Science and Public Policy, 44(1), 2017, 65–74
doi: 10.1093/scipol/scw021
Advance Access Publication Date: 22 June 2016
Article

Early social science research about Big Data

Jan Youtie[1,*], Alan L. Porter[2] and Ying Huang[3]

[1]Enterprise Innovation Institute, Georgia Institute of Technology, 75 Fifth Street, NW, Suite 300, Atlanta, GA 30308, USA; [2]School of Public Policy, Georgia Institute of Technology, Atlanta, GA 30332, USA and Search Technology, Norcross GA 30092, USA and [3]Beijing Institute of Technology, School of Management and Economics, Beijing, 100081, China

*Corresponding author. Email: jan.youtie@innovate.gatech.edu

Abstract

Recent emerging technology policies seek to diminish negative impacts while equitably and responsibly accruing and distributing benefits. Social scientists play a role in these policies, but relatively little quantitative research has been undertaken to study how social scientists inform the assessment of emerging technologies. This paper addresses this gap by examining social science research on 'Big Data', an emerging technology of wide interest. This paper analyzes a dataset of fields extracted from 488 social science and humanities papers written about Big Data. Our focus is on understanding the multi-dimensional nature of societal assessment by examining the references upon which these papers draw. We find that eight sub-literatures are important in framing social science research about Big Data. These results indicate that the field is evolving from general sociological considerations toward applications issues and privacy concerns. Implications for science policy and technology assessment of societal implications are discussed.

Key words: Big Data; bibliometrics; cited references; social science.

1. Introduction

Big Data is a recent emerging technology with broad societal implications in areas such as: privacy and security, ability to address business and medical needs, and the potential to exacerbate or lessen inequalities. Big Data is defined on the basis of size, growth, diversity, and analytic capacities. Big Data has attracted increasing attention (Miller 2014), but to what extent has societal assessment kept pace? Concern about societal implications of emerging technologies has been incorporated into human genetics and nanotechnologies policies in recent years, with these efforts commonly involving social science assessments. Relatively little quantitative research has been performed to study how social scientists inform the emergence of new technologies and guide developmental pathways. The objective of the present work is to address this gap by examining the trajectory of these societal implications through an analysis of a set of papers written by social scientists and humanities scholars about Big Data that have been indexed in the Web of Science (WoS). The paper examines the growth and distribution of social science and humanities papers about Big Data. A key contribution of this paper is an examination of the sub-literatures cited in these papers, which offers a window into the evolution of societal concerns about this emerging technology.

2. Theory

The information technology field has long been concerned with technologies that can advance the storage capacity, processing and analytics in working with information. Although the legacy of information technology developments is long, the term 'Big Data' has a more recent history that some trace to a special issue of *Nature* published in September 2008, on the topic, while others allude to earlier or later references. Indeed the term itself has become a 'meme' for developments in the 21st century that facilitate the procurement, storage, processing, and analysis of large-scale information compilations. Boyd and Crawford (2012) call out the 'mythology' of the term, associating it with an overly optimistic and opportunistic rhetoric. The White House (2014) has drawn on the Gartner Inc. definition of Big Data in terms of the three 'Vs' (although more V's have been added in other definitions):

- volume of data collected and processed at a decreasing cost
- variety of data, including digital data and data originating in analog forms that can be digitized (see President's Council of Advisors on Science and Technology 2014)
- velocity of data that can be obtained nearly in real-time

The ability to process more information, more quickly, and with greater ease of analysis opens up opportunities in areas such as: medical, business, scientific research, environmental, defense, and climate change applications (Bryant et al. 2008).

Yet it is not solely a matter of scale and opportunity that defines Big Data. It also matters how these methods are used, the extent to which they are divorced from context, applications that could violate privacy and security expectations, and inequities and unethical consequences that these capabilities could create. In sum, important

societal considerations accompany Big Data development and applications. Such considerations are not solely the purview of Big Data. Observers underscore the importance of societal considerations in emerging technologies ranging from human genomics to nanotechnology to geoengineering.

The rise of Big Data takes place in the context of new attention being paid to emerging technologies (Rotolo et al. 2015). There is growing concern that these technologies advance responsibly. Emerging technologies come into a regulated world marked by regulatory oversight of societal issues such as: health and safety, security, privacy, and other oversight. In addition to this landscape of regulatory 'hard governance', ever greater concerns have appeared about the coordination and orientation of this governance system. In particular, regulatory systems may lag behind developments in these technologies. While there are issues about the timing of when and how much to intervene, a rise of 'soft governance' mechanisms has occurred alongside emerging technology trajectories, including the creation of voluntary codes of conduct (i.e. 'soft laws') (Kearnes and Rip 2009) and non-governmental portals and councils that promote attention to societal research and the implications of this research. These soft governance approaches highlight the need for spaces and methods for negotiation between science and application advances associated with emerging technologies on the one hand, and societal issues on the other (Kuhlmann 2001).

This tension between the development of emerging technologies and their oversight leads to a debate about the process, given that these two positions are rarely settled questions. Technology assessment is one method designed to provide information to examine these tensions. Technology assessment is a meta-level method used to analyze potential development pathways of a technology and the social and economic implications of this development (Porter et al. 1980; Rip and TeKulve 2008). Technology assessment includes: methods to perform empirical analysis of the emerging technology; methods to engage stakeholders, experts, and publics; and methods to assess future pathways (Porter et al. 2004). Technology assessment does not presume to provide accurate predictions of the future. Rather it seeks to reduce the uncertainties that restrict investment in the technology through revealing and, presumably, encouraging attention to negative societal impacts (Hoppe 2002; Robinson et al. 2012). Technology assessment has traditionally been a central government function (as of 1995 the US Congress no longer has an Office of Technology Assessment to study the likely impacts of new technologies, but other US organizations are involved in technology assessment (or quasi-technology assessment) including the National Academies and the General Accountability Office). However, decentralized methods have arisen to obtain more diverse inputs as the technologies are emerging (Guston and Sarewitz 2002; Kuhlmann 2002).

Thus is laid out the role of social scientists and humanists in assessing the societal implications of emerging technologies. At least two rationales have been given for involvement of social scientists in technology assessment. First, some, if not the majority, of the research underlying the development of many emerging technologies is supported by funding from agencies that serve a broader public mission. Bozeman et al. (2015) argue that as a result, at least some of the criteria used to judge a technology's trajectory should be whether or not the technology furthers this 'public value' mission. Bozeman (2007: 37) defines public values as:

> ... providing normative consensus about (1) the rights, benefits, and prerogatives to which citizens should (and should not) be entitled; (2) the obligations of citizens to society, the state and one another; (3) and the principles on which governments and policies should be based.

