EUROPEAN JOURNAL OF Cultural Studies

Article

Sharing, knowledge management and big data: A partial genealogy of the data scientist

European Journal of Cultural Studies
2015, Vol. 18(4-5) 413–428

DOI: 10.1177/1367549415577385
ecs.sagepub.com
SAGE

Robert W Gehl
The University of Utah, USA

Abstract

This article is a partial genealogy of the data scientist, meant as a contribution to understanding how both big data and the subject who mines it have come to be. It adds to the growing criticism of data mining by considering how big data might be used to manage the very workers who ostensibly command it. The article traces the concept of 'sharing' as it appears in discourses about the knowledge economy, arguing that knowledge sharing produces messy excesses of data. It then traces what is not shared: the knowledge workers capable of mining that data to produce value. It concludes by tracing how the act of sharing knowledge is used to undermine the power of the very subject called forth to command the excesses of sharing. It concludes by describing a reversal: data will become scarce while the ability to mine it ubiquitous and cheap.

Keywords

Big data, data scientist, genealogy, knowledge management, sharing

> If you are looking for a career where your services will be in high demand, you should find something where you provide a scarce, complementary service to something that is getting ubiquitous and cheap. So what's getting ubiquitous and cheap? Data. And what is complementary to data? Analysis.
>
> – Hal Varian, Google Chief Economist

Corresponding author:
Robert W Gehl, Department of Communication, The University of Utah, Salt Lake City, UT 84112, USA.
Email: robert.gehl@utah.edu

This article is a partial genealogy of the data scientist, exploring how both big data and the subject who mines it have come to be, how they interact and how they are managed. My goal is to add to the growing criticism of data mining by considering how big data might be used to control the very workers who ostensibly command it. This is thus meant to complicate the growing (and needed) narrative that data mining, analysis and resulting practices are aimed at 'ordinary' people to manage and exploit their desires, movements and sociality within contemporary informational capitalism. Big data can also be used to manage the very data scientists who aid states and corporations in shaping our software-mediated interactions.

This essay traces one key historical thread out from the complex, emerging assemblage of 'big data': the discourse of sharing as it appears in the knowledge economy. It begins with the advent of the concept of the knowledge economy in the early 1960s, followed by the rise of the field of 'knowledge management' in the 1990s. A key aspect of the knowledge economy is the centrality of sharing to knowledge production. Sharing knowledge is desirable because, as we are told by managers, pundits and social media moguls alike, knowledge has the ability of being non-rival and growing infinitely. The more knowledge is shared, the more economies will grow and innovate, and the more fulfilling life will be.

However, following Colin Koopman's (2013) reading of Foucauldian genealogy, a proper genealogy requires more than one thread. To learn more about sharing, especially the particular form of sharing that appears in the knowledge economy, I pick up a second thread of this genealogy: the production of the subject of knowledge management. This subject is the 'knowledge worker', described in the knowledge economy literature. The knowledge worker is called forth to produce value from previously shared knowledge. While sharing is meant to make knowledge commonplace, distributed and digitized, hence growing the knowledge economy, the knowledge worker capable of mining it for value is presented as rare. In other words, while knowledge might be easily shared, firms will not share the labor used to mine it into value.

These threads meet in contemporary big data. On one side, we see the fruits of the knowledge economy's imperative to share: excessive, messy piles of data living in far-flung databases coming from heterogeneous sources, and the unstructured thoughts of people as expressed in likes, tweets, clicks, text, video, images, movements through space, documents, best practices and stories. On the other side, we see fleeting glimpses of the rare subject capable of mining these messes: the Data Scientist, armed with Hadoop,[1] a large pile of data, algorithms and not a little genius. As with past generations of knowledge workers, the data scientist is called forth to tame the excesses of our constant sharing and mine it for new knowledge and produce valuable new techniques of social management. This subject, we are told, will be well-rewarded; in the United States, for example, the Bureau of Labor Statistics reports data scientists earn double the median income (Royster, 2013, p. 7).

However, in tracing these threads, paying careful attention to the field of knowledge management, and doing a history of the present, I ultimately argue we can also see a future where there will be a reversal, a time when data become closed and rare and the ability to mine it as common as unpaid internships. In other words, we will see a time when sharing – the very practice that partially gives rise to big data – will be folded back

onto the latest knowledge subject, dissolving that subject, making that subject disposable. The production of a glut of data scientists, the abstraction of their skills and the automation of data mining techniques could very well arise through sharing mandated by the manuals of knowledge management. In contemporary capitalism, knowledge workers remain *workers*, which is to say that they are subject to the same processes of exploitation, alienation and management by capitalists as their 'less-skilled' counterparts. This has happened in the knowledge economy before. The data show it can happen again.

What follows is a partial genealogical tracing of the data scientist. I say 'partial' because I focus on only two of the many practices that help create the conditions of possibility for the data scientist; there are of course many more threads that could be considered. However, tracing both knowledge sharing and the production of the knowledge worker reveals some of the many reasons why our contemporary problem of big data is fraught with tensions and frictions between knowledge sharing and the labor of knowledge work, tensions that play out as specific forms of 'knowledge work' are defined, valorized, abstracted or dissolved.

