Towards critical data studies: Charting and unpacking data assemblages and their work

Rob Kitchin and Tracey P. Lauriault, NIRSA, National University of Ireland Maynooth, County Kildare, Ireland

The Programmable City Working Paper 2

http://www.nuim.ie/progcity/

29th July 2014

Pre-print version of chapter to be published in Eckert, J., Shears, A. and Thatcher, J. (eds) *Geoweb and Big Data*. University of Nebraska Press.

Abstract

The growth of big data and the development of digital data infrastructures raises numerous questions about the nature of data, how they are being produced, organized, analyzed and employed, and how best to make sense of them and the work they do. Critical data studies endeavours to answer such questions. This paper sets out a vision for critical data studies, building on the initial provocations of Dalton and Thatcher (2014). It is divided into three sections. The first details the recent step change in the production and employment of data and how data and databases are being reconceptualised. The second forwards the notion of a data assemblage that encompasses all of the technological, political, social and economic apparatuses and elements that constitutes and frames the generation, circulation and deployment of data. Drawing on the ideas of Michel Foucault and Ian Hacking it is posited that one way to enact critical data studies is to chart and unpack data assemblages. The third starts to unpack some the ways that data assemblages do work in the world with respect to dataveillance and the erosion of privacy, profiling and social sorting, anticipatory governance, and secondary uses and control creep. The paper concludes by arguing for greater conceptual work and empirical research to underpin and flesh out critical data studies.

Key words

big data, critical data studies, data assemblages, data infrastructures, civil liberties

A critical approach to data

Societies have collected, stored and analyzed data for a couple of millennia as a means to record and manage their activities. For example, the ancient Egyptians collected administrative records of land deeds, field sizes and livestock for taxation purposes, the 1086 Domesday Book captured demographic data, double entry bookkeeping was used by bankers and insurers in the fourteenth century, and the first national registry was undertaken in Sweden in the seventeenth century (Dupaquier & Dupaquier, 1985; Bard and Shubert 1999; Poovey, 1998; Porter, 1986). However, the term 'data' was only used for the first time in the English language in the seventeenth century with the growth of science, the development of statistics, and the shift from knowledge built from theology, exhortation and sentiment to facts, evidence and the testing of theory through experiment (Poovey 1998; Garvey 2013; Rosenberg 2013). Over time the importance of data has grown, becoming central to how knowledge is produced, business conducted, and governance enacted. Data provide the key inputs to systems that individuals, institutions, businesses and science employ in order to understand, explain, manage, regulate and predict the world we live in, and are used to create new innovations, products, and policies.

The volume, variety and use of data has grown enormously since the seventeenth century. Indeed, there has long been the creation and maintenance of very large datasets, such as censuses or government administrative and natural resource databases. Such databases, however, have typically been generated every few years or are sampled. In contrast, over the past fifty years we have begun to enter the era of big data that are (Kitchin 2013: 262):

- huge in *volume*, consisting of terabytes or petabytes of data;
- high in *velocity*, being created in or near real-time;
- diverse in *variety*, being structured and unstructured in nature;
- *exhaustive* in scope, striving to capture entire populations or systems (n=all);
- fine-grained in *resolution* and uniquely *indexical* in identification;
- *relational* in nature, containing common fields that enable the conjoining of different data sets;

- *flexible*, holding the traits of *extensionality* (can add new fields easily) and *scaleability* (can expand in size rapidly).
  (boyd and Crawford 2012; Dodge and Kitchin 2005; Laney 2001; Marz and Warren 2012; Mayer-Schonberger and Cukier 2013; Zikopoulos *et al*., 2012)

While there are varying estimates, depending on the methodology used, as to the growth of data production caused by big data (e.g., Hilbert and López 2009; Gantz and Reinsel 2011; Short *et al.* 2011), it is clear that there has been a recent step-change in the volume of data generated, especially since the start of the new millennium. Gantz and Reinsel (2011) estimated that data volumes had grown by a factor of nine in the preceding five years, and Manyika *et al.* (2011) projected a 40 percent rise in data generated globally per year. In 2013, EU commissioner for Digital Agenda, Neelie Kroes, reported that 1.7 million billion bytes of data per minute were being generated globally (Rial 2013). Such rises and projections are due to the continuous and exhaustive, rather than sampled, production of data. For example, in 2012, Facebook reported that it was processing 2.5 billion pieces of content (links, comments, etc), 2.7 billion 'Like' actions and 300 million photo uploads *per day* (Constine 2012) and Wal-Mart was generating more than 2.5 petabytes (2^{50} bytes) of data relating to more than 1 million customer transactions *every hour* (Open Data Center Alliance 2012).

These massive volumes of data are being produced by a diverse set of information and communication technologies that increasingly mediate and augment our everyday lives, for example, digital CCTV, retail checkout tills, smart phones, online transactions and interactions, sensors and scanners, and social and locative media. As well as being produced by government agencies, vast quantities of detailed data are now being generated by mobile phone operators, app developers, internet companies, financial institutions, retail chains, and surveillance and security firms, and data are being routinely traded to and between data brokers as an increasingly important commodity. Moreover, more and more analogue data held in archives and repositories are being digitized and linked together and made available through new data infrastructures, and vast swathes of government produced and held data are being made openly accessible as the open data movement gains traction (Lauriault et al. 2007, Kitchin 2014a).