An example of the public values perspective on emerging technology is provided by Slade (2011), who found that public values promoting more equitable access to effective treatments are widely mentioned in policy documents related to nanomedicine, but less prevalent in particular research funding programs.

A second rationale for this involvement is the potential of societal issues to derail emerging technology developments, as was the case with genetically modified organisms. Instead, societal issues should be anticipated alongside the conduct of scientific R&D (Guston and Sarewitz 2002). Analyses of societal aspects can address concerns about: potential benefits; impacts and risks of candidate commercial products; life-cycle and sustainability aspects; regulatory and governance issues, including reporting requirements and informal codes of conduct and suitable standards; disruptive effects on employment; and engagement of publics (from various sectors, including disadvantaged populations) in determining development pathways. Recent examples of these efforts include: chapters on innovation and responsible governance for the international assessment of nanotechnology research needs (Roco et al. 2011); the Oxford Principles in the UK (Rayner et al. 2013) which address responsible research and development in the context of climate change; the European Commission's three pronged definition of responsible research and innovation consisting of ethical promotion of social justice, sustainable development, and socially desirable quality of life (von Schomberg 2013); the eight principles for sustainable innovation proposed by matterforall.org[1] (Maynard 2015); and the Virtual Institute for Responsible Innovation's 14-country network comprised of researchers and practitioners coordinating responsible innovation activities.[2]

Despite the rise in concern about societal aspects of emerging technologies, surprisingly little research has been published which systematically tracks how this societal research arises. Social science research has been shown to involve multiple and distinct literatures (Hicks 2005) and collaborative networks that are increasingly internationalized depending on the country context (Leydesdorff et al. 2014). One study which looked at social science research about societal issues in nanotechnology found that in the early years, social scientists tended to cite physical scientists' work, while eventually citing their own social science literature in subsequent years in areas such as: science visioning, public participation, innovation economics, scientometrics, governance, and ethics (Shapira et al. 2010). Another study found that social scientists studying an emerging technology do not appear to be aware of each other's work (Shumpert et al. 2014). Yet a third study suggests that social scientists studying synthetic biology are influenced by legacy work from prior human genomics work. Thusly, this work is very strong in bioethics and law, but pays less attention to other societal aspects. Moreover, it does not consider societal research findings from other emerging technologies (Shapira et al. 2015).

This paper examines social science work about Big Data in the context of the literature on technology assessment. Big Data is a rapidly emerging area which, at the time of this writing, is in its early stages of emergence, having grown in prominence since 2012. Despite the recent status of growth of the Big Data 'science' field, we argue that it is not too early to understand the rise of societal concerns in social science and humanities works. Moreover, social scientists have themselves used Big Data methods through mining large-scale datasets as an object for studying the emergence of Big

Data, albeit not without concern about issues with the accuracy of, and what can be interpreted from, the data (Schroeder 2014). We are guided by questions about the evolution of the social science and humanities research and the extent to which other social science and humanities research is cited. We expect that social science research will draw on legacy fields in their own domain (such as works in the sociology of science) and in the information technology domain (such as geographic information systems). Drawing on a dataset of social science papers addressing Big Data, we find that a broader set of knowledge sources is used in these social science articles, including (in addition to the expected areas) works on the societal repercussions of the internet, business performance impacts, law and privacy, medical applications, and analytics and software studies. However, there is scope for social scientists writing about Big Data to pay more attention to other fields in order to gain a wider perspective on current and future developments.

3. Methods

To measure scholarship in the social sciences and humanities relating to Big Data concerns, we performed a search in the combined Social Sciences Citation Index and Arts & Humanities Citation Index of the WoS using variations of the term 'Big Data' (with and without spaces between the two words). Our experience with emerging technologies is that social science discussions focus on the general topic, so that intricate search strategies are not helpful (Shapira et al. 2010), hence our use of the term 'Big Data' and its variations to retrieve social science and humanities papers. The search (done on 19 March 2015) retrieved 488 records for the years 2005–2015. Although other sources index social science and humanities papers, most notably Google Scholar, we have selected WoS because it provides higher quality, consistent information (Gardner 2005) especially about cited references, which are a primary data element of interest in this analysis. Comparing the size of the social science research domain with that of the science domain (Porter et al. 2015), we performed a manual review of the resulting papers, which indicated that all of these papers fell within the domain of interest. Of the 488 papers, more than 70% were articles. Another 17% were classified as 'editorial material'. We did not remove these works because more than half had cited references and most appeared in well-regarded academic journals such as *Nature* and the *Journal of the American Medical Informatics Association*. These latter two journals (which have a traditional science orientation) also produce social science and humanities works, including the articles on Big Data which represent our topic of interest. The rest of the social science and humanities papers consisted largely of review articles, meeting abstracts, and book chapters.

We also note that some of the articles we reference concern medical topics (such as the article by Duncan and Keller (2011) on the use of Big Data, which was published in the *American Journal of Psychiatry* and the paper by Lazer et al. (2014) paper on Google Flu published in *Science*). One could presumably exclude these because they do not appear to be social science or humanities articles. We did not because WoS explicitly classified them as social science articles in the Social Science Citation Index or the Arts and Humanities Citation Index. Hicks (2005) notes that one of the characteristics of social science work is a lack of consensus in defining it, so we take these indexes as a standard for inclusion rather than trying to develop our own definitions of these broad areas.

To understand the nature of this dataset, we indicate that it represents a fast growing domain (see Fig. 1). Although our search extended back to 2005, we did not find a substantial number of social science Big Data articles until 2012. In that year, the number of articles grew by a factor of almost four from the previous year's paper count. The paper count grew by a factor of three from 2012 to 2013 and by a factor of 2.7 from 2013 to 2014. Although our search was performed early in 2015, more than 50 social science Big Data articles were already indexed by WoS. This trajectory suggests the early emergence of societal concerns about Big Data and thus, value in examining it while it is under rapid growth.