Thread 1: sharing knowledge

The discourse of 'sharing' is a powerful, if overlooked, concept in data mining discourse. Overlooked concepts are key to genealogy: Michael Mahon (1992) notes, 'The task [for the genealogist] is to dredge up forgotten documents, minor statements, apparently insignificant details, in order to recreate the forgotten historical and practical conditions of our present existence' (p. 9). One such 'minor statement' comes from a response to a question at the very end of an academic presentation at the University of Utah. James Fowler, co-author of *Connected* (Christakis and Fowler, 2009), built up to an enthusiastic crescendo after being asked about the future of social network analysis:

> Nobody wants to share their personal, social data. Nobody wants to share their medical data. They sure as heck don't want to do both at the same time … But, there's a lot of benefits potentially available from doing that merge, not just in terms of research, but eventually things that you could offer people in terms of improving their health or improving their social life. It's coming. This decade's going to be a fantastic decade for [social network analysis]. (Fowler, 2014)

Here is one of the key personages of big data, a recipient of international attention for showing how ideas and practices can be disseminated through our social networks, calling for more sharing. Merging health and social datasets, Fowler argues, would lead to improvements in health and social well-being – not to mention it would help his emerging field of social network analysis. Here is an actor attempting to establish himself and his field as one of Callon's (1986) 'obligatory points of passage' (p. 205). In doing so, he is problematizing (Koopman, 2013, p. 179), positing that there is a problem: we do not share enough, and this is retarding the growth of new knowledge. Sharing – an alliance (Callon, 1986, p. 206) made between different datasets, techniques, institutions and individuals by and through their data – would lead to many benefits. The dataset would get bigger, potentially reaching the magical 'N = all', the transcendent hero of Mayer-Schönberger and Cukier's (2013) influential book *Big Data*. And Fowler is presenting his ilk, data scientists, as the subjects capable of helping society – but only after we agree to the problematization proposed.

This call for sharing has a unique genealogy with ties to discourse about the 'knowledge economy'. This discourse first emerged in Fritz Machlup's (1962) *The Production and Distribution of Knowledge in the United States*. Machlup synthesized epistemological philosophy, cybernetics, economics of information and national accounting into a new theory of knowledge for the information age: knowledge is not knowledge unless it is communicated, or shared (Godin, 2010, p. 261). '"Producing" knowledge [involves] not only discovering, inventing, designing, and planning but also *disseminating and communicating*' (Machlup, 1962, p. 7 my emphasis). As he explains,

> If only one person has a particular piece of knowledge and does not share it with anybody, it may be that no one knows 'about it'. We do not ordinarily take notice of knowledge possessed by only one knower. Only when he [sic] discloses what had been a 'one-man secret' and thus does his part in the production of a state of knowing, in other minds, what he alone has known, will one usually speak of 'socially new knowledge'. (Machlup, 1962, p. 14)

Against an older, Romantic theory that held knowledge is objectively generated through individual inquiry, Machlup argued for knowledge as a problem of communication and transmission of subjectively held ideas. Communicated beliefs, not the laboratory, become the site of knowledge production. This reconceptualization of knowledge prompted Machlup to focus on education, research and development, communication and information systems. All these, Machlup argued, were key components of the 'knowledge economy', a term he is credited with coining (e.g. Drucker, 1992, p. 263). Measured in terms of communications systems, political economies take on a radically new shape, away from production of goods to production of ideas, away from manufacturing to services.

How can we grow this knowledge economy? To do so, we must share and communicate. As sharing scholar Nicholas John notes, this draws on a more recent conceptualization of sharing as a 'communicative' practice in which what we might call 'non-rival' feelings and ideas are shared among people (John, 2013). Such sharing, being non-rival, has the capacity to grow indefinitely – quite attractive in capitalism. In the end, however, while Machlup's theory did much to add new categories to the econometric calculation of gross domestic production (Godin, 2010), it did little to foster the sharing he theorized was central to the knowledge economy.

However, in the 1990s, a new mode of management sought to tackle the sharing problem with new theories and practices. This was 'knowledge management'. An illustrative document in this field is Stephen Denning's *The Springboard* (Denning, 2001; see also King and McGrath, 2004, p. 36). Tasked by his employer, the World Bank, to implement knowledge management, Denning argued that the best method of doing so was through 'knowledge sharing'. *The Springboard* is the story of his attempt to convince executives, managers, domain experts and entry-level bank employees that sharing knowledge (especially in the form of happy little stories) would lead to 'organizational change' in the form of greater efficiency and more customer satisfaction. Denning urged his employees to share 'best practices', mistakes, expertise and new ideas freely throughout the organization and among its clients. Indeed, thanks, in part, to Denning's example, the problem of knowledge sharing is a central concern in the knowledge management

literature (Bock and Kim, 2002; Hall, 2001; Horibe, 1999; Husted and Michailova, 2002; Lee, 2001; Liu and Liu, 2011; Santos et al., 2012).

Because knowledge management arises alongside the popularization of digital technologies, it is no surprise that knowledge management theorists argued for the use of information technologies to foster sharing (e.g. Hall, 2001; Stoddart, 2001). A side-effect of this is the *capture* of knowledge as information, concretizing the 'fuzzy objects' (John, 2013) of communicative sharing (thoughts, feelings, perceptions, etc.) in digital forms such as documents, best practices and databases. In the knowledge management literature, knowledge might appear first as tacit knowledge existing 'between your employee's ears', but it can be *explicated* by directing its flow through digital communications channels such as intranets, groupware, email or documentation (e.g. Davenport and Prusak, 1998, p. 70; O'Dell et al., 1998, pp. 86–87; Santos et al., 2012, p. 30; Tiwana, 2000, p. 45). Knowledge management is thus partially about the production of masses of data qua informationalized knowledge sharing. Indeed, during the 1990s, as knowledge management scholars constructed their field, their colleagues in information technology were busily constructing 'knowledge warehouses' and standardized relational databases (e.g. Offsey, 1997; O'Leary, 1998). It is notable, then, that Denning's storytelling thesis does not end with mere sharing among employees, but the *banking* of such knowledge, famously making the World Bank into the 'Knowledge Bank' (Mehta, 2001). Such 'banked' (or 'warehoused') knowledge is made possible only after knowledge is shared in particular, digitizable ways.

