Small data, data infrastructures and big data

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

The Programmable City

The Programmable City Working Paper 1

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

Abstract

The production of academic knowledge has progressed for the past few centuries using small data studies characterized by sampled data generated to answer specific questions. It is a strategy that has been remarkably successful, enabling the sciences, social sciences and humanities to advance in leaps and bounds. This approach is presently being challenged by the development of big data. Small data studies will, however, continue to be important in the future because of their utility in answering targeted queries. Nevertheless, small data are being made more big data-like through the development of new data infrastructures that pool, scale and link small data in order to create larger datasets, encourage sharing and re-use, and open them up to combination with big data and analysis using big data analytics. This paper examines the logic and value of small data studies, their relationship to emerging big data and data science, and the implications of scaling small data into data infrastructures, with a focus on spatial data examples. The final section provides a framework for conceptualizing and making sense of data and data infrastructures.

Key words

big data, small data, data infrastructures, data politics, spatial data infrastructures, cyber-infrastructures, epistemology

Small data, big data

Until a couple of years ago data were not considered in terms of being 'small' or 'big'. All data were, in effect, what is now sometimes referred to as 'small data' regardless of their size.

Due to factors such as cost, resourcing, and difficulties of generating, processing, analyzing and storing data, data were produced in tightly controlled ways using sampling techniques that limited their scope, temporality and size (Miller 2010). However, in the last decade or so, new technological developments have led to the production of what has been termed 'big data', which have very different characteristics to small datasets (see Table 1). As detailed in Kitchin (2013), big data are:

- huge in *volume*, consisting of terrabytes or petabytes of data;
- high in *velocity*, being created in or near real-time;
- diverse in *variety*, being structured and unstructured in nature, and often temporally and spatially referenced;
- *exhaustive* in scope, striving to capture entire populations or systems (n=all) within a given domain such as a nation state or a platform such as Twitter users, or at least much larger sample sizes than would be employed in traditional, small data studies;
- fine-grained in *resolution*, aiming to be as detailed as possible, 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 scalability (can expand in size rapidly).
  (Boyd and Crawford 2012; Dodge and Kitchin 2005; Marz and Warren 2012; Mayer-Schonberger and Cukier 2013).

Table 1: Comparing small and big data

Characteristic	**Small data**	**Big data**
Volume	Limited to large	Very large
Exhaustivity	Samples	Entire populations
Resolution and identification	Course & weak to tight & strong	Tight & strong
Relationality	Weak to strong	Strong
Velocity	Slow, freeze-framed/bundled	Fast, continuous
Variety	Limited to wide	Wide
Flexible and scalable	Low to middling	High

The term 'big' then is somewhat misleading as big data are characterized by much more than volume. Indeed, some 'small' datasets can be very large in size, such as national censuses that also seek to be exhaustive and have strong resolution and relationality.

However, census datasets lack velocity (usually conducted once every ten years), variety (usually c.30 structured questions), and flexibility (once a census is set and is being administered it is all but impossible to tweak the questions or add new questions or remove others and generally the fields are fixed, typically across censuses, to enable time-series analysis). Other small datasets also consist of a limited combination of big data's characteristics. For example, a qualitative dataset such as interview transcripts have strong resolution and flexibility but are usually relatively small in size (perhaps a couple of dozen respondents), have no or spaced-out temporality (one-off interviews or a sequence over a number of months), possess weak relationality, and are limited in variety (text transcripts). In contrast, big data have all these characteristics, or nearly all depending on their form (for example, sensor data are lacking in variety but have the other characteristics), with the crucial qualities being velocity and exhaustivity.

The rapid growth of big data has arisen due to the simultaneous development of a number of enabling technologies, infrastructures, techniques and processes, and their rapid embedding into everyday business and social practices and spaces. These include:

- the widespread roll-out of a diverse set of information and communication technologies, especially fixed and mobile internet;
- the embedding of software into all kinds of objects, machines and systems transforming them from 'dumb' to 'smart', as well as the creation of purely digital devices and systems;
- the development of ubiquitous computing and the ability to access networks and computation in many environments and on the move, including the creation of new social media platforms;
- advances in database design (especially the creation of NoSQL databases) and systems of information management;
- distributed and forever storage of data at affordable costs;
- and new forms of data analytics designed to cope with data abundance as opposed to data scarcity.

The practices of everyday life and the places in which we live are now augmented, monitored and regulated by dense assemblages of data-enabled and data-producing infrastructures and technologies, such as traffic and building management systems, surveillance and policing systems, government databases, customer management and logistic chains, financial and payment systems, and locative and social media. Within these socio-technical systems much of the data generation is automated through algoritmically-controlled cameras, sensors, scanners, digital devices such as smart phones, clickstreams, or are the by-product of networked interactions (such as the records of online transactions), or are volunteered by users through social media or crowdsourcing initiatives.

