI: 10.1177/2053951716646135.

Borup M, Brown N, Kondad K, et al. (2006) The sociology of expectations in science and technology. *Technology Analysis & Strategic Management* 18(3–4): 285–298.

Boyd D and Crawford K (2012) Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15(5): 662–679.

Brown N and Michael M (2003) A sociology of expectations: Retrospecting prospects and prospecting retrospects. *Technology Analysis and Strategic Management* 15(1): 3–18.

Callon M (1998) *Law of Markets*. Oxford: Blackwell Publishers.

Calvert J (2010) Systems biology, interdisciplinarity and disciplinary identity. In: Parker J, Vermeulen N and Penders B (eds) *Collaboration in the New Life Sciences*. Farnham: Ashgate, pp. 201–219.

Campbell A, Eisenman S, Lane N, et al. (2008) The rise of people-centric sensing. *IEEE Internet Computing* 12(4): 12–21.

Chang J (2015) Core services: Reward bioinformaticians. *Nature* 520: 151–152.

Check Hayden E (2014) The $1000 genome. *Nature* 507: 294–295.

Cockayne DG (2016) Affect and value in critical examinations of the production and 'prosumption' of Big Data. *Big Data & Society* 3(2). DOI: 10.1177/2053951716640566.

Coffey A (2014) Analysing documents. In: Flick U (ed.) *The Sage Handbook of Qualitative Data Analysis*. London: Sage, pp. 367–379.

Cole S (1983) The hierarchy of the sciences? *American Journal of Sociology* 89(1): 111–139.

Collins FS, Morgan M and Patrinos A (2003) The human genome project: Lessons from large-scale biology. *Science* 300(5617): 286–290.

Collins HM and Evans R (2007) *Rethinking Expertise*. Chicago, IL: University of Chicago Press.

Collins HM, Evans R, Ribeiro R, et al. (2006) Experiments with interactional expertise. *Studies in History and Philosophy of Science Part A* 37(4): 656–674.

Collins H, Bartlett A and Reyes-Galindo L (2017) Demarcating fringe science for policy. *Perspectives on Science* 25(4): 411–438.

Conte R, Gilbert N, Bonelli G, et al. (2012) Manifesto of computational social science. *The European Physical Journal Special Topics* 214(1): 325–346.

Cooke Bailey JN, Pericak-Vance MA and Haines (2014) Genome-wide association studies: Getting to pathogenesis, the role of inflammation/complement in age-related macular degeneration. *Cold Spring Harbor Perspectives in Medicine* 4(12): a017186.

Delfanti A (2016) Beams of particles and papers: How digital preprint archives shape authorship and credit. *Social Studies of Science* 46(4): 629–645.

Durkheim E (1897 [2006]) *On Suicide*. London: Penguin Books.

Emerging Technology from the arXiv (2016) Neural network learns to identify criminals by their faces. *MIT Technology Review*, 22 November. Available at: https://www.technologyreview.com/s/602955/neural-network-learns-to-identify-criminals-by-their-faces/ (accessed 22 December 2016).

Ensmenger NL (2012) *The Computer Boys take Over: Computers, Programmers, and the Politics of Technical Expertise*. Cambridge, MA: MIT Press.

Eyal G (2013) For a sociology of expertise: The social origins of the autism epidemic. *American Journal of Sociology* 118(4): 863–907.

Franks B (2012) *Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics*. Hoboken, NJ: John Wiley & Sons.

Galison P (1997) *Image and Logic: A Material Culture of Microphysics*. Chicago, IL: University of Chicago Press.

Garcia-Sancho M (2012) *Biology, Computing and the History of Molecular Sequencing: From Proteins to DNA, 1945–2000*. Basingstoke: Palgrave Macmillan.

Gieryn TF (1983) Boundary-work and the demarcation of science from non-science: Strains and interests in professional ideologies of scientists. *American Sociological Review* 48(6): 781–795.

Giles J (2006) Sociologist fools physics judges. *Nature* 442(7098): 8.

Giles J (2012) Making the links. *Nature* 488(7412): 448–450.

Gitler E and Klapp J (2016) High performance computer applications. In: *6th international conference, ISUM 2015. Communications in computer and information science*, Mexico City, Mexico, 9–13 March 2015, p. 595, Switzerland: Springer International Publishing.

González-Bailón S (2013) Social science in the era of big data. *Policy & Internet* 5(2): 147–160.

Grada A and Weinbrecht K (2013) Next-generation sequencing: Methodology and application. *Journal of Investigative Dermatology* 133: e11.

Harvey M and McMeekin A (2002) *UK Bioinformatics: Current Landscapes and Future Horizons*. London: Department of Trade and Industry.