With such knowledge concretized and informationalized, a new call for sharing appears, especially after the terrorist attacks of 9/11: inter-institutional sharing. A central example of this is the call for intelligence agencies to share information and 'connect the dots' of terrorist networks and activities (Chen et al., 2008). Monahan, Palmer and Regan's studies of Department of Homeland Security 'Fusion Centers' illustrate this (Monahan and Palmer, 2009; Regan and Monahan, 2013). These 'centers of concatenation' (Monahan and Regan, 2011) were intended to bring US federal and local law enforcement officers – and their databases – in close proximity to one another to increase sharing among agencies and help them coordinate the war on terror and crime. But this form of inter-institutional sharing also appears in more benign forms, such as when firms linked together in a supply chain share data (Bell et al., 2002). Indeed, this is the sort of sharing James Fowler desires when he calls for the sharing of health and social data; Fowler wants firms (such as wearable computer company FitBit and Facebook) to allow researchers to merge their datasets (similarly, see Mayer-Schönberger and Cukier, 2013, p. 138). In a sense, these institutional forms of sharing are macro-level repetitions of micro-level individual sharing. Sharing means growing knowledge and thus the knowledge economy.

Of course, we cannot overlook a more recent phenomenon, sharing via social media (John, 2013). Facebook's 'Share' button is emblematic of this. As Mark Zuckerberg (2009) puts it,

> As people share more, the timeline gets filled in more and more with what is happening with everything you're connected to. The pace of updates accelerates. This creates a continuous stream of information that delivers a deeper understanding for everyone participating in it. As

this happens, people will no longer come to Facebook to consume a particular piece or type of content, but to consume and participate in the stream itself.

Here is the logic of Facebook and many other forms of social media: a call for users to share digitized details of their lives in a social stream, reaching a point of 'deeper understanding' through constant sharing. More importantly, variations in the phrase 'you agree to share' are central to social media sites' Terms of Service agreements, the moment when users agree to transmit their ideas to social media server farms (Gehl, 2014b, p. 65). While social media are new, as should be clear in this genealogy, the social media practice of sharing has its roots in the half-century history of the knowledge economy.

Indeed, sharing is so central to the knowledge economy that it is taken for granted. Knowledge becomes a non-rivalrous, measurable, disembodied thing that both moves from person to person or institution to institution, but also can be duplicated across those entities ad infinitum. In other words, of course, this is a conception of knowledge as information. The informationalization of knowledge, coupled with advances in digital storage and the concretization of so many utterances – from the corporate world to anti-terrorist agencies to the world of digitized online interaction – leads to our current *epistemé*, that of big data, the dream of N = all, wherein every social problem can be solved. If only we keep sharing.

Thread 2: knowledge workers

However, the need for more sharing to grow the knowledge economy presented a problem: how to produce *value* from all that informationalized, shared knowledge. The solution proposed to this problem is the production of a unique subject, the knowledge worker. This is the second thread of this genealogy: the articulation of a subject capable of negotiating the fruits of knowledge sharing. Tracing this articulation allows us a chance to observe a chemical reaction with the other thread, knowledge sharing, allowing us to better know our contemporary data scientists.

Again, the rise of knowledge management in the 1990s is a key moment in this history, because this field creates the subject it seeks to manage. As knowledge management theorists Davenport and Prusak (1998) note, knowledge workers take data and information and convert them into knowledge. They do so by applying their skills, experiences and judgments to the glut of information available to produce not just more information, but 'knowledge in action'. This is a valuable commodity: 'one of the reasons we find knowledge valuable is that it is close – and closer than data or information – to action. Knowledge can and should be evaluated by the decisions or actions to which it leads' (Davenport and Prusak, 1998, p. 6). Such actions, when tied to corporate strategies, result in new products and services, potential new vectors of value-realization.

And, because a core aspect of the knowledge economy is sharing knowledge (qua information), knowledge workers appear to become more valuable in relation to all the information being captured as knowledge is shared. To put it simply, the more knowledge qua information produced, the more valuable the worker who can mine it and find value. In the knowledge management literature, the knowledge worker is arranged hierarchically above not only industrial workers but also with respect to mere data or information workers (Dalkir,

2011, p. 23; Davenport and Prusak, 1998). Mere-data and mere-information workers certainly produce information, either implicitly through their actions – what is now called 'data exhaust' (Mayer-Schönberger and Cukier, 2013, pp. 113–115) – or explicitly by entering data into a database. However, they are presented as incapable of making *knowledge* out of that information, even as they make more of it. In contrast, the rare knowledge worker is one who can look at the growing piles of information and create new ideas.

However, there is a fear at the center of the knowledge management literature: the loss of knowledge that can happen when workers leave a firm (Gehl, 2014a). As knowledge management theorist Amrit Tiwana (2000) notes, 'When the person having that critical piece of knowledge quits to join a competitor, that knowledge also walks out the door' (p. 36). Unlike tangible assets such as machinery, embodied knowledge can get up and leave a firm, thus depriving the firm of its investment in that 'human capital' (Folbre, 2012). This rare worker is thus highly mobile, a 'free agent' in much the same sense the term is used in professional sports (Pink, 2001). Along those lines, corporations, universities and governments are said to be fighting over the best brains around, hoping to benefit from their skills before they move on. Hence, the hoopla over Richard Florida's (2002, 2005) so-called 'Creative Class' enjoying life in 'Creative Cities' the world over. This rare subject is to be 'attracted' to regions, cities and employers, rather than compelled. An illustrative example of this comes from patent-tracking data from the World Intellectual Property Organization, which shows that North America and Europe attracted over 90 percent of the world's patent holders away from less-developed states, with China leading the world in having inventors leave for wealthier countries (Fink et al., 2013, pp. 4–6). Countries and regions that cannot attract these valuable workers are said to face a 'brain drain' and are only left to hope that the developed world continues to 'transfer knowledge' to them at a discount (Cherlet, 2014; King and McGrath, 2004).