Collectively, such systems produce massive, exhaustive, dynamic, varied, detailed, indexical, inter-related, low cost per data point datasets that are flexible and scalable. To take just two examples as way of illustration. In 2011, Facebook's active users spent more than 9.3 billion hours a month on the site (Manyika *et al.* 2011), and by 2012 Facebook reported that it was processing 2.5 billion pieces of content (links, stores, photos, news, etc) and 500+ terabytes of data, 2.7 billion 'Like' actions and 300 million photo uploads *per day* (Constine 2012), each accompanied by associated metadata. Wal-Mart was generating more than 2.5 petabytes of data relating to more than 1 million customer transactions *every hour* in 2012 ("equivalent to 167 times the information contained in all the books in the Library of Congress"; Open Data Center Alliance 2012: 6).

Whereas small datasets were largely oases of data within data deserts, big data produce a veritable data deluge that seemingly enables research to shift from: data-scarce to data-rich; static snapshots to dynamic unfoldings; coarse aggregation to high resolution; relatively simple hypotheses and models to more complex, sophisticated simulations and theories (Kitchin 2013). This has led some to question whether big data might lead to the demise of small data or whether the stature of studies based on small data might be diminished due to their limitations in size, scope and temporality. For example, Sawyer (2008) notes that funding agencies are increasingly pushing their limited funding resources to data-rich areas, perhaps conflating data volume with insight, utility and value, and consigning research questions for which it is difficult to generate big data to a funding desert and marginal position within and outside the academy.

Such a re-prioritization, however, misunderstands both the nature of big data and the value of small data. Big data may seek to be exhaustive, but as with all data they are both a representation and a sample. What data are captured is shaped by:

- the field of view/sampling frame (where data capture devices are deployed and what their settings/parameters are; who uses a space or media, e.g., who belongs to Facebook or shops in Walmart);

- the technology and platform used (different surveys, sensors, lens, textual prompts, layout, etc. all produce variances and biases in what data are generated);
- the context in which data are generated (unfolding events mean data are always situated with respect to circumstance);
- the data ontology employed (how the data are calibrated and classified), and;
- the regulatory environment with respect to privacy, data protection and security (Kitchin 2013, 2014).

Indeed, all data provide oligoptic views of the world: views from certain vantage points, using particular tools, rather than an all-seeing, infallible god's eye view (Haraway 1991; Amin and Thrift 2002). As such, big data constitute a 'series of partial orders, localised totalities, with their ability to gaze in some directions and not others' (Latour cited in Amin and Thrift 2002: 92). Big data undoubtedly strive to be more exhaustive and provide dynamic, fine-grained insight but, nonetheless, their promise can never be fully fulfilled. Big data generally capture what is easy to ensnare -- data that are openly expressed (what is typed, swiped, scanned, sensed, etc.; people's actions and behaviours; the movement of things) -- as well as data that are the 'exhaust', a by-product, of the primary task/output. Tackling a question through big data often means repurposing data that were not designed to reveal insights into a particular phenomenon, with all the attendant issues of such a maneuver, for example creating ecological fallacies.

In contrast, small data may be limited in volume and velocity, but they have a long history of development across science, state agencies, non-governmental organizations and business, with established methodologies and modes of analysis, and a record of producing meaningful answers. Small data studies can be much more finely tailored to answer specific research questions and to explore in detail and in-depth the varied, contextual, rational and irrational ways in which people interact and make sense of the world, and how processes work. Small data can focus on specific cases and tell individual, nuanced and contextual stories. Small data studies thus seek to mine gold from working a narrow seam, whereas big data studies seek to extract nuggets through open-pit mining, scooping up and sieving huge tracks of land.

These two approaches of narrow versus open mining have consequences with respect to data quality, fidelity and lineage. Given the limited sample sizes of small data, data quality -- how clean (error and gap free), objective (bias free) and consistent (few discrepancies) the data are; veracity -- the authenticity of the data and the extent to which they accurately (precision) and faithfully (fidelity, reliability) represent what they are meant to; and lineage -- documentation that establishes provenance and fit for use; are of paramount importance (Lauriault 2012). Much work is expended on limiting sampling and methodological biases as well as ensuring that data are as rigorous and robust as possible before they are analyzed or shared. In contrast, it has been argued by some that big data studies do not need the same standards of data quality, veracity and lineage because the exhaustive nature of the dataset removes sampling biases and more than compensates for any errors or gaps or inconsistencies in the data or weakness in fidelity (Mayer-Schonberger and Cukier 2013). The argument for such a view is that "with less error from sampling we can accept more measurement error" (p.13) and "tolerate inexactitude" (p. 16). Viewed in this way, Mayer-Schonberger and Cukier (2013: 13) thus argue "more trumps better." Of course, this presumes that all uses of big data will tolerate inexactitude, when in fact many big data applications do require precision (e.g., finance data), or at least data with measurable error parameters.