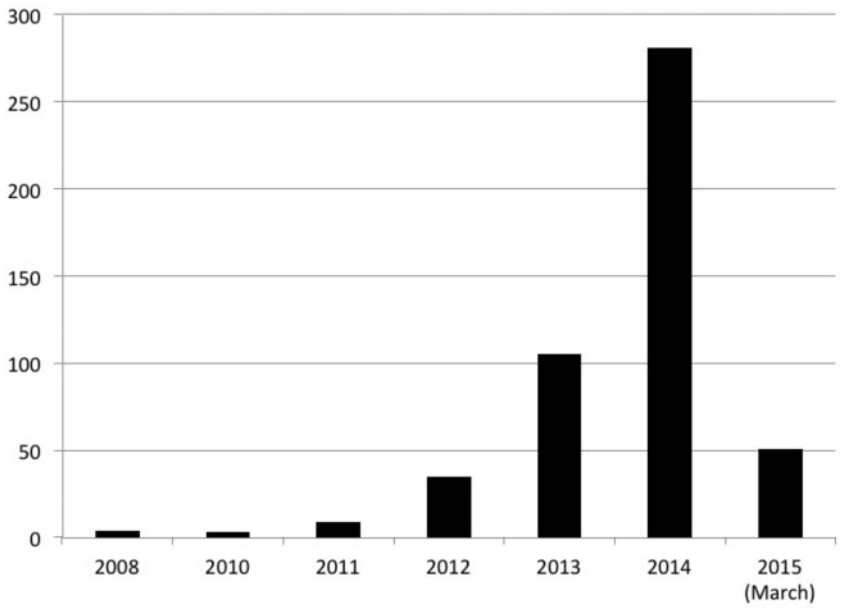

Figure 1. Number of social science publications about Big Data by year (as of March 2015)
488 publications sourced from the WoS

Table 1. Top 20 journals based on number of social science papers about Big Data

Journal	No. of papers
Health Affairs	17
Behavioral and Brain Sciences	11
Harvard Business Review	11
Journal of the American Medical Informatics Association	11
Value in Health	11
International Journal of Communication	10
Computer Law & Security Review	8
EContent	8
Review of Policy Research	8
Forbes	7
Information Communication & Society	7
Journal of Business Logistics	7
Nature	7
PS-Political Science & Politics	7
Scientometrics	6
Decision Support Systems	5
MIT Sloan Management Review	5
Proceedings of the National Academy of Sciences of the United States of America	5
Science	5
Transactions in GIS	5

Source: 488 publications sourced from WoS

US-based authors accounted for 63% of the papers, followed by UK-based authors who account for 18% of the papers. A dramatic drop-off occurred thereafter to 5% each for authors based in Australia and China; 4% each for those based in Germany, South Korea and Canada; and 3% each for those based in the Netherlands, Spain, and Switzerland. The prominence of US- and UK-based authors has been observed in other studies of social scientists writing about emerging technology (Shapira et al. 2015).

Four of the five most common journals in terms of number of Big Data social science papers are in the health area (see Table 1). These journals (with more than 10 Big Data social science papers) are: *Health Affairs*, *Behavioral and Brain Sciences*, *Journal of the American Medical Informatics Association*, and *Value in Health*. The lone non-health journal among the top five is *Harvard Business Review*. Other business and management journals with at least four Big Data social science papers are *Journal of Business Logistics*, *MIT Sloan Management Review*, *MIS Quarterly Executive*, and *Technological Forecasting & Social Change*. The communications area is represented by *International Journal of Communication, Information Communication & Society*, and *Journal of Communication*, each having at least four Big Data social science papers. Big Data social science papers in the information science and computer science areas, with at least four papers, can be found in *EContent*, *Scientometrics*, *Decision Support Systems*, *International Journal of Geographical Information Science*, and *Online*. Journals in the health and business areas also overlap with these WoS categories. In the law area, *Computer Law and Security Review* has eight Big Data social science papers, while in the political science area, *Review of Policy Research* also has eight papers and *PS-Political Science & Politics* has seven. We further note that 16% of the articles concern Big Data health applications, while 15% concern privacy issues.

Because of the newness of the domain, one would not expect that these articles would attract many citations. Indeed that is the case for 64% of the articles, which have no forward citations. However, three articles were cited by more than 50 other papers (see Table 1). The first is about the ability to use Big Data to assess the effects of genes versus the environment published by Duncan and Keller (2011) in the *American Journal of Psychiatry*. The second, one of the earliest papers (which is about the need for appropriate infrastructure for the use of Big Data in science) by Lynch (2008), appears in *Nature*. The third is an introduction to a special issue in *MIS Quarterly* about the use of Big Data and analytics for business intelligence by Chen et al. (2012). Another article, cited by 47 other papers, is a classic social science assessment of societal concerns about Big Data, written by Boyd and Crawford (2012), 'Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon' published in *Information Communication & Society*. Another four articles were cited by 30 or more papers: a *Harvard Business Review* paper written by McAfee and Brynjolfsson (2012) about cases of effective use of Big Data for making improved business decisions; work by Kosinski et al. (2013) on using Facebook Likes to predict behavior in the *Proceedings of the National Academy of Sciences of the United States of America*; a paper by Lavalle et al. (2011) in *MIT Sloan Management Review* that explains the results of an IBM survey on how business executives use Big Data analytics; and a paper by Lazer et al. (2014) in *Science* about errors in Google Flu Trends results (see Table 2).

Only one of these authors has more than five publications in the Big Data social science domain (Michal Kosinski). Stanley Fawcett, John Ioannidis and Matthew Waller each have five publications. Fawcett and Waller's works concern the use of Big Data in business logistics while Ioannidis's work is prominent in the biomedical area. Two more, Thomas Davenport (whose work on Big Data appears in *Harvard Business Review* and *MIT Sloan Management*) and Lucila Ohno-Machado (whose articles are on the subject of bioinformatics) have four publications each (see Fig. 1).

Given these characteristics of the dataset, our analytical focus is on the type of information sources used in these social science publications. Here cited references are taken as a proxy for knowledge bases and flows (Leydesdorff 1998). We posit that cited references (also termed 'backward citations') can be usefully employed to uncover key societal dimensions of an emerging technology including for technology assessment, in this case Big Data. The dataset includes more than 17,500 cited references. Eighty-three percent of the publications have at least one cited reference. The average article has 40 cited references, while the median is 30. Nine articles have 200 or more references; eight of these nine fall into the legal category.