Notably, then, the knowledge worker appears in a 'reciprocal and incompatible' (Koopman, 2013) tension with sharing. In these discourses, a skilled knowledge worker is someone that is hard to share, while knowledge is something that cannot exist unless it is shared. The latter grows and the former seems to become rare. Knowledge work is an elite practice, valorized in government funding schemes and in the business literature as the engine of economic growth (Ross, 2010). In this view, knowledge workers cannot be shared between firms or states. This point will be picked up below.

Tying the threads: 'the sexiest job of the 21st century'

It should be obvious how the current obsession with big data fits into this genealogy. (Digitized) sharing in all its forms – from social media sharing to inter-institutional sharing to the sharing of knowledge in firms – has gotten out of hand. 'Data Warehouses' struggle to store everything from tweets to likes to shares to comments to best practices to emails to documentation. Structured data are outstripped by semi-structured and even unstructured data (Coté, 2013). Governments, employers and marketers are recognizing that these piles of data can tell them about the sentiments, health, dispositions and proclivities of citizens, employees and consumers – but they do not know how to learn these details. In order to even work with such masses of data, new computational techniques, database structures and network architectures are required; computers are disciplined to

communicate, to share their processors and hard drives (Gehl, 2013; Jacobs, 2009). Meanwhile, new technologies of sharing – social media, smartphones, wearable computers, networked objects – allow for more and more data to be added to the existing streams such as government and business recordkeeping. In this sense, N = all is a nightmare.

But the nightmare can be abolished. To use the hyperbolic language of the business literature, the 'race is on' to hire the latest knowledge worker capable of creating value from this mess: data scientists. Data science is now famous for being dubbed the 'sexiest job of the 21st century' by the *Harvard Business Review* (Davenport and Patil, 2012). The Data Scientist – an ideal subject being constructed in myriad books, academic articles, conferences and popular press reports – is a heterogeneous assemblage of multiple threads: centuries of knowledge statistical analysis (Cleveland, 2001; Hacking, 1990), computational power, advances in algorithms within the field of computer science, new forms of information storage and retrieval such as Hadoop (Coté, 2013), new programming languages (Mackenzie, 2013), entrepreneurialism and, not to be overstated, the tendency for the field of business to produce novel-sounding concepts.

In the literature about this subject, the data scientist is capable of mining messy masses of data and producing knowledge that states and corporations crave. Data scientists are valuable because they are capable of discerning the 'option value' of large datasets (Mayer-Schönberger and Cukier, 2013, p. 102). As Mayer-Schönberger and Cukier (2013) put it, this is a move beyond analysis of 'superficial sharing of photos, status updates, and "likes"' (p. 92) to see how N = all could tell us something new about credit systems, the spread of disease, navigation, logistics or the allocation of public resources. With this new knowledge, a firm can create new products and services and maintain its competitive edge while governments can better manage their populations (even with fewer resources) (Royster, 2013). A common refrain among those who proclaim the rise of the data scientist is that we are in for a revolution, but only if we recognize the problem of too much data and accept the impartial findings of data science for the good of us all.

Moreover, the Data Scientist is seen to comprise specialized and hard-earned training and talents that are not easily replicable. As Jeff Hammerbacher (2009) of Facebook proclaims, 'the future belongs to the data scientist!' (p. 84). The future is a matter of attracting this rare subject who can navigate the petabytes of shared data. As Davenport and Patil (2012) put it, 'If capitalizing on big data depends on hiring scarce data scientists, then the challenge for managers is to learn how to identify that talent, attract it to an enterprise, and make it productive' (p. 72). Like other knowledge workers before them, data scientists are poised to benefit from what was shared by bringing rarified talents to bear on the fruits of sharing.

We can thus read the data scientist as the latest idealized subject arising from the sharing problem: there is a glut of data, and there is need for a subject to command, comprehend, mine and extract value from it. The data are copious and overwhelming; the Data Scientist is sexy and rare.

The threads frayed: from 'tacit knowledge' to 'process knowledge'

Critical work on data mining and analytics is engaged with these techniques as new forms of social management (e.g. Andrejevic, 2013; Mosco, 2014). I am wholly

supportive of this line of inquiry. However, here I want to ask questions about the labor of doing data mining. Will firms continue to fight over the rare subject capable of making sense of big data? Will data remain 'ubiquitous and cheap' while data scientists remain 'scarce and complementary'? Once again, genealogy can guide us to a potential answer. In this case, the question is as follows: how do sharing of knowledge and the production of a rare, idealized subject react with one another as they meet in the data scientist?

Again, I turn to the knowledge management literature, specifically the repeated discussions of the dangers of *knowledge hoarding* – the refusal of knowledge workers to share their knowledge. Barriers to knowledge sharing, including the 'lack of initiative and strategy by the workers', must be overcome (Santos et al., 2012). As Frances Horibe argues,

> Often people or groups of [the knowledge-hoarding] mind-set claim their knowledge is so special that it can't (or mustn't) be shared. They are the corporate equivalent of the mystic healer with knowledge too arcane and abstract for our poor brains to comprehend. (Horibe, 1999, p. 185)

To combat this, 'one of the challenges of knowledge management is to ensure that knowledge sharing is rewarded more than knowledge hoarding' (Davenport and Prusak, 1998, p. 28). In other words, the knowledge worker is to be rewarded if she or he shares her or his knowledge. Conversely, any knowledge worker who 'hoards' knowledge is to be punished. But the challenge is great. Management scholars Husted and Michailova (2002) put it bluntly: 'individuals in firms are inherently hostile to knowledge-sharing' (p. 61). Such hoarders must be compelled to share; key techniques to compel sharing include 'after-action reviews' to document projects postmortem, bringing in new knowledge workers who are willing to share, decomposing tasks into small parts to force multiple people to work together on them, requiring the use of particular IT systems and of course public shaming or firing of employees who fail to share (Horibe, 1999; Husted and Michailova, 2002, p. 71; Santos et al., 2012).