Harvey M and McMeekin A (2010) Public or private economies of knowledge: The economics of diffusion and appropriation of bioinformatics tools. *International Journal of the Commons* 4(1): 481–506.

Hilgartner S (2013) Constituting large-scale biology: Building a regime of governance in the early years of the human genome project. *BioSocieties* 8: 397–416.

Holmwood J (2010) Sociology's misfortune: Disciplines, interdisciplinarity and the impact of audit culture. *The British Journal of Sociology* 61(4): 639–658.

Khan I, Fraser A, Bray M, et al. (2014) ProtocolNavigator: Emulation-based software for the design, documentation and reproduction biological experiments. *Bioinformatics* 30(23): 3440–3442.

Kitchin R (2014a) Big Data, new epistemologies and paradigm shifts. *Big Data & Society* 1(1): 1–12.

Kitchin R (2014b) *The Data Revolution: Big Data, Open Data, Data Infrastructures and their Consequences*. London: Sage.

Kitchin R and McArdle G (2016) What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society* 3(1). DOI: 10.1177/2053951716631130.

Knorr-Cetina K (1999) *Epistemic Cultures: How the Sciences make Knowledge*. Cambridge, MA: Harvard University Press.

Kriege J (1996) *History of CERN, Volume III. The Years of Consolidation 1966–1980*. Amsterdam: North Holland Publishers.

Krishnamurthy V and Poor V (2014) A tutorial on interactive sensing in social networks. *IEEE Transactions on Computational Social Systems* 1(1): 3–21.

Latour B and Woolgar S (1986) *Laboratory Life: The Construction of Scientific Fact*. Princeton, NJ: Princeton University Press.

Law J and Urry J (2004) Enacting the social. *Economy and Society* 33(3): 390–410.

Lazer D and Radford J (2017) Data ex machina: Introduction to Big Data. *Annual Review of Sociology* 43: 19–39.

Leonelli S (2012) Making sense of data-driven research in the biological and biomedical sciences. *Studies in the History and Philosophy of Biological and Biomedical Sciences* 43: 1–3.

Leonelli S (2013) Global data for local science: Assessing the scale of data infrastructures in biological and biomedical research. *BioSocieties* 8(4): 449–465.

Leonelli S (2016) *Data-centric Biology: A Philosophical Study*. Chicago, IL: Chicago University Press.

Lewis J (2008) *Computing genomic science: Bioinformatics and standardisation in proteomics*. Unpublished doctoral dissertation, Cardiff University.

Lewis J and Bartlett A (2013) Inscribing a discipline: Tensions in the field of bioinformatics. *New Genetics and Society* 32(3): 243–263.

Lewis J, Bartlett A and Atkinson P (2016) Hidden in the middle: Culture, value and reward in bioinformatics. *Minerva* 54(4): 471–490.

Liben-Nowell D and Kleinberg J (2007) The link-prediction problem for social networks. *Journal of the Association for Information Science and Technology* 58(7): 1019–1031.

Liu Y and Xu S (2016) Detecting rumors through modelling information propagation networks in a social media environment. *IEEE Transactions on Computational Social Systems* 3(2): 46–62.

Lynch C (2008) Big data: How do your data grow? *Nature* 455(7209): 28–29.

McFarland DA, Lewis K and Goldberg A (2016) Sociology in the era of big data: The ascent of forensic social science. *The American Sociologist* 47(1): 12–35.

MacKenzie D (2006) *An Engine, not a Camera*. Cambridge, MA: MIT Press.

Merz M and Knorr-Cetina K (1997) Deconstruction in a 'thinking' science: Theoretical physicists at work. *Social Studies of Science* 27(1): 73–111.

Metzler K (2016) "The Big Data rich and the Big Data poor": The new digital divide raises questions about future academic research. *The Impact Blog, London School of Economics and Political Science*. Available at: http://blogs.lse.ac.uk/impactofsocialsciences/2016/11/22/the-big-data-rich-and-the-big-data-poor-the-new-digital-divide-raises-questions-about-future-academic-research/ (accessed 28 November 2016).

Metzler K, Kim DA, Allum N, et al. (2016) *Who is Doing Computational Social Science? Trends in Big Data Research (White paper)*. London: SAGE Publishing.

Michael M (2016) Enacting big futures, little futures: Toward an ecology of futures. *The Sociological Review* 65(3): 509–524.

Murray DE (2014) *Knowledge Machines: Language and Information in a Technological Society*. London: Routledge.

Niederer S and Taudin Chabot R (2015) Deconstructing the cloud: Responses to Big Data phenomena from social sciences, humanities and the arts. *Big Data & Society* 2(2). DOI: 10.1177/2053951715594635.