Our dimensional analysis of cited references draws on an approach used in our earlier nanotechnology research (Shapira et al. 2010, 2015). This approach is based on a multi-dimensional scaling (MDS) analysis of co-citations of all authors with more than 10 mentions in these papers. This filter amounts to a focus on 43 authors. Half of the Big Data social science papers cited at least one of these 43 authors. Analysis of these authors' works was performed using VantagePoint desktop text analysis software.[3] VantagePoint examines the associations from a matrix comprised of these cited authors. The nodes in the network drawing represent the number of papers citing the author while the extent to which two authors were cited in the same paper comprises the links. Visualization of the resulting cited reference network only shows the strongest links, using VantagePoint's path erasing algorithm to highlight the most important links based on the proportion where 'elbows' occur in a

Table 2. Most highly cited social science papers about Big Data

Author	Article, journal	Year of publication	Number of citations (as of March 2015)
Duncan, L. E. and Keller, M. C.	A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry, *American Journal of Psychiatry*	2011	176
Lynch, C.	Big Data: How do your data grow?, *Nature*	2008	64
Chen, H., Chiang, R. H. and Storey, V. C.	Business intelligence and analytics: From Big Data to big impact, *MIS Quarterly*	2012	53
Boyd, D. and Crawford, K.	Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon, *Information, Communication & Society*	2012	47
McAfee, A. and Brynjolfsson, E.	Big data: The management revolution, *Harvard Business Review*	2012	37
Kosinski, M., Stillwell, D. and Graepel, T.	Private traits and attributes are predictable from digital records of human behaviour, *Proceedings of the National Academy of Sciences*	2013	36
LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S. and Kruschwitz, N.	Big data, analytics and the path from insights to value, *MIT Sloan Management Review*	2013	34
Lazer, D., Kennedy, R., King, G. and Vespignani, A.	The parable of Google Flu: Traps in Big Data analysis, *Science*	2014	32

Source: 488 publications sourced from WoS

similarity plot. Thus, non-appearance of links does not mean a lack of co-citation, but rather fewer co-citations. MDS reduces the multiple dimensions in this map into two-dimensional space. The proximity of two nodes indicates association, though the exact x–y positioning has little meaning.

The MDS results are used to identify clusters that represent each of the dimensions of knowledge flows. To interpret these clusters, we coded each author into a category based on that author's background. This assignment includes an examination of the author's curriculum vitae or other biographic information, degree, research area, and type of literature in which the author has published. Although assigning authors to categories is not without subjectivity, we believe our assignment approach is reasonable for the purpose that it is intended, which is to help us interpret the map by labeling clusters of nodes. One limitation is that WoS only shows the first cited author (which is the one that we coded), hence our analysis is missing information on the other authors. Despite this limitation, the dimensions do a good job of reflecting our interpretation of the aforementioned author background characteristics.

4. Results

The results of the cluster analysis are shown in Fig. 2. The exploratory disposition of the cluster analysis means that a number of solutions may be considered. For example, a seven-cluster solution could be presented around a natural science and engineering field, but it would group a large share of papers together, thus obscuring potentially useful variations. After considering multiple possibilities, we deem the eight category solution to be the most useful for understanding variation in social science cited references. The eight categories are: Internet, Society, New Media (hereafter Internet & Society), Business Impacts of IT/Management of Technology (hereafter Business Impacts), Big Data & Medicine, Law & Privacy, Internet & Science/Sociology of Science (hereafter Sociology of Science), Analytics/Software, Decision-making, and Geographic Information Systems (GIS).

Some of these clusters are distinct and self-contained, for example the GIS cluster, albeit there are a notable number of common papers that cite Batty (in the GIS cluster) and Newman (in the Analytics/Software cluster). The Law & Privacy cluster also has strong commonalities in papers citing four of its authors (Richards, Solove, Ohm, and Tene). In some cases, there are a sufficient number of papers from two clusters that the author becomes situated at the intersection of these two clusters. For example Viktor Mayer-Schönberger is at the intersection of the Law & Privacy cluster and the Internet & Society cluster. Helen Nissenbaum is situated in the Law & Privacy cluster but is also cited by papers in the Internet & Society cluster. The map includes some interesting two-cluster adjacencies: the Internet & Society and Sociology of Science clusters; the Law & Privacy and the Internet & Society clusters, the Big Data & Medicine and the Business Impacts clusters, and the GIS and Analytics/Software clusters. There also is a chained relationship between the Law & Privacy, Decision-making, Big Data & Medicine, and Business Impacts clusters—suggesting some commonality in the papers citing the authors in these four clusters. In contrast, the Sociology of Science cluster is most distant from the Business Impacts cluster, indicating that these two clusters have the fewest papers in common.

Taking the clusters one by one, Internet & Society is the largest cluster (see Table 3). This cluster comprises 83 papers in which authors in this category were cited. Examples from this cluster are Boyd and Crawford's work about the societal implications of Big Data and Anderson's article in *Wired* about the ability to analyze societal data without a theoretical context (Boyd and Crawford 2012; Anderson 2008). The Business Impacts cluster includes 72 papers which cited authors in this category, exemplified by the McKinsey paper (lead author Manyika) on business innovation through Big Data and the article by Davenport and Patil in *Harvard Business Review* on the business contributions of data scientists (Manyika et al. 2011; Davenport and Patil 2012). Big Data & Medicine represents 67 papers, including the work of Ginsberg et al. (2009) on Google Flu Trends and Christakis and Fowler (2007) on the use of social networks to analyze the proliferation of obesity. The Law & Privacy cluster comprises 64 papers, including the book by Mayer-Schönberger and Cukier *Big Data: A Revolution*