Why is there such attention in the knowledge management literature paid to the problem of knowledge hoarding? After all, if we agree with Machlup that knowledge sharing will grow the economy, with the Department of Homeland Security that intelligence sharing will help fight terrorism, or with Fowler that sharing data will improve public health, why *wouldn't* knowledge workers be willing to share?

The answer is that part of the knowledge that is to be shared involves *techniques* as well as facts, perceptions or ideas. The embodied techniques and practices held by workers are what knowledge management theorists call 'tacit knowledge'. As management scholar Linda Stoddart (2001) puts it, 'The essential challenge is to turn tacit knowledge into usable information that can be shared in order to stimulate innovation and create new products and services. To accomplish this, attitudes towards sharing information and knowledge need to change' (p. 20). If such attitudes are changed, the result is that the firm can 'be liberated from the fear of losing important intellectual assets if valued colleagues leave the firm' (Hall, 2001, p. 139). In other words, if the knowledge worker shares knowledge about practices and techniques in such a way that the firm can capture it, then the knowledge worker becomes less valuable.

If shared and captured, the tacit knowledge of the worker is transformed into what the knowledge management literature calls 'process knowledge' – firm-owned techniques

for producing products and services (Amaravadi and Lee, 2005; Bell et al., 2002; Cheng et al., 2014). As Amaravadi and Lee (2005) note, 'some amount of process knowledge is tacitly held by employees and presumably acquired through training and experience … the need to structure and organize it is vital to the [knowledge management] effort' (p. 66). Knowledge management, then, is not simply about creating databases for workers to consult, nor is it about sharing mere thoughts or feelings; it is about managing the process by which workers' skills can be transferred from their minds and bodies and into concrete objects, documents and processes controlled by the firm. This is, of course, an uneven and complicated process, but at least in the pages of the management literature it is an attainable and necessary goal.

The transformation of employee-embodied tacit knowledge into firm-controlled process knowledge is part of the means by which knowledge workers are drained of the very resource that allowed them to command high salaries and job security. The subject called forth to tame the excesses of sharing is also compelled to share her or his skills, ideas, techniques and 'best practices' (e.g. O'Dell et al., 1998) by which she or he was able to tame sharing. 'Sharing' – the informationalization of subjective knowledge – dissolves the unique subject meant to transcend it.

But to what ends? Returning to the idea of the 'free agent', knowledge workers enjoy free agency with no loyalty to the firm. However, a knowledge economy firm has a major advantage over these 'free agents': the ability to capture ideas, techniques, tools and inventions during the short period the free agent is a member of the corporate team. Seemingly ephemeral skills of knowledge workers can be abstracted and privatized through the advanced techniques of intellectual property capture. *So long as these skills are shared*. Such knowledge then becomes firm-owned process knowledge, 'essential to training employees, establishing standards and communicating best practices within the organization' (Amaravadi and Lee, 2005, p. 65). To use another example, it is telling that when one technology firm purchases another, it is very often to buy patents and other intellectual property. In other words, the firm is buying previously shared knowledge in the form of techniques. A knowledge worker is transient, but (thanks to global intellectual property trade agreements) if skills, practices and techniques are codified into intellectual property, they last nearly forever. 'No loyalty' goes both ways.

In addition, the development, sharing and abstraction of intellectual techniques can give rise to training systems meant to produce more of the sort of worker an industry desires (Allen, 2002; Cheng et al., 2014). Recalling an earlier wave of knowledge workers who were in their day sexy and rare, Davenport and Patil remind us of the Wall Street 'quants' of the 1980s and 1990s, the physicists and mathematicians who gave us the new algorithms of trading and hedging that have since paid so many dividends in our global economy. While these quants were once rare, once their value was recognized, 'a variety of universities developed master's programs in financial engineering, which churned out a second generation of talent that was more accessible to mainstream firms' (Davenport and Patil, 2012, p. 76). 'More accessible' meaning, of course, cheaper. Or, not even needed: Wall Street hiring is down since 2007, even though stocks are hitting record highs (Lopez, 2013). This same pattern of wealth growth and labor stagnation has taken place with other formerly 'sexy' knowledge fields, such as 'software engineering, architectural design, financial analysis, diagnostic radiology, and legal services' (Mosco and Stevens, 2007; see also Ross, 2010).

Finally, social media and concomitant practices such as crowdsourcing enable firms to tap into the collective – and freely shared – practices and techniques of 'ordinary' people, their 'collective intelligence'. Where a firm might have to pay experts to do a task, now they can call on the crowd to make designs, create new advertising campaigns or solve intellectual problems. The benefit to end users? The satisfaction of sharing their insights with firms that were previously oblivious to them.

If knowledge sharing – specifically the sharing of skills and techniques – can be folded back upon the very subject called forth to both enable and contain knowledge, any particular knowledge worker du jour would be vulnerable to being made ubiquitous and cheap, vulnerable to what Adrian Mackenzie (2013) called the 'recursive loop' of the analytics of analytics (p. 392). To be fair, it is impossible to say with any certainty that the data scientist will be dominated by data-minded corporations; it may not be possible for such firms to simply pick up the software, algorithms and techniques data scientists produce and turn them back on them. Data science is a new, unique field, and knowledge management theorists note that their managerial techniques must be concretized to work with specific knowledge fields (Davenport, 2005). However, what can be said with certainty is that executives at big data firms *desire* to dominate their data scientists. How can we tell? They say so in their writings.