This step-change in data production has prompted critical reflection on the nature of data and how they are employed. As the concept of data developed, data largely came to be understood as being pre-analytical and pre-factual, that which exists prior to interpretation

and argument; the raw material from which information and knowledge are built. From this perspective, data are understood as being representative, capturing the world as numbers, characters, symbols, images, sounds, electromagnetic waves, bits, etc, and holding the precepts of being abstract, discrete, aggregative (they can be added together), non-variant, and meaningful independent of format, medium, language, producer and context (i.e., data hold their meaning whether stored as analog or digital, viewed on paper or screen, or expressed in any language) (Floridi 2008, 2010; Rosenberg 2013). Data are viewed as being benign, neutral, objective and non-ideological in essence, reflecting the world as it is subject to technical constraints; they do not hold any inherent meaning and can be taken at face value (Pérez-Montoro and Díaz Nafría 2010). Indeed, the terms commonly used to detail how data are handled suggest benign technical processes: 'collected', 'entered', 'compiled', 'stored', 'processed' and 'mined' (Gitelman and Jackson 2013). In other words, it is only the uses of data that are political, not the data themselves.

This understanding of data has been challenged in recent years. Contra, the notion that data is pre-analytic and pre-factual, it has been argued that data are constitutive of the ideas, techniques, technologies, people, systems and contexts that conceive, produce, process, manage, and analyze them (Bowker and Star 1999; Lauriault 2012; Ribes and Jackson 2013; Kitchin 2014a). In other words, how data are conceived, measured and employed actively frames their nature. Data do not pre-exist their generation; they do not arise from nowhere and their generation is not inevitable: protocols, organisational processes, measurement scales, categories, and standards are designed, negotiated and debated, and there is a certain messiness to data generation. As Gitelman and Jackson (2013: 2, following Bowker 2005) put it, "raw data is an oxymoron"; "data are always already 'cooked'." Data then are situated, contingent, relational, and framed, and used contextually to try and achieve certain aims and goals.

Similarly, databases and repositories are not simply a neutral, technical means of assembling and sharing data, but are bundles of contingent and relational processes that do work in the world (Star and Ruhleder 1996; Kitchin and Dodge 2011). They are complex socio-technical systems that are embedded within a larger institutional landscape of researchers, institutions and corporations (Ruppert 2012) and are subject to socio-technical regimes "grounded in ... engineering and industrial practices, technological artifacts, political programs, and institutional ideologies which act together to govern technological development" (Hetch, 2001:257). Databases and repositories are expressions of knowledge/power, shaping what questions can be asked, how they are asked, how they are

answered, how the answers are deployed, and who can ask them (Lauriault 2012; Ruppert 2012).

Beyond this philosophical rethinking of data, scholars have begun to make sense of data ethically, politically/economically, spatially/temporally, and technically (Kitchin 2014a). Data can concern all aspect of everyday life, including sensitive issues, and be used in all kinds of ways, including to exploit, discriminate against and persecute people. There are then a series of live moral and ethical questions concerning how data are produced, shared, traded and protected; how data should be governed by rules, principles, policies, licenses and laws; and under what circumstances and to what ends data can be employed. There are no simple answers to such questions, but the rise of more widespread and invasive data generation and more sophisticated means of data analysis creates an imperative for public debate and action. Related, data are framed by political concerns as to how they are normatively conceived and contested as public and private goods. The open data and open government movements, for example, cast data as a public commons that should be freely accessible. In contrast, business views data as a valuable commodity that, on the one hand, needs to be protected through intellectual property regimes (copyright, patents, ownership rights) and, on the other, be exploitable for capital gain. Indeed, data often constitute an economic resource: for government they are sold under cost recovery regimes and for business they are tradable commodities to which additional value can be added and extracted (e.g., derived data, analysis, knowledge). In the present era, data are a key component of the emerging knowledge economy enhancing productivity, competiveness, efficiencies, sustainability and capital accumulation. The ethics, politics and economics of data develops and mutates across space and time with changing regimes, technologies and priorities. From a technical perspective, there has been a focus on how to handle, store and analyze huge torrents of data, with the development of data mining and data analytics techniques dependent on machine learning, and concerns with respect to data quality, validity, reliability, authenticity, usability, and lineage.

In sum, we are starting to witness the development of what Dalton and Thatcher (2014) call critical data studies -- research and thinking that applies critical social theory to data to explore the ways in which they are never simply neutral, objective, independent, raw representations of the world, but are situated, contingent, relational, contextual, and do active work in the world. In their analysis, Dalton and Thatcher set out seven provocations needed to provide a comprehensive critique of the new regimes of data:

1. situate data regimes in time and space
2. expose data as inherently political and whose interests they serve
3. unpack the complex, non-deterministic relationship between data and society
4. illustrate the ways in which data are never raw
5. expose the fallacies that data can speak for themselves and that big data will replace small data
6. explore how new data regimes can be used in socially progressive ways
7. examine how academia engages with new data regimes and the opportunities of such engagement.