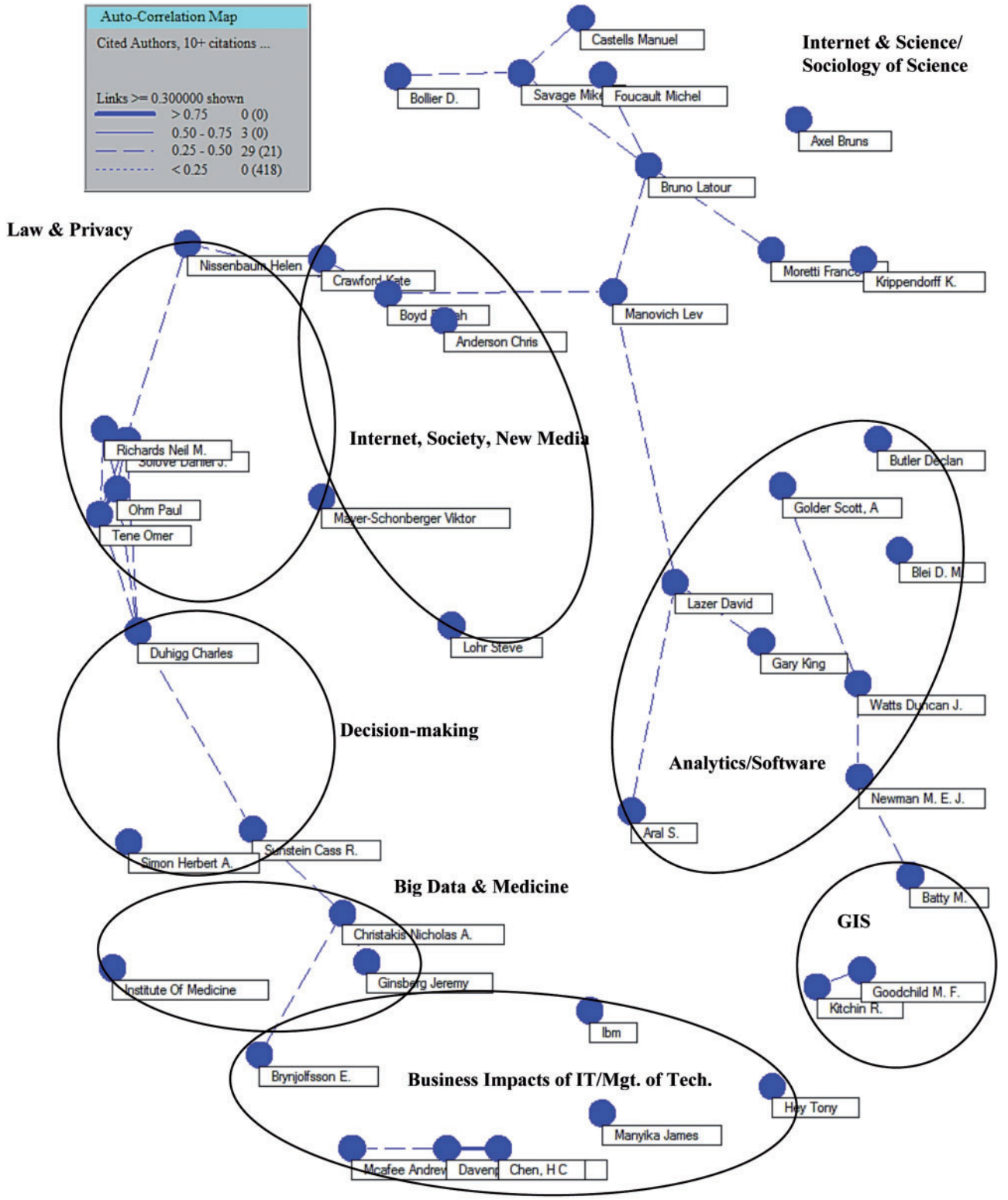

Figure 2. Auto-correlation map of clusters of cited authors receiving 10 or more citations in Big Data social science papers (produced using VantagePoint with circles and labels added by the present author)

That Will Transform How We Live, Work, and Think (2013) about internet life and governance and Solove's paper on the flow of aggregated personal data from information providers to the government (Solove 2002). Sociology of science, representing 63 papers, draws on foundational thinkers such as Bruno Latour and Michel Foucault through works such as those by Manovich (2012) and Savage and Burrows (2007) about how Big Data affects practices in the social sciences. Fifty-six articles fall into the Analytics/Software cluster including the critique by Lazer et al. (2014) of the Google Flu Trends project and analytics issues presented by King (2011) in *Science* which draws in part