Consider DJ Patil, Silicon Valley mogul formerly with LinkedIn. As he notes, even if a data scientist is with a firm for a few short years, having that person on the firm's data team can have the paradoxical result where

> you see data products being built in all parts of the company. When the company sees what can be created with data, when it sees the power of being data enabled, you'll see data products appearing everywhere. That's how you know when you've won. (Patil, 2011, p. 17)

In other words, your firm 'wins' when the techniques and skills of data scientists become firm-controlled process knowledge distributed throughout the firm. Patil recommends capturing data scientists' knowledge and distributing it widely, either to other people or through new tools or processes. Thus, if data science can be abstracted from the scientist and engineered into easy-to-use big data tools (such as Facebook-born software packages Hive and HiPal), then non-data scientists (marketers, managers and salespeople) can do big data queries and analysis themselves (Hammerbacher, 2009, p. 81). In their perspective, any given data scientist might be in the firm for only a brief period, but if the result is a 'data-minded firm', then overpaying for free-agent knowledge workers is no longer a concern.

In addition, given the hype over big data, universities will train the second generation of data scientists (Cleveland, 2001; Royster, 2013, p. 9). Responding to a 2011 McKinsey report claiming there will be nearly 200,000 unfilled data science jobs in the United States alone (Manyika et al., 2011), universities are starting data science programs across North America (Henschen, 2013; 'Map of University Programs in Big Data Analytics', 2014). Just as the ranks of the quants were grown with university training, so too will the ranks of data scientists. Moreover, in line with the historical push for 'corporate universities' – that is, training centers operated by corporations to train their workers (Allen, 2002) – big data scholars Davenport and Harris (2007) note that data scientists will be

expected to share their knowledge within firms, training 'analytics amateurs' in the ways of data science (p. 185) and thus explicating their process knowledge. Indeed, sharing of techniques is seen in the knowledge management literature as a key duty of the knowledge worker and is part of the overall emphasis on 'learning' within corporate culture (Bell et al., 2002; Cheng et al., 2014; e.g. Horibe, 1999, pp. 185–186).

Finally, calls are being made to 'the crowd' to gain their collective intelligence in big data analysis problems (Mackenzie, 2013, p. 400). Kaggle.com, a data science crowdsourcing site proclaiming itself to be 'the leading platform for predictive modeling competitions', invites firms to share datasets with 'the world's largest community of data scientists' comprising 'tens of thousands of PhDs from quantitative fields such as computer science, statistics, econometrics, maths and physics, and industries such as insurance, finance, science, and technology. They come from over 100 countries and 200 universities' (see kaggle.com/about). As I write, there are a few competitions for prizes ranging from US$30,000 to US$100,000. However, the majority of contests feature 'knowledge' as the prize, which is to say that the winner will learn something (rather than be paid for his or her work). Just as Flickr has for photography and Google ReCaptcha has for digitizing texts, Kaggle promises to radically reduce the cost of data analysis by creating a market for budding young data scientists to give their work away for free.

These are all methods by which the very practice that gives rise to big data, the informationalization of techniques through sharing, could reduce the value of the data scientist, the subject called upon to mine those data. Beyond this, data-minded firms will have another major advantage in managing and mass-producing such subjects precisely because of their command of data. Consider 'Big Data Management' (McAfee and Brynjolfsson, 2012), a practice that helps manage knowledge workers by using the very data they are asked to produce. Thanks to big piles of data,

> we can measure and therefore manage more precisely than ever before. We can make better predictions and smarter decisions. We can target more-effective interventions, and can do so in areas that so far have been dominated by gut and intuition rather than by data and rigor … Simply put, because of big data, managers can measure, and hence know, radically more about their businesses, and directly translate that knowledge into improved decision making and performance. (McAfee and Brynjolfsson, 2012, p. 4)

'N = all' here means a desire for total knowledge of the firm's intellectual capacity. Such data can be mined in order to better manage knowledge workers, including the data scientists tasked with doing the mining. As Adrian Mackenzie (2013) notes in his analysis of machine learning (a technique for dealing with big data), 'The generic predictive capacity of machine learning can be turned in almost any direction, including on programmers themselves' (p. 403). As Stephen Baker notes in his homage to data science, *The Numerati*, no matter how creative a worker is, data can be used to dissolve that worker. 'In a workplace defined by metrics, even those of us who like to think that we're beyond measurement will face growing pressure to build our case with numbers of our own' (Baker, 2008, p. 40). The data scientist, Baker notes, is not an exception – even as Baker profiles and praises the data scientists who are bringing this quantification of labor about.

Conclusion

Ultimately, what big data firms crave from their data scientists is precisely what they enjoy with their data: cheapness and ubiquity. They may not have that now, but their desire is manifested in their calls for university programs in big data, easy-to-use data analysis software systems and managerial texts explaining how to hire and manage data scientists. The first generation of data scientists is being confronted with ways to share their knowledge, techniques and ideas with their employers, just as past generations of knowledge workers were compelled to share their knowledges and techniques. This repeats what has happened to many 'sexy' knowledge workers over the past several decades, all of whom were in their turn darlings of the knowledge economy, all of whom were soon subject to the logics of casualization and precarity. It remains to be seen whether data science can be an exception to this historical arc.