Moreover, the warning "garbage in, garbage out" still holds. Big datasets that generate dirty, gamed or biased data, or data with poor fidelity, are going to produce analysis and conclusions that have weakened validity and deliver fewer benefits to those that analyze and seek to exploit them. And by dint of their method of production big data can suffer from all of these ails. The data can be dirty through instrument error or biased due to the demographic being sampled (e.g., not everybody uses Twitter) or the data might be gamed or faked through false accounts or hacking (e.g., there are hundreds of thousands of fake Twitter accounts seeking to influence trending and direct clickstream trails) (Bollier 2010; Crampton *et al*. 2012). With respect to fidelity there are question marks as to the extent to which social media posts really represent peoples' views and the faith that should be placed on them. Manovich (2011: 6) warns that "[p]eoples' posts, tweets, uploaded photographs, comments, and other types of online participation are not transparent windows into their selves; instead, they are often carefully curated and systematically managed."

There are issues of access to both small and big data. Small data produced by academia, public institutions, non-governmental organizations and private entities can be restricted in access, limited in use to defined personnel or available for a fee or under license. Increasingly, however, public institution and academic data are becoming more open. Big data are, with a few exceptions such as satellite imagery and national security and policing, mainly produced by the private sector. Access is usually restricted behind pay walls and proprietary licensing, limited to ensure competitive advantage and to leverage income through their sale or licensing (CIPPIC 2006). Indeed, it is somewhat of a paradox that only a handful of entities drowning in the data deluge (boyd and Crawford 2012) and companies such as mobile phone operators, app developers, social media providers, financial institutions, retail chains, and surveillance and security firms are under no obligations to share freely the data they collect through their operations. In some cases, a limited amount of the data might be made available to researchers or the public through Application Programming Interfaces (APIs). For example, Twitter allows a few companies to access its firehose (stream of data) for a fee for commercial purposes (and have the latitude to dictate terms with respect to what can be done with such data), but researchers are restricted to a ‘gardenhose’ (c. 10 percent of public tweets), a ‘spritzer’ (c. one percent of public tweets), or to different subsets of content (‘white-listed’ accounts), with private and protected tweets excluded in all cases (boyd and Crawford 2012). The worry is that the insights that privately owned and commercially sold big data can provide will be limited to the business sector, or maybe only opened to a privileged set of academic researchers whose findings cannot be replicated or validated (Lazer *et al*. 2009).

Given the concerns and limitations of big data, small data studies will continue to be an important component of the research landscape. Such data, however, will increasingly come under pressure to be scaled-up within digital data infrastructures in order that they are preserved for future generations, become accessible to re-use and combination with other small and big data, and more value and insight can be extracted from them through the application of big data analytics. Already, considerable resources have been invested in creating such data infrastructures. In the remainder of this paper we examine the scaling of small data into data infrastructures, the implications of such a scaling with respect to exposing small data to new big data epistemologies and repurposing, and provide a framework for conceptualizing and making sense of data infrastructures, focusing on spatial data examples.

Scaling, preserving, sharing and re-using small data: creating data infrastructures

Data have been collected together and stored for much of recorded history. Such practices have been both informal and formal in nature. The former consists simply of gathering data and storing them, whereas the latter consists of a set of curatorial practices and institutional structures designed to ensure that data are preserved for future generations. The former might best be described as data holdings, or backups, whereas the latter are data archives. Archives are formal collections of data that are actively structured, curated and documented, are accompanied by appropriate metadata, and where preservation, access and discoverability are integrated into technological systems and institutions designed to last the test of time (Lauriault *et al.* 2013). Archives explicitly seek to be long term endeavours, preserving the full record set -- data, metadata and associated documentation -- for future re-use.

The ability to store data digitally and to structure them within databases has radically transformed the volume of data that can be stored and efficiently and effectively handled and queried and has enabled the creation of extensive digital holdings and archives. Such digital data can be easily shared and re-used for a low marginal cost, although the cost of both the soft (institutional, policies, standards, human resources) and hard (technology, servers, software, delivery mechanisms, portals) infrastructures are not in the least bit inexpensive. Moreover, these data can be manipulated and analyzed by exposing them to computational algorithms. As such, procedures and calculations that would be difficult to undertake by hand or using analogue technologies become possible in just a few microseconds, enabling more and more complex analysis to be undertaken or the replication of objects (i.e. an atlas) and results. Further the data can also be relatively easily linked together and scaled into other forms of data infrastructure.