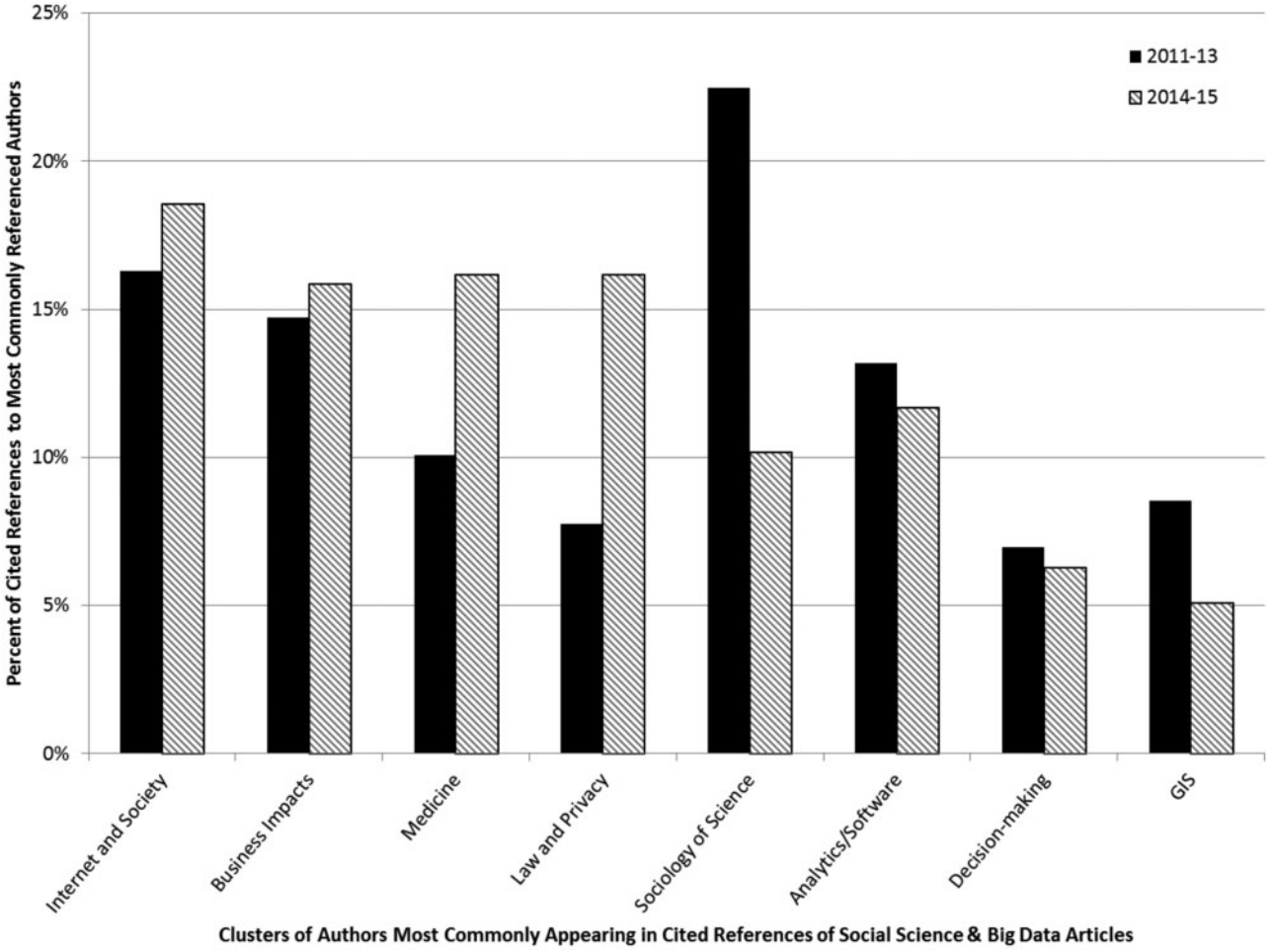

Figure 3. Percentage of Big Data social science papers most commonly referencing clusters in two time periods (2011–2013 and 2015–2015 (part year))

on his work with nonstandard data types such as Twitter (and the Crimson Hexagon analytic tool developed in his research group). Popular books on how Big Data can be used to improve behaviors and decision-making form the foundation as cited references in 30 papers: Duhigg (2012) *The Power of Habit*, Thaler and Sunstein (2008) *Nudge*, and the work of Herbert Simon, a Nobel Prize laureate, on topics such as information overload. Twenty-eight papers cite references concerning geography, including the paper by Goodchild (2007) in *GeoJournal* about how devices carried by individuals create analyzable geographic information that raises concerns about the accuracy of the information and threats to individual privacy and the book *Code/Space* by Kitchin and Dodge (2011) which explores the interaction of digital devices with physical lifestyle.

We tested this manual background-based coding with the clustering algorithm in VOSviewer (see Appendix). The test indicates that the two clustering methods produce fairly consistent results. However, we observe some differences. Overlap is observed in the VOSviewer density cluster maps among the Business Impacts, Big Data & Medicine, and Decision-making clusters; between the Sociology of Science and Internet & Society clusters; and between the GIS and Sociology of Science clusters. The VOSviewer algorithm also split the Analytics/Software cluster into two, albeit these clusters are adjacent. The algorithm did the same for the Internet & Society cluster.

We sought to understand changes in the distribution of clusters over time (see Fig. 3). Because of the small numbers of papers in the earlier years, we group the years into two periods. These two periods are unequal in terms of inclusion of number of years, but represent two breaks in the growth trajectory that suggest different pathways of emergence: 2011–2013 represents the earliest growth period from a small base, while 2014–2015 (part year) represents rapid takeoff. The time divisions suggest that the distribution of references to these clusters has changed over time. In the early period (2011–2013), Big Data social science papers were most apt to reference Sociology of Science papers (23%), followed by Internet & Society papers (16%), Business Impacts papers (15%), and Analytics/Software papers (13%). In the later period (2014–2015), Big Data social science papers were most apt to reference Internet & Society (18%) closely followed (at 16% each) by Law & Privacy, Big Data & Medicine, and Business Impacts papers. The percentage of papers citing Law & Privacy works rose from 8% in the earlier period to 16% in the more recent period and those citing Big Data & Medicine grew from 10% in the earlier period to 16% in the more recent period. In contrast, comparatively less emphasis was given to Sociology of Science papers (from 23% in the earlier period to 10% in the recent period) and GIS papers (from 9% of papers in the earlier period to 5% in the recent period). References to these two clusters of paper increased between the two periods but at a slower rate. Sociology of Science references grew by 17% between the two periods and GIS papers by 55% compared to Internet & Society (which doubled), Big Data & Medicine (which more than

Table 3. Clusters of cited authors in Big Data social science papers and reference examples

Cluster	Citing papers	Top cited references in cluster
Internet and Society	83	Boyd, D. and Crawford, K. (2012) Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15: 662–79
		Anderson, C. (2008) The end of theory. *Wired Magazine*, 16(7), 16–07
Business Impacts	72	Manyika, J., Chui, M., Brown, B., Bughin, J. et al. (2011) Big data: The next frontier for innovation, competition, and productivity
		Davenport, T. H. and Patil, D. J. (2012) Data scientist. *Harvard Business Review*, 90: 70–6
Big Data & Medicine	67	Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. and Brilliant, L. (2009) Detecting influenza epidemics using search engine query data. *Nature*, 457(7232), 1012–14
		Christakis, N. A. and Fowler, J. H. (2007) The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357(4): 370–9
Law & Privacy	64	Mayer-Schönberger, V. andCukier, K. (2013) *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt
		Solove, D. J. (2002). Digital dossiers and the dissipation of fourth amendment privacy. *Southern California Law Review*, *75*
Sociology & Science	63	Manovich, L. (2012) *Trending: The Promises and the Challenges of Big Social Data, Debates in the Digital Humanities*, M. K. Gold (ed.). University of Minnesota Press
		Savage, M. and Burrows, R. (2007) The coming crisis of empirical sociology. *Sociology*, 41: 885–99.
Analytics/Software	56	Lazer, D. M., Kennedy, R., King, G. and Vespignani, A. (2014) The parable of Google Flu: Traps in big data analysis. 343(6176): 1203–5
		King, G. (2011) Ensuring the data-rich future of the social sciences. *Science*, 331(6018): 719–21
Decision-making	30	Duhigg, C. (2012) *The Power of Habit: Why We Do What We Do In Life and Business*. Random House
		Sunstein, C. R. (2011) Empirically informed regulation. *University of Chicago Law Review*, 1349–429
		Thaler, R. H. and Sunstein, C. R. (2008) *Nudge*. Yale University Press
GIS	28	Goodchild, M. F. (2007) Citizens as sensors: The world of volunteered geography. *GeoJournal*, 69: 211–21
		Kitchin, R. and Dodge, M. (2011) *Code/Space: Software and Everyday Life*. MIT Press

Source: 488 publications sourced from WoS

tripled), and Law & Privacy (which more than quadrupled). One interpretation of this finding is that Big Data social science papers are spreading from foundational social science and information systems works to increasingly address application and privacy issues.