Conversely, if data is seen to be a major corporate asset, it is hard to imagine firms continuing to share it. Data will no doubt continue to be ubiquitous. Was it ever cheap? It is beyond the scope of this article to trace all the ways in which firms will lock down data, but a few things can be mentioned here. As Mayer-Schönberger and Cukier (2013) note, naive firms might have initially shared their data with third parties, but now are increasingly privatizing such data with intellectual property mechanisms and selling it (pp. 134–138).[2] Twitter, for example, is a big data darling, but only those who can afford to pay get to see the full 'firehose' of tweets. Many of the technologies that allow for the capture of ideas can be used to enclose big datasets and protect them with the full force of international intellectual property laws, limiting access to this ostensibly 'ubiquitous and cheap' resource and thus reversing Hal Varian's proverb.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Notes

1. Hadoop is an open-source software system designed for storing and analyzing large datasets in server farms.
2. However, see Davenport and Harris' (2007, pp. 182–183) discussion of sharing analytics and data. They suggest that firms will continue to share data among one another, making no mention of leasing such data. Indeed, this tension between sharing and making proprietary is central to discourses about knowledge (Bell et al., 2002) and now, it appears, big data.

References

Allen M (ed.) (2002) *The Corporate University Handbook: Designing, Managing, and Growing a Successful Program*. New York: AMACOM.

Amaravadi CS and Lee I (2005) The dimensions of process knowledge. *Knowledge and Process Management* 12: 65–76.

Andrejevic M (2013) *Infoglut: How Too Much Information Is Changing the Way We Think and Know*. New York: Routledge.

Baker S (2008) *The Numerati*. Boston, MA: Houghton Mifflin.

Bell DG, Giordano R and Putz P (2002) Inter-firm sharing of process knowledge: Exploring knowledge markets. *Knowledge and Process Management* 9: 12–22.

Bock GW and Kim Y-G (2002) Breaking the myths of rewards: An exploratory study of attitudes about knowledge sharing. *Information Resources Management Journal* 15: 14–21.

Callon M (1986) Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St Brieuc Bay. In: Law J (ed.) *Power, Action, and Belief: A New Sociology of Knowledge?* London; Boston, MA: Routledge & Kegan Paul, pp.169–233.

Cheng C-Y, Ou T-Y, Chen T-L, et al. (2014) Transferring cognitive apprenticeship to manufacturing process knowledge management system. *VINE* 44: 420–444.

Chen H, Reid E, Sinai J, et al. (eds) (2008) *Terrorism Informatics: Knowledge Management and Data Mining for Homeland Security*. New York: Springer.

Cherlet J (2014) Epistemic and technological determinism in development aid. *Science Technology Human Values* 39: 773–794.

Christakis NA and Fowler JH (2009) *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives*. New York: Little, Brown and Co.

Cleveland WS (2001) Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review* 69: 21–26.

Coté M (2013) Data motility: The materiality of big social data. Available at: https://www.academia.edu/4767377/Data_Motility_The_Materiality_of_Big_Social_Data (accessed 11 March 2014).

Dalkir K (2011) *Knowledge Management in Theory and Practice* (2nd edn). Cambridge, MA: MIT Press.

Davenport TH (2005) *Thinking for a Living: How to Get Better Performance and Results from Knowledge Worker*. Boston, MA: Harvard Business School Press.

Davenport TH and Harris JG (2007) *Competing on Analytics: The New Science of Winning*. Boston, MA: Harvard Business School Press.

Davenport TH and Patil DJ (2012) Data scientist. *Harvard Business Review* 90: 70–76.

Davenport TH and Prusak L (1998) *Working Knowledge: How Organizations Manage What They Know*. Boston, MA: Harvard Business School Press.

Denning S (2001) *The Springboard: How Storytelling Ignites Action in Knowledge-Era Organizations*. Boston, MA: Butterworth-Heinemann.

Drucker PF (1992) *The Age of Discontinuity: Guidelines to Our Changing Society*. New Brunswick, NJ; London: Transaction Publishers.

Fink C, Raffo J and Miguelez E (2013) *Study on Intellectual Property and Brain Drain – A Mapping Exercise* (No. CDIP/12/inf/4). Geneva: World Intellectual Property Organization.

Florida R (2002) *The Rise of the Creative Class: And How It's Transforming Work, Leisure, Community and Everyday Life*. New York: Basic Books.

Florida RL (2005) *The Flight of the Creative Class: The New Global Competition for Talent*. New York: HarperBusiness.

Folbre N (2012) The political economy of human capital. *Review of Radical Political Economics* 44: 281–292.

Fowler, JH (2014, February). *Connected: The surprising power of our social networks and how they shape our lives*. Presentation, University of Utah, Boston, MA.

Gehl RW (2013) Server farms: Disciplined machines behind noopower. *Media Fields Journal*, no. 6 Available at: http://mediafieldsjournal.squarespace.com/storage/issue-6/PDFs/Gehl_Final.pdf

Gehl RW (2014a) Power from the c-suite: The chief knowledge officer and chief learning officer as agents of noopower. *Communication and Critical/Cultural Studies* 11, 175–194.

Gehl RW (2014b) *Reverse Engineering Social Media: Software, Culture, and Political Economy in New Media Capitalism*. Philadelphia, PA: Temple University Press.

Godin B (2010) The knowledge economy: Fritz Machlup's construction of a synthetic concept. In: Viale R and Etzkowitz H (eds) *The Capitalization of Knowledge: A Triple Helix of University-Industry-Government*. Cheltenham; Northampton, MA: Edward Elgar, pp.261–290.

Hacking I (1990) *The Taming of Chance*. Cambridge: Cambridge University Press.

Hall H (2001) Input-friendliness: Motivating knowledge sharing across intranets. *Journal of Information Science* 27: 139–146.