A data infrastructure is a digital means for storing, sharing and consuming data across networked technologies. Over the past two decades in particular, considerable effort has been expended on developing and promoting data access and discovery infrastructures, which take a number of forms: catalogues, directories, portals, clearinghouses and repositories (Lauriault *et al.* 2007). These terms are often used interchangeably and are confused for one another, though they are slightly different types of entities. Catalogues, directories and portals are centralized resources that may detail and link to individual data archives (e.g., Earth Observation Data Management Service of the Canada Centre for Remote Sensing) or data collections held by individual institutions (e.g. Australian National Data Service) or are federated infrastructures which provide the means to access the collections held by many (e.g., US National Sea Ice Data Center). They might provide fairly detailed inventories of the datasets held, and may act as metadata aggregators but do not necessarily host the data (e.g., GeoConnections Discovery Portal; Europeana) (O'Carroll *et al.* 2013). Single site repositories host all the data sets in a single site, accessible through a web interface, though they may maintain back-up or mirror sites in multiple locations (e.g., The UK Data Archive). A federated data repository or clearing house can be a shared place for storing and accessing data (e.g. US National Database for Autism Research (NDAR), NASA's Global Change Master Directory). It might provide some data services in terms of search and retrieval, and data management and processing, but each holding or archive has been produced independently and may not share data formats, standards, metadata, and policies. Nevertheless, the repository seeks to ensure that each archive meets a set of requirement specifications and uses audit and certification to ensure data integrity and trust amongst users (Dasish 2012).

A cyber-infrastructure is more than a collection of digital archives and repositories. It consists of a suite of dedicated networked technologies, shared services (relating to data management and processing), analysis tools such as data visualizations (e.g., graphing and mapping apps), and shared policies (concerning access, use, IPR, etc) which enable data to be distributed, linked together and analyzed (e.g. a spatial data infrastructure). Whilst it is sometimes used to denote the infrastructure that enables a federated repository to function, here we use it to denote a data infrastructure in which data share common technical specifications relating to formats, standards, and protocols. In other words, there are strong rules relating to data standardization and compliance within the infrastructure. Such cyber-infrastructures include those implemented by national statistical agencies and national spatial data infrastructures that require all data stored and shared to comply with defined parameters in order to maximize data interoperability and ensure data quality, fidelity and integrity that promotes trust. The objectives of spatial data infrastructures are to ensure that users from multiple sectors and jurisdictions can seamlessly re-use these data and link them into their systems. A cross border natural disaster, for instance, would require multiple agencies, in different countries along with sub-national entities, under severe time constraints and pressures, to access, model and visualize spatial data in near real time while also inputting newly acquired data to respond to and inform an emergency response arena. In less stressful environments, SDIs enable the management of cross border shared services and natural resources (e.g. EU Water Framework Directives).

The arguments for the storing, sharing and scaling of small data

The arguments for building repositories and data infrastructures centre on the promises of new discoveries and innovations through the combination of datasets and the crowdsourcing of minds. Individual datasets are valuable in their own right, but when combined with other datasets or examined in new ways fresh insights can potentially be discerned and new questions answered (Borgman 2007). By combining datasets, it is contended that the cumulative nature and pace of knowledge building is accelerated (Lauriault *et al.* 2007). Moreover, by preserving data over time it becomes possible to track trends and patterns, and the longer the record, the greater the ability to build models and simulations and have confidence in the conclusions drawn (Lauriault *et al.* 2007). Over time then, the cumulative value of data infrastructures increases as the data become more readily and broadly available, both in scope and temporality. Such a sharing strategy is also more likely to spark new interdisciplinary collaborations between researchers and teams and to foster enhanced skill through having access to new kinds of data (Borgman 2007). Moreover, the sharing of data and the adoption of infrastructure standards, protocols and policies increases data quality and enables third party data and study verification, thus increasing data integrity (Lauriault *et al.* 2007).

The financial benefits of data infrastructures centre on the scales of economy created by sharing resources and avoiding replication, the leveraging effects of re-using costly data, the generation of wealth through new discoveries, and producing more efficient societies. Research and the production of administrative, statistical and geomatics data are typically costly undertakings, with various funding agencies collectively spending billions of dollars every year to fund research activities. Rather than creating a plethora of ad hoc archives, it makes more sense to establish a smaller number of dedicated institutional repositories or infrastructures which undertake basic data standardization and produce significant efficiencies in effort, as well as enable broader access to data for individual researchers/institutions where entry costs to a field would normally be prohibitive (Fry *et al.* 2008). As well as reducing wastage, preserving and sharing the fruits of such endeavours is more likely to maximize the return on investment by enabling as much value as possible to be extracted from the data (Lauriault *et al.* 2007). That said, the sustainability of these infrastructures are often an issue. Research data infrastructures are funded through a mix of mechanisms such as state and research funds, community organized infrastructures rely on small grants and membership fees, while the open data infrastructures run by civil society organizations are built by volunteers.