5. Discussion and conclusions

This paper proposes an empirically-based method for understanding societal dimensions in the assessment of emerging technologies. The focus of the method is on the knowledge sources that social scientists studying Big Data cite. The results indicate that societal assessment of Big Data draws on multiple dimensions. Some of the citations involve foundational sociology of science works, while others involve legacy research in the GIS area. The emphasis these dimensions have change over time, with social scientists increasingly drawing out from these foundational fields to address Internet & Society, Law & Privacy, Big Data & Medicine, and Business Impacts papers.

Social scientists conducting research about other emerging technologies cite papers that fall into similar clusters to those which this study of Big Data social science papers has brought to light. For example, Shapira et al. (2010), who examined social science papers on the subject of nanotechnology, also identified cited references about sociology/philosophy/history of science. In addition, they identified measurement efforts by scientometrics specialists and economists that are analogous to the Business Impacts cluster in the present paper. Another work using the same method to examine social scientists conducting research about synthetic biology similarly uncovers a sociology/philosophy/history of science cluster as well as a Law cluster (Shapira et al. 2015). Although there are several common knowledge bases used by social scientists studying these technologies, these studies also identified knowledge sources that are not much used as reference bases for social scientists studying Big Data. For example, social scientists studying nanotechnology draw on additional knowledge clusters concerning future science visions, ethics, public perception and deliberation, and the work of a prominent individual policy entrepreneur.

These knowledge base gaps suggest that there may be opportunities for social scientists to make use of other literatures in investigating Big Data societal questions. It is possible that some of these topics are partially covered in existing knowledge sources such as in the Law & Privacy or Internet & Society clusters. Nevertheless, it is also possible that additional research into ethics, public perception and deliberation, and science visions could open up a broader consideration of societal impacts. More focused efforts to engage with these external perspectives could encourage a wider-ranging base from which to obtain insights into the societal implications of Big Data today and in the future.

One limitation of our use of social science data to uncover dimensions of societal implications is that a 'Western bias' occurs in some social science fields, including an English language bias (Leydesdorff et al. 2014). This limitation suggests that efforts to generalize its findings to other national contexts may be incomplete. Co-authorships with US- and/or UK-based researchers may be one way to reach investigators in other countries that seek to examine social science issues in the rollout of emerging technologies.

Our work ultimately informs an emerging science policy debate about how best to organize technology assessment of societal implications to achieve more responsible research and innovation. Should it be incorporated into a central, science-based initiative or should it take a decentralized approach dedicated to assessing societal implications (Calvert and Martin 2009)? The centralized approach presumes that there is a single social scientist, or small number of them

(often ethicists) who represents societal implications. This approach has merit in terms of close integration of scientists and the single social scientist. That said, our results suggest that societal implications are highly multi-dimensional. It would be difficult for any single person to represent them. A decentralized approach that involves multiple social science perspectives and knowledge sources seems best to obtain a more complete assessment of the societal implications of Big Data and other emerging technologies.

Funding

This research is conducted with support from the US National Science Foundation (NSF) Award # 1527370, "Forecasting Innovation Pathways of Big Data & Analytics." The findings and observations contained in this paper are those of the authors and do not necessarily reflect the views of the NSF.

Notes

1. See <matterforall.org> accessed 12 Jan 2016.
2. See <https://cns.asu.edu/viri> accessed 12 Jan 2016.
3. See <www.theVantagePoint.com> accessed 12 Jan 2016.

References

Anderson, C. (2008) 'The end of theory'. *Wired Magazine*, 16(7): 16–07.

Boyd, D. and Crawford, K. (2012) 'Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon'. *Information, Communication and Society*, 15: 662–79.

Bozeman, B. (2007) *Public Values and Public Interest: Counterbalancing Economic Individualism*. Washington, DC: Georgetown University Press.

——, Rimes, H. and Youtie, J. (2015) 'The evolving state-of-the-art in technology transfer research: Revisiting the contingent effectiveness model'. *Research Policy*, 44: 34–49.

Bryant, R., Katz, R. H. and Lazowska, E. D. (2008) 'Big-data computing: Creating revolutionary breakthroughs in commerce, science and society'. <http://www.datascienceassn.org/sites/default/files/Big%20Data%20Computing%202008%20Paper.pdf> accessed 12 Jan 2016.

Calvert, J. and Martin, P. (2009) 'The role of social scientists in synthetic biology'. *EMBO reports*, 10: 201–4.

Chen, H., Chiang, R. H. and Storey, V. C. (2012) 'Business intelligence and analytics: From big data to big impact'. *MIS Quarterly*, 36: 1165–88.

Christakis, N. A. and Fowler, J. H. (2007) 'The spread of obesity in a large social network over 32 years'. *New England Journal of Medicine*, 357: 370–9.

Davenport, T. H. and Patil, D. J. (2012) 'Data scientist'. *Harvard Business Review*, 90: 70–6.

Duhigg, C. (2012) *The Power of Habit: Why We Do What We Do in Life and Business*. New York: Random House.

Duncan, L. E. and Keller, M. C. (2011) 'A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry'. *American Journal of Psychiatry*, 168: 1041–9.

Gardner, S. and Eng, S. (2005) 'Gaga over Google? Scholar in the social sciences'. *Library Hi Tech News*, 22: 42–5.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L. et al. (2009) 'Detecting influenza epidemics using search engine query data'. *Nature*, 457(7232): 1012–14.

Guston, D. H. and Sarewitz, D. (2002) 'Real-time technology assessment'. *Technology in Society*, 24: 93–109.

Goodchild, M. F. (2007) 'Citizens as sensors: The world of volunteered geography'. *GeoJournal*, 69: 211–21.

Hicks, D. (2005) 'The four literatures of social science' in *Handbook of Quantitative Social Science and Technology Research*, H. Moed, W. Glänzel and U. Schmoch (eds), pp. 473–496. Dordrecht, the Netherlands: Springer.

Hoppe, H. C. (2002) 'The timing of new technology adoption: Theoretical models and empirical evidence'. *The Manchester School*, 70: 56–76.

Kearnes, M. and Rip, A. (2009) 'The emerging governance landscape of nanotechnology' in *Jenseits von Regulierung: Zum politischen Umgang mit der Nanotechnologie*. S. Gammel, A. Lösch and A. Nordmann (eds), pp. 97–121. Heidelberg, Germany: Aka Verlag.

King, G. (2011) 'Ensuring the data-rich future of the social sciences'. *Science*, 331(6018): 719–21.