We agree with the need for all of these provocations. In a short presentation at the Association of American Geographers, one of us (Kitchin 2014b) set out a vision for what critical data studies might look like: to unpack the complex assemblages that produce, circulate, share/sell and utilise data in diverse ways; to chart the diverse work they do and their consequences for how the world is known, governed and lived-in; and to survey the wider landscape of data assemblages and how they interact to form intersecting data products, services and markets and shape policy and regulation. It is to this endeavour that we now turn.

Charting and unpacking data assemblages

Kitchin (2014a: 24) defines a 'data assemblage' as a complex socio-technical system, composed of many apparatuses and elements that are thoroughly entwined, whose central concern is the production of a data (see Table 1). A data assemblage consists of more than the data system/infrastructure itself, such as a big data system, an open data repository, or a data archive, to include all of the technological, political, social and economic apparatuses that frames their nature, operation and work. The apparatuses and elements detailed in Table 1 interact with and shape each other through a contingent and complex web of multifaceted relations. And just as data are a product of the assemblage, the assemblage is structured and managed to produce those data (Ribes and Jackson 2013). Data and their assemblage are thus mutually constituted, bound together in a set of contingent, relational and contextual discursive and material practices and relations. For example, the data assemblage of a census consists of a large amalgam of apparatuses and elements that shape how it is formulated, administered, processed, communicated, and how its findings are employed. A census is underpinned by a realist system of thought; it has a diverse set of accompanying forms of

supporting documentation; its questions are negotiated by many stakeholders; its costs are a source of contention; its administering and reporting is shaped by legal frameworks and regulations; it is delivered through a diverse set of practices, undertaken by many workers, using a range of materials and infrastructures; and its data feed into all kinds of uses and secondary markets. Data assemblages evolve and mutate as new ideas and knowledges emerge, technologies are invented, organisations change, business models are created, the political economy alters, regulations and laws introduced and repealed, skill sets develop, debates take place, and markets grow or shrink. They are thus always in a state of becoming.

Table 1: The apparatus and elements of a data assemblage

Apparatus	**Elements**
Systems of thought	Modes of thinking, philosophies, theories, models, ideologies, rationalities, etc.
Forms of knowledge	Research texts, manuals, magazines, websites, experience, word of mouth, chat forums, etc.
Finance	Business models, investment, venture capital, grants, philanthropy, profit, etc.
Political economy	Policy, tax regimes, incentive instruments, public and political opinion, etc.
Governmentalities and legalities	Data standards, file formats, system requirements, protocols, regulations, laws, licensing, intellectual property regimes, ethical considerations, etc.
Materialities and infrastructures	Paper/pens, computers, digital devices, sensors, scanners, databases, networks, servers, buildings, etc.
Practices	Techniques, ways of doing, learned behaviours, scientific conventions, etc.
Organisations and institutions	Archives, corporations, consultants, manufacturers, retailers, government agencies, universities, conferences, clubs and societies, committees and boards, communities of practice, etc.
Subjectivities and communities	Of data producers, experts, curators, managers, analysts, scientists, politicians, users, citizens, etc.
Places	Labs, offices, field sites, data centres, server farms, business parks, etc, and their agglomerations
Marketplace	For data, its derivatives (e.g., text, tables, graphs, maps), analysts, analytic software, interpretations, etc.

Source: Kitchin (2014a: 25)

This notion of a data assemblage is similar to Foucault's (1977) concept of the 'dispositif' that refers to a 'thoroughly heterogeneous ensemble consisting of discourses, institutions, architectural forms, regulatory decisions, laws, administrative measures, scientific statements, philosophical, moral and philanthropic propositions' (in Gordon 1980:194) which enhance and maintain the exercise of power within society. The dispositif of a data infrastructure produces what Foucault terms 'power/knowledge', that is knowledge that fulfils a strategic function: 'The apparatus is thus always inscribed in a play of power,

but it is also always linked to certain coordinates of knowledge which issue from it but, to an equal degree, condition it. This is what the apparatus consists in: strategies of relations of forces supporting, and supported by, types of knowledge' (in Gordon 1980:196). In other words, data infrastructures are never neutral, essential, objective; their data never raw but always cooked to some recipe by chefs embedded within institutions that have certain aspirations and goals and operate within wider frameworks.

This cooking of data is revealed through the work of Ian Hacking (1982, 1986, 1991, 2002, 2007), who drew inspiration from Foucault's thinking on the production of knowledge. Hacking posits that two interrelated processes are at work within a data assemblage that produce and legitimate its data and associated apparatuses/elements, and shapes how its data do work in the world, that in turn influences future iterations of data and the constitution of the assemblage. In both cases he posits that a dynamic nominalism is at work, wherein there is an interaction between data and that they represent, leading to mutual changes.