Examples of data infrastructures

Given the anticipated gains from sharing data, over the past three decades supranational bodies such as the European Union, national governments, researchers, philanthropic and civil society organizations, have invested extensively in funding a wide variety of data and cyber-infrastructure initiatives.

Spatial data infrastructures (SDIs), are the achetype cyber-infrastructure. National scale SDIs are normally institutionally located in national mapping organizations, national surveys, or the departments that manage natural resources. They are an assemblage of institutions (e.g., government, geomatics), policies (e.g., data sharing protocols), laws (e.g. licenses, legislation, regulation), technologies (e.g., data portals, storage, software), processes (e.g., web mapping, metadata aggregation), standards (e.g., metadata, file transfer, data quality) and specifications (e.g., interoperability), scientific and computing knowledge, skilled human resources, discovery and access portals, framework data (e.g., common datasets upon which others can build such as road networks) and mapping services that direct the who, how, what and why geospatial data are collected, stored, manipulated, analyzed, transformed and shared. They are inter-sectoral, cross-domain, trans-disciplinary, interdepartmental, and require much consensus building. Supranational SDIs such as the Infrastructure for Spatial Information in the European Community (INSPIRE), are very similar to national SDIs, however in the case of INSPIRE, it governs how nations are to construct their infrastructures via rules, directives and policies, that will lead to data, geomatics systems and services being seamlessly interoperable across 27 member states. In addition, INSPIRE includes a GeoPortal which is a federated catalog that aggregates the metadata of member state SDIs thus providing users with a single point to discover and view EU geospatial data.

On a smaller scale, and in a different domain, the UK Data Archive, is an example of a research data infrastructure that acquires, curates and provides access to social science and humanities data. Data are discovered via the UK Data Service which is a catalogue that provides access to hosted national and international survey data collections, international databanks, census data and qualitative data. Secure data services for access and use of more sensitive research data are also provided. Data are described with standard metadata, and a number of educational resources are provided for users to work with the data once they have been downloaded. The UK Data Archive, although, not a certified trusted digital repository, has as its objective to maintain its large collections of data for long-term reuse, and provides a number of capacity building resources to enable researchers to manage and deposit their data.

There are not many examples of data infrastructures in the non-profit and charitable sector. The Canadian Council on Social Development, Community Data Program (CDP), is however an example of a small data infrastructure created for the specific purpose of enabling small area, evidence based decision making in the social sector. It is funded by its members through a consortia model. Members are city based networks of municipal administrators, school boards, community health centres, social planning councils and a number of charitable and non-profit organizations. The CDP acquires and disseminates mostly public sector data and custom ordered cross tabulated data aggregated into neighbourhood, city ward, small area census geographies and postal codes. These are stored into a database and delivered to members via an online catalog. In this instance, members not only benefit from the data, but also from services where experts negotiate data acquisition based on community needs and specifications, and a knowledge sharing network between super users and novices.

Finally, in the last four years a number of open data infrastructures have been created by national governments, sub national governments such as cities, provinces, counties and states, and civil society organizations such as the UK-based Open Knowledge Foundation (OKNF) and to a lesser extent research and private sector entities. The objectives of these data infrastructures are to unlock access to public sector datasets and make them accessible via a discovery and access portals for free and under open licences. The OKNF is an open data supranational organization which provides direction to governments and civil society groups and helps build capacity in terms of the deployment of catalogs (e.g., CKAN), and has created a set of open data principles and open license specifications. Open data portals have not yet matured into a cyber-infrastructures, although government funded open data portals do manifest some of their qualities. Unlike SDIs, these are not grounded in a domain, discipline or the sciences, and often open data infrastructures are administered in information management/technology departments and championed by chief technology officers, or are created and supported by volunteers groups composed of new media enthusiasts and app developers.

These four cases are but a small sample of the innumerable data and cyber-infrastructures currently in operation. In all four cases, the data found in their portals are small data, SDIs being the exception as remote sensing data and many environmental sensors produce data that have the qualities of big data. Alternatively, geodemographic data infrastructures, discussed later, exemplify the scaling of small data with big data.

Implications of scaling small data into data infrastructures

Whilst the scaling of small data into data infrastructures does not create big data, in the sense that the data still lack velocity, it does make them more big data-like by making them more extensive, relational and interconnected, varied, and flexible. This enables two effects to occur. First, it opens scaled small data to new epistemologies and, in particular, to new forms of big data analytics. Second, it facilitates small data being conjoined with big data to produce more complex, inter-related and wide-ranging data infrastructures that are presently driving the rapid growth of commercial data brokers, including the burgeoning geodemographics industry (also known as locational targeted niche marketing tools). Both have consequences with respect to how small data are being used and raise normative questions concerning the creation and use of data infrastructures.