Hammerbacher J (2009) Information platforms and the rise of the data scientist. In: Segaran T and Hammerbacher J (eds) *Beautiful Data*. Sebastopol, CA: O'Reilly Media, Inc, pp.73–84.

Henschen D (2013) Big data analytics master's degrees: 20 top programs. *InformationWeek*. Available at: http://www.informationweek.com/big-data/news/big-data-analytics/big-data-analytics-masters-degrees-20-top-programs/240145673 (accessed 21 April 2014).

Horibe F (1999) Managing knowledge workers: New skills and attitudes to unlock the intellectual capital in your organization. Toronto, ON, Canada; New York: John Wiley.

Husted K and Michailova S (2002) Diagnosing and fighting knowledge-sharing hostility. *Organizational Dynamics* 31: 60–73.

Jacobs A (2009) The pathologies of big data. *Communications of the ACM* 52: 36–44.

John NA (2013) Sharing and Web 2.0: The emergence of a keyword. *New Media Society* 15: 167–182.

King K and McGrath SA (2004) *Knowledge for Development? Comparing British, Japanese, Swedish and World Bank Aid*. London: Zed Books; Cape Town, South Africa: HSRC Press.

Koopman C (2013) *Genealogy as Critique: Foucault and the Problems of Modernity*. Bloomington, IN: Indiana University Press.

Lee J-N (2001) The impact of knowledge sharing, organizational capability and partnership quality on IS outsourcing success. *Information & Management* 38: 323–335.

Liu N-C and Liu M-S (2011) Human resource practices and individual knowledge-sharing behavior – An empirical study for Taiwanese R&D professionals. *International Journal of Human Resource Management* 22: 981–997.

Lopez L (2013) *Here's How Much the Average Wall Streeter Makes Compared to the Average New Yorker*. Slate. Available at: http://www.slate.com/blogs/business_insider/2013/10/22/wall_street_jobs_and_salaries_what_s_in_the_latest_new_york_state_comptroller.html

McAfee A and Brynjolfsson E (2012) Big data: The management revolution. *Harvard Business Review* 90: 60–68.

Machlup F (1962) *The Production and Distribution of Knowledge in the United States*. Princeton, NJ: Princeton University Press.

Mackenzie A (2013) Programming subjects in the regime of anticipation: Software studies and subjectivity. *Subjectivity* 6: 391–405.

Mahon M (1992) *Foucault's Nietzschean Genealogy: Truth, Power, and the Subject*. Albany, NY: State University of New York Press.

Manyika J, Chui M, Brown B, et al. (2011) *Big Data: The Next Frontier for Innovation, Competition, and Productivity.* Washington, DC: McKinsey and Company.

Map of University Programs in Big Data Analytics (2014) DataInformed. Available at: http://data-informed.com/bigdata_university_map/ (accessed 21 April 2014).

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.

Mehta L (2001) Commentary: The World Bank and its emerging knowledge empire. *Human Organization* 60: 189–196.

Monahan T and Palmer NA (2009) The emerging politics of DHS fusion centers. *Security Dialogue* 40: 617–636.

Monahan T and Regan PM (2011) Centers of concatenation: Fusing data in post-9/11 security organizations. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1913470

Mosco V (2014) *To the Cloud: Big Data in a Turbulent World*. Boulder, CO: Paradigm.
Mosco V and Stevens A (2007) Outsourcing knowledge work: Labor responds to the new international division of labor. In: McKercher C and Mosco V (eds) *Knowledge Workers in the Information Society*. Lanham, MD: Lexington Books, pp.147–162.
O'Dell CS, Grayson CJ and Essaides N (1998) *If Only We Knew What We Know: The Transfer of Internal Knowledge and Best Practice*. New York: Free Press.
Offsey S (1997) Knowledge management: Linking people to knowledge for bottom line results. *Journal of Knowledge Management* 1: 113–122.
O'Leary DE (1998) Enterprise knowledge management. *Computer* 31: 54–61.
Patil DJ (2011) *Building Data Science Teams*. Sebastopol, CA: O'Reilly Media, Inc.
Pink DH (2001) *Free Agent Nation: How America's New Independent Workers Are Transforming the Way We Live*. New York: Warner Books.
Regan PM and Monahan T (2013) Beyond counterterrorism: Data sharing, privacy, and organizational histories of DHS fusion centers. *International Journal of E-Politics* 4: 1–14.
Ross A (2010) *Nice Work If You Can Get It: Life and Labor in Precarious Times*. New York: New York University Press.
Royster S (2013) Working with Big Data. *Occupational Outlook Quarterly*, pp. 2–10. Available at: http://www.bls.gov/careeroutlook/2013/fall/art01.pdf
Santos VR, Soares AL and Carvalho JÁ (2012) Knowledge sharing barriers in complex research and development projects: An exploratory study on the perceptions of project managers. *Knowledge and Process Management* 19: 27–38.
Stoddart L (2001) Managing intranets to encourage knowledge sharing: Opportunities and constraints. *Online Information Review* 25: 19–29.
Tiwana A (2000) *The Knowledge Management Toolkit: Practical Techniques for Building a Knowledge Management System*. Upper Saddle River, NJ: Prentice Hall PTR.
Zuckerberg M (2009) Improving your ability to share and connect. In: Facebook Blog. Available at: http://blog.facebook.com/blog.php?post=57822962130 (accessed 13 May 2009).

Biographical note

Robert W Gehl is currently an assistant professor in the Department of Communication at the University of Utah. His research draws on science and technology studies, software studies and critical/cultural studies and focuses on the intersections between technology, subjectivity and practice. His book, *Reverse Engineering Social Media* (2014, Temple University Press), explores the architecture and political economy of social media.