The first of these processes is what Hacking (1991, 2002 and 2007) terms 'the looping effect'. The looping effect concerns how data are classified and organised; how a data ontology comes into existence and how it can reshape that which has been classified. The loop has five stages (see Figure 1):

(i) classification, wherein things that are regarded as having shared characteristics are grouped together, or in cases of deviance forced into groupings;
(ii) objects of focus (e.g., people, spaces, fashions, diseases, etc) that in the case of people eventually start to self-identify with the class into which they are assigned, or for non-human, people come to understand and act toward through the classification;
(iii) institutions, who institutionalise classifications and manage the data infrastructures;
(iv) knowledge, that is used to formulate, reproduce and tweak classifications; and
(v) experts, those within institutions who produce and exercise knowledge, implementing the classification.

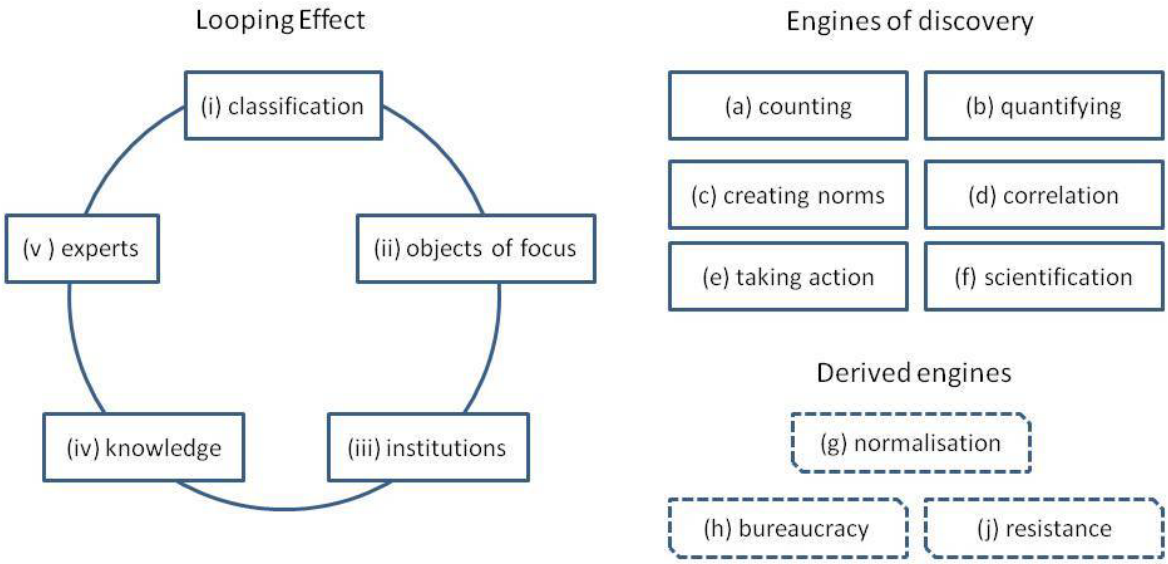
Through this looping effect Hacking argues a process of 'making people up' occurs in data systems such as the census or assessing mental health, wherein the systems of classification work to reshape society in the image of a data ontology; for example, people defining themselves or being defined by mental health symptoms, and a system of mental health facilities being built, staffed by specialist professionals.

The second of the processes are what Hacking terms ‘engines of discoverability’, that extend beyond simply methods, which he discusses using a medical lens, which Lauriault (2012) has modified to incorporate the making up of spaces as well as people. Hacking (2007) posits that there are a number of such engines, the last three of which are derived engines:

(a) counting the volumes of different phenomena;
(b) quantifying, turning counts into measures, rates, classifications;
(c) creating norms, establishing what might or should be expected;
(d) correlation, determining relationships between measures;
(e) taking action, employing knowledge to tackle and treat issues;
(f) scientification, establishing and adopting scientific knowledge;
(g) normalization, seeking to fashion the world to fit norms (e.g., encouraging diets to meet expected body mass indices);
(h) bureaucratization, putting in place institutions and procedures to administer the production of expectations and undertake action; and
(j) resistance to forms of knowledge, norms, bureaucracy by those that are affected in negative ways (e.g., homosexual and disabled people’s resistance to medicalized models that class, position and treat them in particular ways) or those forwarding alterative systems, interpretations and visions.

Taken together these engines undertake the work of a data assemblage at the same time as it legitimates and reproduces such work and the assemblage itself. For example, a census counts a population and aspects of their lives, turns them into measures, establishes baseline rates, assesses relationships between factors, and is transformed into knowledge, which leads to practices of normalization, and is enacted by dedicated and related bureaucracy. Each stage reinforces the previous and collectively they justify the work it does. The knowledge produced, and indeed the whole assemblage, can be resisted, as with the census boycotts in Germany in the 1980s (Hannah 2011) and the recent cancellation of the census in Canada.

Figure 1: The working of a data assemblage, following Ian Hacking (Lauriault 2012)



Data assemblages form part of a wider data landscape composed of many inter-related and interacting data assemblages and systems. Within the public sector, for example, there are thousands of data systems, each surrounded by a wider assemblage, that interact and work in concert to produce state services and forms of state control at local, regional and national scales. Often, this data landscape extends to the pan-national and the global, through inter-regional and worldwide data sets, data sharing arrangements and infrastructures, and the formulation of protocols, standards and legal frameworks (e.g., Global Spatial Data Infrastructures, INSPIRE). Likewise, within industry, firms create and occupy a complex data landscape, selling, buying and sharing data from millions of data systems, all part of wider socio-technical assemblages. For example, the data landscape of big data consists of hundreds of companies, ranging from small and local to large and global, who provide a range of complementary and competing services, such as cooked data, speciality compilers and aggregators, data analytics, segmentation tools, list management, interpretation and consulting, marketing, publishing, and research and development. We have barely begun to map out various data landscapes, their spatialities and temporalities, their complex political economy, and the work that they do in capturing, analyzing and reshaping the world. It to the latter we now turn.