New epistemologies

Traditional small data methods of analysis have primarily been designed to extract insights from scarce, static, clean and weak relational data sets that have been sampled and adhere to strict assumptions (such as independence, stationarity, and normality), and were generated and analyzed with a specific question in mind (Miller 2010). The challenge with big data is to cope with abundance and exhaustivity (including sizable amounts of data with low utility and value), timeliness and dynamism, messiness and uncertainty, high relationality, semi-structured or unstructured content, and the fact that much of them are generated with no specific question in mind or are a by-product of another activity. The solution has been new data analytics that utilize the power of algorithms and computation to process and provide insight into datasets that would simply be too costly, difficult and time-consuming to analyze otherwise. Such analytics scale-up existing statistical methods, such as regression, model building, data visualization and mapping, as well as employing new machine learning and visual analytics techniques that computationally mine meaning from data and detect, classify and segment meaningful patterns, relationships, associations and trends between variables, and build predictive, simulation and optimization models (Han *et al.* 2011).

Machine learning generally consists of two broad types: supervised (using training data) and unsupervised (using self-organization). In supervised learning, a model is trained to match inputs to certain known outputs (Hastie *et al.* 2009). For example, the model might be trained to match patterns on an aerial photograph with building shapes, or to predict a certain outcome. In contrast, in unsupervised learning the model seeks to teach itself to spot patterns and find structure in data without the use of training data. In general, this is achieved through identifying clusters and relationships between the data where the characteristics of similarity or associations were not known in advance. For example, the model might learn how to segment customers into self-similar groups and to predict purchases amongst those groups (Han *et al.* 2011). In both cases, the model is created through a learning process shaped by learning rules and weightings that direct how the model is built in relation to the data (Hastie *et al.* 2009). The process of building the model starts with a simple construction and then tweaks it repeatedly using the learning rules, as if applying ‘genetic mutations’, until it evolves into a robust model (Siegel 2013: 122). Using a machine learning approach hundreds of different types of models can be applied to a dataset in order to determine which best explain or perform optimally. Indeed, an ensemble approach builds multiple models using a variety of statistical techniques (e.g. regression, neural network, nearest neighbour, factor analysis, and decision tree models) to predict the same phenomena, rather than selecting a single approach and building a handful of models (Siegel 2013). These data analytics can equally be applied to scaled small data to extract and model insights.

Data analytics are reflective of a particular way of making sense of the world; they are the manifestation of a particular epistemology. Some envisage them as a new form of empiricism that enables data to speak for themselves free of theory. For example, Anderson (2008) argues that "the data deluge makes the scientific method obsolete". He continues, "We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot... Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all. There's no reason to cling to our old ways." In other words, rather than testing whether certain hypothesized patterns or relationships exist within a dataset, algorithms are set to work on big data to discover meaningful associations between data without being guided by hypotheses. In this epistemological vision, scaled small data are made sense of through a purely inductive approach.

In contrast, data-driven science seeks to hold to the tenets of the scientific method, but uses a combination of abductive, inductive and deductive approaches to advance the understanding of a phenomenon. It differs from the traditional deductive approach in that it seeks to generate hypotheses and insights 'born from the data' rather than 'born from the theory' (Kelling *et al.* 2009: 613). In other words, it seeks to incorporate a mode of induction into the initial stages of the research design, though explanation through induction is not the intended end-point. This process of induction does not arise from nowhere, but is situated and contextualized within a highly evolved theoretical domain. The patterns, associations and trends identified through initial data analytics are thus used to identify potential hypotheses worthy of further examination and testing. As such, the epistemological strategy adopted within data-driven science is to use guided knowledge discovery techniques to identify valuable insights that traditional 'knowledge-driven science' might fail to spot and then to investigate these further (Kelling *et al.* 2009; Miller 2010; Loukides 2010).

With respect to the social sciences and humanities, data infrastructures, new data analytics and associated epistemologies offer the potential to transform the research landscape. As noted, data infrastructures provide access to large collections of data for re-use and analysis. These data can be conjoined in new ways and the relationships and associations between them explored using data analytics. With respect to structured data, it becomes possible to produce more refined and sophisticated models and to test the veracity of these models across a multitude of groups, settings and situations (Lazer *et al*. 2009). This includes the production of more elaborate and robust spatial models (Batty 2013). Access to unstructured data multiplies, including both new sources of information (e.g., social media) and many which have heretofore been difficult to access (e.g., millions of books, documents, newspapers, photographs, art works, and material objects; Cohen 2008). These data are thus opened up to the power of computation, including sophisticated tools for handling, searching, linking, sharing and analyzing data that seek to complement and augment existing humanities methods and traditional forms of interpretation and theory building (Berry 2011; Manovich 2011), as well utilizing new data analytics that provide new means to make sense of such data (Moretti 2005).