November J (2012) *Biomedical Computing: Digitizing Life in the US*. Baltimore, MD: The John Hopkins University Press.

Pentland A (2014) *Social Physics: The Lesson from a New Science*. London: Penguin.

Pickering A (1995) *The Mangle of Practice: Time, Agency, and Science*. Chicago, IL: University of Chicago Press.

Pilato G and Maniscalco U (2015) *Soft Sensors for Social Sensing in Cultural Heritage. 2015 Digital Heritage 2*, Granada, Spain, 28 September–2 October 2015. IEEE Publishing, pp. 749–750.

Pinar WF, Reynolds WM, Slattery P, et al. (2008) *Understanding Curriculum: An Introduction to the Study of Historical and Contemporary Curriculum Discourses*. New York, NY: Lang.

Reyes-Galindo L (2011) *The sociology of theoretical physics*. Unpublished doctoral dissertation, Cardiff University.

Reyes-Galindo L (2014) Linking the subcultures of physics: Virtual empiricism and the bonding role of trust. Social *Studies of Science* 44(5): 736–757.

Reyes-Galindo L (2016) Automating the Horae: Boundary-work in the age of computers. *Social Studies of Science* 46(4): 586–606.

Ruppert E (2015) Who owns big data. *Discover Society,* 23. Available at: http://discoversociety.org/2015/07/30/who-owns-big-data/.

Savage M and Burrows R (2007) The coming crisis of empirical sociology. *Sociology* 41(5): 885–899.

Slayton R (2013) *Arguments that Count: Physics, Computing, and Missile Defense, 1949–2012*. Cambridge, MA: MIT Press.

Smith RJ (2014) Missed miracles and mystical connections: Qualitative research, digital social science and big data. In: Hand M and Hillyard S (eds) *Big Data? Qualitative Approaches to Digital Research*. Bradford: Emerald.

Steimle J (2015) Drowning in Big Data – Finding insight in a digital sea of information. *Forbes*, 25 March. Available at: http://www.forbes.com/sites/joshsteimle/2015/03/25/drowning-in-big-data-finding-insight-in-a-digital-sea-of-information/#3e6509a76d90 (accessed 1 October 2016).

Stein L (2008) Bioinformatics: alive and kicking. *Genome Biology* 9(12): 114.

Stephens N and Lewis J (2017) Doing laboratory ethnography: Reflections on method in scientific workplaces. *Qualitative Research* 19(2): 202–216.

Stephens N, Khan I and Errington R (Forthcoming) Analysing The Role of Virtualisation and Visualisation on Interdisciplinary Knowledge Exchange in Stem Cell Research Processes. Palgrave Communications.

Stevens H (2013) *Life Out of Sequence: A Data-driven History of Bioinformatics*. Chicago, IL: Chicago University Press.

Storer NW (1967) The hard sciences and the soft: Some sociological observations. *Bulletin of the Medical Library Association* 55(1): 75–84.

Strasser B (2010) Collecting, comparing and computing sequences: The making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954–1965. *Journal of the History of Biology* 43: 623–660.

Suarez-Diaz E (2010) Making room for new faces: Evolution, genomics and the growth of bioinformatics. *History and Philosophy of the Life Sciences* 32: 65–90.

Symons J and Alvarado R (2016) Can we trust Big Data? Applying philosophy of science to software. *Big Data & Society* 3(2). DOI: 10.1177/2053951716664747.

Thrift N (2005) *Knowing Capitalism*. London: Sage.

Tripp S and Greuber M (2011) *Economic Impact of the Human Genome Project*. Columbus, OH: Battelle Memorial Institute. Available at: http://www.battelle.org/docs/default-document-library/economic_impact_of_the_human_genome_project.pdf.

Ugander J, Backstrom L, Marlow C, et al. (2012) Structural diversity in social contagion. *Proceedings of the National Academy of Sciences* 109(16): 5962–5966.

Wallis JC, Rolando E and Borgman C (2013) If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS One* 8(7): e67332.

Wang Y and Kosinski M (2018) Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology* 114(2): 246–257.

Williams ML, Burnap P and Sloan L (2017) Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology* 51(6): 1149–1168.

Wu X and Zhang X (2016) Automated inference on criminality using face images. arXiv: 1611.04135.

Xu Z, Zhang H, Hu C, et al. (2016) Building knowledge base of urban emergency events based on crowdsourcing of social media. *Concurrency and Computation: Practice and experience* 28: 4038–4052.

Zhang Y and Zhao Y (2015) Astronomy in the Big Data era. *Data Science Journal* 14(11): 1–9.