Uncovering the work of data assemblages

As noted in the previous section, data assemblages do work in the world. Data are being leveraged to aid the tasks of governing people and territories, managing organisations,

producing capital, creating better places, improving healthcare, advancing science, and so on. This leveraging takes many forms, but the central tenet is that data if analyzed and exploited appropriately produce information and knowledge that can be used to reshape operating procedures and organizational structure, identify new products, segment markets, reduce uncertainty and risk, and increase efficiency, productivity, competitiveness and sustainability (Kitchin 2014a). Whilst much of the work to which data are put are beneficial to wider society, being used to improve quality of life and to tackle humanitarian and environmental issues, there is also a darker side to much data work. Here, we want to consider the latter, highlighting four ways in which data are being employed to produce pernicious social and economic relations: dataveillance and the erosion of privacy; profiling and social sorting; anticipatory governance; and secondary uses and control creep. These practices are presently the subject of much debate and there is an urgent need for critical studies that can inform the arguments being made.

As the revelations of Wikileaks, Edward Snowden and other whistleblowers, the Maher Arar case and other legal challenges with respect to erroneous record keeping and the mistreatment of individuals have demonstrated, from 9/11 onwards there has been a step change in the extent and nature of state-led surveillance and securitisation in many nations. Vast quantities of everyday communications (telephone calls, text messages, emails, social media), as well as general internet use, are being routinely and systematically gathered by organisations, such as the US National Security Agency, and analyzed for strategic intelligence (Amoore 2006, Bamford 2009). All nation states similarly gather large databases of information about citizens with respect to all aspects of their lives -- income, tax, welfare, health, education, and so on. Likewise, companies now routinely generate data with respect to all aspects of their business, including their customers and their patterns of consumption. Indeed, given the mediating role of software in tasks such as working, travelling, consuming, communicating and playing it is increasingly difficult to take part in daily life without leaving a digital trace (Kitchin and Dodge 2011). For example, the Dutch Data Protection Authority estimates that the average Dutch citizen is recorded in 250-500 databases, with some in up to 1000 databases and growing (Koops 2011). These databases not only include individuals' digital footprints (data they themselves leave behind) but also individuals' data shadows (information about them generated by others). Those to whom the data refer often have little control over the data generated, their form, extent, or how they are used (CIPPIC 2006). Individually these databases provide limited views of people, but gain power when combined revealing detailed patterns and enabling what has been termed dataveillance -- the sorting and

sifting of datasets in order to identify, monitor, track, regulate, predict and prescribe (Clarke 1988; Raley 2013). The widespread generation of data and the practices of dataveillance raise many questions concerning privacy and rights of anonymity and confidentiality which are only just starting to be thought through and responded to (Solove 2006, Elwood and Leszczynski 2011).

Data have long been used to profile, segment and manage populations, but these processes have become much more sophisticated, fine-grained, widespread and routine with the application of data analytics employing machine learning techniques. Whilst the state might profile its citizens for the purposes of security and policing, commercial enterprises are seeking to reduce risk and maximize yield through more effective targeting of products. Whereas earlier generations of profiling sought to create aggregated population or area profiles, which then shaped decision making with regards to marketing and product placement (e.g., geodemographic profiling), new generation analytics can work at the level of the individual, combining data from various sources such as credit and store card transactions, clickstreams, social media posts, and other kinds of personal data to produce a detailed customer profile (Siegel 2013). These profiles are used to socially sort customers, identifying some for preferential treatment and excluding others, and to predict the likelihood that customers might be able to meet payments, or their projected lifetime value if they remain loyal, and how likely they are to move their custom (Graham 2005; Minelli *et al.* 2013). They are also being used to underpin new forms of dynamic and personalised pricing, tailored to a consumer's profile and purchase history, that are designed to leverage optimal spending (Tene and Polonetsky 2012). Consumers are thus being routinely measured and ranked, and receive differential services, based on their associated data and where they live.

One particularly pernicious form of predictive profiling is anticipatory governance. Here, predictive analytics are used to assess likely future behaviours or events and to direct appropriate action. Such anticipatory governance has been a feature of air travel for a number of years, with passengers profiled for risk and levels of security checks prior to starting their journey (Dodge and Kitchin 2004, Amoore 2006). More recently it has been extended to general policing, with it being used by a number of US police forces to identify potential future criminals and to direct the patrolling of areas based on an analysis of historical crime data, records of arrests, and the known social networks of criminals (Siegel 2013; Stroud 2014). In such cases, a person's data shadow does more than follow them; it precedes them, seeking to police behaviours that may never occur (Stalder 2002; Harcourt 2006). As a consequence, people are treated differently in anticipation of something they

may or may not do. Given their effects vis-a-vis individual lives and their black-boxed nature, the practices of predictive profiling, social sorting and anticipatory governance require much more attention, as do the companies that develop and undertake such tasks.