Such approaches are not without critique, with detractors arguing that data analytics are mechanistic, reductionist, functionalist, and parochial, reducing diverse individuals and complex, multidimensional social structures to mere data points (Wyly, 2013), thus fostering weak, surface analysis, rather than deep, penetrating insight; that they sacrifice specificity, context and depth for scale, automation and breadth. Indeed, Brooks (2013) contends that data analytics: struggle with the social (people are not rational and do not behave in predictable ways; human systems are incredibly complex, having contradictory and paradoxical relations); and with context (data are largely shorn of the social, political and economic and historical context); create bigger haystacks (consisting of many more spurious correlations making it difficult to identify needles); have trouble addressing big problems (especially social and economic ones); favour memes over masterpieces (identifies trends but not necessarily significant features that may become a trend); and obscure values (of the data producers and those that analyze them and their objectives). Such debates over the value and appropriateness of new analytics and epistemologies, and their application to scaled small data, seem set to continue for the foreseeable future.

New hybrid small and big data infrastructures

Scaled small data also gain in value as a commodity, especially when they can be conjoined with big data. In contrast to academic, research-orientated or governmental data infrastructures, data brokers (sometimes called data aggregators, consolidators or resellers) gather together data into privately held infrastructures for re-sale on a for-profit basis. They source data from both public and private sources. For example, from public sector sources they gather data relating to individuals and aggregates (e.g., groups, places) concerning health, education, crime, property, travel, environment, etc., matching these with private sector data related to or captured within retail, financial, logistics, business intelligence, real estate, private security, political polling, transportation, media, and so on. The potential to link data across domains is high. For example, the Dutch Data Protection Authority estimates that the average Dutch citizen is included in 250-500 databases, with more socially active people included in up to 1000 databases (Koops 2011). More recently, data brokers have been combining these data with the metadata and content from locative (e.g., smart phone apps) and social media (e.g. Twitter and Facebook). For example, Facebook is partnering with large data brokers and marketers in order to merge together the profiles, networks and uploaded content of its billion users (their likes, comments, photos, videos, etc) with non-Facebook purchasing and behaviour data (Edwards 2013). Such data form a set of vast relational data infrastructures. Acxiom (one of Facebook's partners) is reputed to have constructed a data infrastructure concerning 500 million active consumers worldwide (about 190 million individuals and 126 million households in the United States), with about 1,500 data points per person, its servers processing over 50 trillion data transactions a year and its turnover exceeding one billion dollars (Singer 2012). It also manages separately customer databases for or works with 47 of the Fortune 100 companies (Singer 2012).

These vast relational data infrastructures can be used to construct a suite of derived data products, wherein value is added through integration, and may be subjected to data analytics to profile individuals, groups and places, and to predict what people might do under different circumstances. In the main, profiles are used to micro-target advertising and niche marketing campaigns, assess how such targets might behave and be nudged into a particular response (e.g., selecting and purchasing a particular item), assess credit worthiness and socially sort individuals (determine whether one might receive a service or set personalized pricing), and provide detailed business analytics, whilst reducing their overheads in terms of wastage and loss through risky investments (Lyon 2002, Graham 2005, Siegel 2013). Acxiom, for instance, seeks to mesh offline, online and mobile data in order to create a '360-degree view' of consumers, using these data to create detailed profiles and robust predictive models which it sells to interested parties (Singer 2012).

Geodemographic segmentation is a data analytical process which can combine both small and big data in order to create quantitatively based classification systems of groups of people at a particular geographic unit of analysis, often at postal code geographies. Once classification systems are developed, primarily with small data inputs, big data such as purchasing histories, which use postal codes as unique identifiers, can be matched to these classifications to assess consumption patterns and to refine the groupings. These data infrastructures, while they can be used to better understand population dynamics in cities, are mostly developed by the private sector to geo-target marketing. As an illustration, the Environics Analytics PRiZMc2 Lifestyle segmentation tool classifies Canadians into 66 lifestyle types such as 'cosmopolitan elite' or 'Les Chics and Lunch at Tim's' (short for Tim Horton Donuts) "based on their demographics, marketplace preferences and psychographic Social Values". This company also produces a product called Wealth$capes Dollar and Sense which provides marketers with a similar service (Environics Analytics, 2013). The algorithms, methodological assumptions and the mix of datasets used to produce the geodemographic profiles are proprietary and protected by intellectual property regimes and are not subject to public scrutiny. Irrespective, by using such products companies seek become more effective and efficient in their operations with respect to targeting customers and siting stores.