Kitchin, R. and Dodge, M. (2011) *Code/Space: Software and Everyday Life*. Cambridge, MA: MIT Press.

Kosinski, M., Stillwell, D. and Graepel, T. (2013) 'Private traits and attributes are predictable from digital records of human behavior'. *Proceedings of the National Academy of Sciences*, 110: 5802–5.

Kuhlmann, S. (2001) 'Future governance of innovation policy in Europe—three scenarios'. *Research Policy*, 30: 953–76.

—— (2002) 'Distributed techno-economic intelligence for policymaking. RTD Evaluation Toolbox. Assessing the socio-economic impact of RTD-policies', pp. 210–7. Seville, Spain: European Commission-Joint Research Centre.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S. and Kruschwitz, N. (2011) 'Big data, analytics and the path from insights to value'. *MIT Sloan Management Review*, 52: 21.

Lazer, D., Kennedy, R., King, G. and Vespignani, A. (2014) 'The parable of Google Flu: Traps in big data analysis'. *Science*, 343(6176): 1203–5.

Leydesdorff, L. (1998) 'Theories of citation?'. *Scientometrics*, 43(1): 5–25.

——, Park, H. W. and Wagner, C. (2014) 'International coauthorship relations in the Social Sciences Citation Index: Is internationalization leading the network?'. *Journal of the Association for Information Science and Technology*, 65: 2111–26.

Lynch, C. (2008) 'Big data: How do your data grow?'. *Nature*, 455(7209): 28–9.

McAfee, A. and Brynjolfsson, E. (2012) 'Big data: The management revolution'. *Harvard Business Review*, (90): 60–6.

Manovich, L. (2012) 'Trending: The promises and the challenges of big social data' in *Debates in the Digital Humanities*, M. K. Gold (ed.), pp. 460–475. Minneapolis, MN: University of Minnesota Press.

Manyika, J., Chui, M., Brown, B., Bughin, J. et al. (2011) 'Big data: The next frontier for innovation, competition, and productivity'. <http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation> accessed 12 Jan 2016.

Mayer-Schönberger, V. and Cukier, K. (2013) *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston, MA: Houghton Mifflin Harcourt.

Maynard, A. D. (2015) 'The (nano) entrepreneur's dilemma'. *Nature Nanotechnology*, 10: 199–200.

Miller, R. (2014) 'If you think Big Data's big now, just wait'. *USA Today*, 2 October 2014.

Park, H. W. and Leydesdorff, L. (2013) 'Decomposing social and semantic networks in emerging "big data" research'. *Journal of Informetrics*, 7: 756–65.

Porter, A. L., Ashton, B., Clar, G., Coates, J. et al. (2004) 'Technology futures analysis: Toward integration of the field and new methods'. *Technological Forecasting & Social Change*, 71: 287–303.

——, Huang, Y. and Schuehle, J. (2015) 'MetaData: BigData research evolving across disciplines, players, and topics'. Paper presented at the IEEE International Conference on Big Data, held 29 October –1 November 2015, Santa Clara, CA.

——, Rossini, F. A., Carpenter, S. R. and Roper, A. T. (1980) *A Guidebook for Technology Assessment and Impact Analysis*. New York: North Holland.

President's Council of Advisors on Science & Technology (2014) 'Big Data and Privacy: A Technological Perspective'. Washington, DC: The White House, 1 May 2014.

Rayner, S., Heyward, C., Kruger, T., Pidgeon, N. et al. (2013) 'The Oxford principles'. *Climatic Change*, 121: 499–512.

Rip, A. and TeKulve, H. (2008) 'Constructive Technology Assessment and Socio-Technical Scenarios' in . *The Yearbook of Nanotechnology in Society, Volume 1: Presenting Futures*. E. Fisher, C. Selin and J. M. Wetmore (eds), pp. 49–70. Berlin: Springer.

Robinson, D. K. R., Le Masson, P. and Weil, B. (2012) 'Waiting games: Innovation impasses in situations of high uncertainty'. *Technology Analysis and Strategic Management*, 24: 543–7.

Roco, M. C., Harthorn, B., Guston, D. and Shapira, P. (2011) 'Innovative and responsible governance of nanotechnology for societal development'. *Journal of Nanoparticle Research*, 13: 3557–90.

Rotolo, D., Hicks, D. and Martin, B. (2015) 'What is an emerging technology?'. *Research Policy*, 44: 1827–43.

Savage, M. and Burrows, R. (2007) 'The coming crisis of empirical sociology'. *Sociology*, 41: 885–99.

Schroeder, R. (2014) 'Big Data: Towards a more scientific social science and humanities' in *Society and the Internet: How Networks of Information are Changing our Lives*, M. Graham and W. H. Dutton (eds), pp. 163–76. Oxford, UK: OUP.

Shapira, P., Youtie, J., Porter, A. L. (2010) 'The emergence of social science research in nanotechnology'. *Scientometrics*, 85: 595–611.

———, ——— and Li, Y. (2015) 'Social science contributions compared in synthetic biology and nanotechnology'. *Journal of Responsible Innovation*, 2: 143–8.

Shumpert, B. L., Wolfe, A. K., Bjornstad, D. J., Wang, S. and Campa, M. F. (2014) 'Specificity and engagement: Increasing ELSI's relevance to nano–scientists'. *NanoEthics*, 8: 193–200.

Slade, C. P. (2011) 'Exploring societal impact of nanomedicine using public value mapping' in *Nanotechnology and the Challenges of Equity, Equality and Development*, S. Cozzens and J. Wetmore (eds), pp. 69–88. Berlin: Springer.

Solove, D. J. (2002) 'Digital dossiers and the dissipation of fourth amendment privacy'. *Southern California Law Review*, 75: 1084–1168.

Sunstein, C. R. (2011) 'Empirically informed regulation'. *University of Chicago Law Review*, 78: 1349–429.

Thaler, R. H. and Sunstein, C. R. (2008) *Nudge*. New Haven, CT: Yale University Press.

von Schomberg, R. (2013) 'A vision of responsible research and innovation' in *Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society*, R. Owen, J. Bessant and M. Heintz (eds), pp. 51–74. Hoboken, NJ: Wiley.

White House (2014) 'Big Data: Seizing opportunities, preserving values'. Washington, DC: Executive Office of the President.

Appendix

Density cluster map of cited authors appearing in 10 or more Big Data social science cited references using VOSviewer