The work that data systems do in all of these cases is based on a generating an excess of data. Indeed, big data is premised on generating, hoarding and linking as much data as possible in the hope that value and insight can be leveraged from them. Rather than being generated and used to fulfil a specific task, data can be repackaged, sold and repurposed for all kinds of secondary uses. Such a strategy runs counter to the policy of data minimization, one of the foundations of privacy and data protection in the European Union and North America. This stipulates that data should only be generated and used to perform a particular task, and that they should only retained for as long as they are required to perform that task (Tene and Polonetsky 2012, CIPPIC 2012). A clear example of where the premise of data minimization is being breached is with respect to control creep, where data generated for one form of governance is appropriated for another (Innes 2001). This has mostly clearly occurred with respect to security, with airline industry and government administrative data being repurposed for profiling and assessing passenger risk (Lyon 2007). Given the implications to civil liberties from secondary data use, there is a need to examine its consequences and new approaches to data protection such as privacy by design (Cavoukian 2009).

Conclusion

Dalton and Thatcher (2014) conclude their call for critical data studies by setting out five questions that they believe require further study, all relating to big data:

- What historical conditions lead to the realization of 'big data' such as they are?
- Who controls 'big data', their production and analysis? What motives and imperatives drive their work?
- Who are the subjects of 'big data' and what knowledges are they producing?
- How is 'big data' actually applied in the production of spaces, places and landscapes?
- What is to be done with 'big data' and what other kinds of knowledges could it help produce?

There are many more questions that can be added to this list, not least by widening the lens to open data and data archives and repositories, but also to consider the wider data landscape,

data assemblages, and data markets. Rather than produce an extensive list of questions, we want to conclude by calling for greater conceptual work and empirical research to underpin and flesh out critical data studies.

The ways in which data are being generated, the analytics used to process and extract insight from them, the industries growing up around them, their wider political economic framing, and how they are employed all demand critical engagement. Whilst there is a rich and diverse tradition of critical social theory which can be directed towards data assemblages and the wider data landscape, such theory needs to be refined and fine-tuned to make sense of data and their work in the world, with new theory developed where needed. Yet we have barely begun to critically conceptualize data and their apparatus and elements. Such thinking needs to be complemented with more normatively orientated reflection on the ethics and politics of big data, open data, and data systems of different varieties.

Such conceptual and normative assessments needs to be accompanied by a diverse set of empirical case studies that examine all facets of data-driven governance, business and science and unpacks data assemblages and maps the wider data landscape. Our suggested approach is to employ methods such as ethnographies, interviews, focus groups, participant observation to delve into the workings of assemblages; to trace out genealogies of how the data landscape has changed over time and space; to map the materialities and infrastructures that constitute data infrastructures; and to deconstruct the discursive regime accompanying data-driven initiatives (Kitchin 2014a).

Undertaking this conceptual and empirical work is what our own research will focus on over the next few years as part of the *Programmable City* project (http://www.nuim.ie/progcity/), building on our initial large-scale studies (Lauriault 2012; Kitchin 2014a). This extensive project is examining the intersections of big and open data, ubiquitous computing, software and algorithms, and smart city developments in Dublin and Boston, unpacking a set of data assemblages and charting the data landscape of each city. We have no doubt that many others will be engaging in similar studies given the growth in data-driven driven forms of science, business and government. Hopefully, what this research will produce are a diverse set of vibrant critical data studies.

References

Amoore, L. (2006) Biometric borders: Governing mobilities in the war on terror. *Political Geography* 25: 336-351.

Bamford, J. (2009) *The Shadow Factory: The Ultra-Secret NSA from 9/11 to the Eavesdropping on America*. Anchor Books, New York.

Bard, K.A. and Shubert, S.B. (1999) *Encyclopedia of the Archaeology of Ancient Egypt*. Routledge, London.

Bowker, G. (2005) *Memory Practices in the Sciences.* MIT Press, Cambridge, MA.

Bowker, G. and Star, L. (1999) *Sorting Things Out: Classification and Its Consequences*. MIT Press, Cambridge, MA.

boyd, D. and Crawford, K. (2012) 'Critical questions for big data', *Information, Communication and Society,* 15(5): 662–79.

Cavoukian, A. (2009) *Privacy by Design: A Primer*. http://www.privacybydesign.ca/content/uploads/2013/10/pbd-primer.pdf (last accessed 17th July 2014)

CIPPIC (2006) *On the Data Trail: How detailed information about you gets into the hands of organizations with whom you have no relationship. A Report on the Canadian Data Brokerage Industr*y. Canadian Internet Public Policy Interest Clinic https://www.cippic.ca/sites/default/files/May1-06/DatabrokerReport.pdf . (last accessed 17 July 2014).

CIPPIC (2012) *Submissions to the House of Commons Standing Committee on Access to Information, Privacy & Ethics*. Canadian Internet Public Policy Interest Clinic https://cippic.ca/en/ETHI_hearings_on_Social_Media_Privacy. (last accessed 17 July 2014).

Clarke, R. (1988) Information Technology and Dataveillance. *Communications of the ACM*: 31(5): 498-512.

Constine, J. (2012) 'How big is Facebook's data? 2.5 billion pieces of content and 500+ terabytes ingested every day', *TechCrunch,* 22 August, http://techcrunch. com/2012/08/22/how-big-is-facebooks-data-2-5-billion-pieces-of-content-and- 500-terabytes-ingested-every-day/ (last accessed 17 July 2014).

Dalton, C. and Thatcher, J. (2014) What does a critical data studies look like, and why do we care? Seven points for a critical approach to 'big data'. *Space and Society Open Site*. http://societyandspace.com/material/commentaries/craig-dalton-and-jim-thatcher-what-does-a-critical-data-studies-look-like-and-why-do-we-care-seven-points-for-a-critical-approach-to-big-data/ (last accessed 17 July 2014)

Dodge, M. and Kitchin, R. (2004) Flying through code/space: the real virtuality of air travel, *Environment and Planning A* 36: 195-211.

Dodge, M. and Kitchin, R. (2005) 'Codes of life: identification codes and the machine-readable world', *Environment and Planning D: Society and Space*, 23(6): 851–81.

Dupaquier, M. and J. (1985), *Histoire de la démographie*, Librairie Académique Perrin, Paris.

Elwood, S. and Leszczynski, A. (2011) Privacy reconsidered: reconsidered: New representations, data practices, and the geoweb. *Geoforum* 42: 6–15.

Floridi, L. (2008) 'Data', in W.A. Darity (ed.), *International Encyclopedia of the Social Sciences*, 2nd edition. Detroit: Macmillan. Preprint online: http://www.philosophyofinformation. net/publications/pdf/data.pdf. (last accessed 17 July 2014)

Floridi, L. (2010) *Information: A Very Short Guide*. Oxford University Press, Oxford.

Foucault, M. (1977). The confession of the flesh. In C. Gordon (Ed.), (1980) *Power/Knowledge.* New York, Pantheon Books, pp. 194–228.

Gantz, J. and Reinsel, D. (2011) *Extracting value from chaos*. IDC, Framingham, MA http://idcdocserv.com/1142 (last accessed, 1st October 2012)

Garvey, E.G. (2013) "facts and FACTS": Abolitionists' database innovations. In Gitelman, L. (ed) *"Raw Data" is an Oxymoron.* MIT Press, Cambridge. pp 89-102.

Gitelman, L. and Jackson, V. (2013) 'Introduction', in L. Gitelman (ed.), *'Raw Data' is an Oxymoron*. MIT Press, Cambridge, MA, pp. 1–14.

Graham, S. (2005) Software-sorted geographies. *Progress in Human Geography* 29(5): 562-80.

Hacking, I. (1982) Biopower and the avalanche of numbers, in *Humanities in Society* 5(3&4): 279-295.

Hacking, I. (1986) Making up people, in Heller, T. et al. (eds) *Reconstructing Individualism*, , Stanford University Press, Stanford. pp. 222-236.Hacking, I. (1991), A tradition of natural kinds, *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 61(1/2) 109-126.

Hacking, I. (2001-2002), *Façonner les gens* (2001-2002), École de France Philosophie et histoire des concepts scientifiques (2001-2006), accessed May 10, 2011 from http://www.college-de-france.fr/media/historique/UPL46356_UPL35836_ihackingres0102.pdf.

Hacking, I. (2007), *Kinds of People, Moving Targets*, British Academy Lecture, Read at the Academy 11 April 2006, accessed September 13, 2011 from http://socio-logic.jp/2013-

hacking-sympo/reading/Hacking_2006_KindsofPeople_OCR.pdf, pp. 285-318. (last accessed 17 July 2014).

Hannah, M. (2011) *Dark Territory in the Information Age*. Ashgate, Farham.

Harcourt, B.E. (2006) *Against Prediction: Profiling, Policing and Punishing in an Actuarial Age*. Chicago University Press, Chicago.

Hecht, G. (2001) Technology, Politics, and National Identity in France, Chapter 8 in Technologies *of Power: Essays in Honor of Thomas Parke Hughes and Agatha Chipley Hughes*, eds. by Allen, Michael Thad and Gabrielle Hecht, MIT Press, pp. 145-174.

Hilbert, M. and López, P. (2011) The world's technological capacity to store, communicate, and compute information, *Science*, February 10, Vol. 331: 703-705

Innes, M. (2001) Control creep. *Sociological Research Online* 6(3), http://www.socresonline.org.uk/6/3/innes.html (last accessed 14th January 2014)

Kitchin, R. (2013) 'Big data and human geography: opportunities, challenges and risks', *Dialogues in Human Geography,* 3(3): 262–7.

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

Kitchin, R. (2014b) Short presentation on the need for critical data studies. *The Programmable City* blog. http://www.nuim.ie/progcity/2014/04/short-presentation-on-the-need-for-critical-data-studies/ (last accessed 17 July 2014).

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

Koops, B.J. (2011) Forgetting Footprints, Shunning Shadows: A Critical Analysis of the 'Right to Be Forgotten' in Big Data Practice. *SCRIPTed* 8(3): 229-256.

Laney, D. (2001) '3D data management: controlling data volume, velocity and variety', *Meta Group*, http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf. (last accessed 17 July 2014).

Lauriault, T.P. (2012) *Data, Infrastructures and Geographical Imaginations*. Ph.D. Thesis, Carleton University, Ottawa.

Lauriault, T.P., Craig, B.L., Taylor, D.R.F. and Pulsifier, P.L. (2007) Today's Data are Part of Tomorrow's Research: Archival Issues in the Sciences. *Archivaria* 64: 123–179.

Lauriault, T.P. (2014) Critical Analysis of the Irish Big Data Skills Report. Programmable City blog. http://www.nuim.ie/progcity/2014/05/big-data-series-critical-analysis-of-the-irish-big-data-skills-report/ (last accessed 17 July 2014)

Lyon, D. (2007) *Surveillance Studies: An Overview*. Polity, Cambridge.

Manyika, J., Chiu, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Hung Byers, A. (2011) *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. McKinsey Global Institute.

Marz, N. and Warren, J. (2012) *Big Data: Principles and Best Practices of Scalable Realtime Data Systems*. MEAP edition. Manning, Shelter Island, New York.

Mayer-Schonberger, V. and Cukier, K. (2013) *Big Data: A Revolution that will Change How We Live, Work and Think*. John Murray, London.

Minelli, M., Chambers, M. and Dhiraj, A. (2013) *Big Data, Big Analytics*. Wiley, Hoboken, NJ.

Open Data Center Alliance (2012) *Big Data Consumer Guide*. Open Data Center Alliance, http://www.opendatacenteralliance.org/library/odca-documents/big-data-consumer-s-guide/open-data-center-alliance-big-data-consumer-s-guide (last accessed 17 July 2014).

Pérez-Montoro, M.G. and Díaz Nafría, J.M. (2010) 'Data', in J.M. Díaz Nafría, M.G. Pérez-Montoro and F.S. Alemany (eds), *Glossary of Concepts, Metaphors, Theories and Problems Concerning Information*. León: Universidad de León. http://glossarium.bitrum.unileon.es/Home/dato/data (last accessed 17 July 2014).

Poovey, M. (1998) *A History of the Modern Fact: Problems of Knowledge in the Sciences of Wealth and Society*. University Chicago Press, Chicago.

Porter, T.M. (1986) *The Rise of Statistical Thinking*. New Jersey: Princeton University Press.

Raley, R. (2013) Dataveillance and countervailance, in Gitelman, L. (ed) *"Raw Data" is an Oxymoron*. MIT Press, Cambridge, pp 121-146.

Rial, N. (2013) 'The power of big data in Europe', *New Europe*, 24 May, http://www.neurope.eu/article/power-big-data-europe (last accessed 17 July 2014).

Ribes, D. and Jackson, S.J. (2013) 'Data bite man: the work of sustaining long-term study', in L. Gitelman (ed.), *'Raw Data' is an Oxymoron*. MIT Press, Cambridge, MA, pp. 147–66.

Rosenberg, D. (2013) Data before the fact. In Gitelman, L. (ed) *"Raw Data" is an Oxymoron*. MIT Press, Cambridge, pp 15-40.

Ruppert, E. (2012) 'The governmental topologies of database devices', *Theory, Culture Society*, 29: 116–36.

Siegel, E. (2013) *Predictive Analytics*. Wiley, Hoboken, NJ.

Short, J.E., Bohn, R.E. and Chaitanya, C. (2011) *How Much Information? 2010: Report on Enterprise Server Information*. Global Information Industry Center, UC San Diego. http://hmi.ucsd.edu/pdf/HMI_2010_EnterpriseReport_Jan_2011.pdf (last accessed 1st October 2012)

Smolan, R. and Erwitt, J. (2012) *The Human Face of Big Data*. Sterling, New York.

Solove, D.J. (2006) A Taxonomy of Privacy. *University of Pennsylvania Law Review* 154(3): 477-560.

Stalder, F. (2002) Privacy is not the antidote to surveillance. *Surveillance and Society* 1(1): 120-124.

Star, S.L. and Ruhleder, K. (1996) 'Steps Toward an Ecology of Infrastructure: design and access for large information spaces', *Information Systems Research,* 7(1): 111–34.

Stroud, M. (2014) The minority report: Chicago's new police computer predicts crimes, but is it racist? *The Verge*, http://www.theverge.com/2014/2/19/5419854/the-minority-report-this-computer-predicts-crime-but-is-it-racist (last accessed 17 July 2014)

Tene, O, and Polonetsky, J. (2012) Big Data for All: Privacy and User Control in the Age of Analytics. *Social Sciences Research Network.* http://ssrn.com/abstract=2149364 (last accessed 17 July 2014)

Zikopoulos, P.C., Eaton, C., deRoos, D., Deutsch, T. and Lapis, G. (2012) *Understanding Big Data*. McGraw Hill, New York.