The scaling of small data, mashing them with big data, and subjecting them to data analytics, can have profound implications for citizens and the services and opportunities extended to them. The worry for some is that a form of 'data determinism' is being practiced in which individuals are not profiled and judged just on the basis of what they have done, but on the prediction of what they might do in the future (Ramirez 2013). A new probability market is emerging – although gambling industry odds compilers and security markets have been around for some time – which constitutes a new phase in the era of probabilistic thinking (Hacking 1975, 1990), one that is making up new kinds of people (Hacking 2007) and new kinds of places (Lauriault 2012), led by the private sector and surveillance institutions, mostly for the purpose of marketing products and security. Moreover, there are concerns to the extent to which scaled small data data infrastructures facilitate dataveillance (surveillance enacted through the processing and analyzing of data records), infringe on privacy and other human rights, affect access to private health insurance and its rates, stigmatize and redline areas, pose significant data security concerns with regards to data being stolen and exploited criminally, and enable control creep wherein data generated for one purpose is used for another (Clarke 1988, Innes 2001, Solove 2006, CIPPIC 2006). Citizens may also have not agreed with the entities producing the data as to how data about themselves are used (CIPPIC 2006). As such, whilst scaling small data does offer a number of benefits they also can have differential and negative consequences. There are thus a number of fundamental normative questions that need urgent reflexive consideration concerning the production of data infrastructures if we are to maximize their benefits whilst minimizing their more pernicious effects.

Making sense of data and data infrastructures

As the last section makes clear, the scaling of small data has many potential implications. It is thus important to think critically about the nature of data, databases and data infrastructures, their socio-technical production, and how they reflect rationalities about the world at the same time as they reproduce and reinforce such rationalities. In this section, we provide a framework for conceptualizing and making sense of data infrastructures.

Data are often understood as the raw material produced by abstracting the world into categories, measures and other representational forms; they are pre-analytical, pre-factual and rhetorical in nature; that is, they are simple, straightforward, and unadorned, pre-existing argument or interpretation that converts them to facts, evidence and information; they speak for themselves (Rosenberg 2013). From this perspective, data are understood as being benign and lacking in political ideology. Likewise, the algorithms used to process these data are viewed as neutral and non-ideological in their formulation and operation, grounded in scientific objectivity (Kitchin and Dodge 2011). As already intimated, however, data are more complicated than that. Data do not exist independently of the ideas, techniques, technologies, people and contexts that conceive, produce, process, manage, analyze and store them (Bowker and Star 1999; Lauriault 2012; Ribes and Jackson 2013). As Gitelman and Jackson (2013: 2, following Bowker) put it, "raw data is an oxymoron"; "data are always already 'cooked' and never entirely 'raw'." What data are generated is the product of choices and constraints, shaped by a system of thought, models and methodologies, techniques and technical know-how, public and political opinion, ethical considerations, the regulatory environment, and funding and resourcing. Data then are situated, contingent, relational, and framed and used contextually to try and achieve certain aims and goals (Poovey 1998; Latour 1987, Hacking 1982, Anderson 1991).

Similarly, databases and data infrastructures are not simply neutral, technical means of assembling and sharing data; and they are not merely products that store captured data about the world, 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 and pursue technopolitics" (Hetch, 2001:257). They are essential tools in the production of knowledge, governance and capital.

Databases are designed and built to hold certain kinds of data and enable certain kinds of analysis. How they are structured has profound consequences as to what queries and analysis can be performed conditioning the work that can be done on and through them (Ruppert 2012). Databases then are not neutral containers; they are expressions of knowledge/power and they enact and reproduce such relations (Ruppert 2012) and are biopolitical objects (Foucault 2010). At the same time, databases unmoor data analysis from the data by enabling complex queries and calculations without those conducting such analyses having to peruse and work the data themselves or even understand how the data have been compiled and organized (Gitelman and Jackson 2013). This unmooring is aided by techniques such as standardization of formats and metadata and works to decontextualize and depoliticize the data contained within (Wilson 2011). Importantly, such unmooring enables the power/knowledge of the database to travel and be deployed by others shorn of its complex inner workings and history and politics of production. Although, if data curation and preservation are properly practiced, databases are accompanied by methodological guidebooks and use instructions which seek to reduce the effects of this unmooring.

Data infrastructures host and link databases into a more complex data assemblage. As with databases, there is nothing inherent or given about how such archiving and sharing structures are composed. Indeed, the design and management of data infrastructures are riddled with technical and political challenges that are tackled through messy and contested negotiations that are contextualized by various agendas and governmentalities. The solutions created in terms of standards, protocols and policies inherently have normalizing effects in that they seek common shared ground and to universalize practices amongst developers and users (Lauriaut 2012), glossing over and ameliorating the tension between enabling interoperability and limiting customization and constraining innovation (Star and Ruhleder 1996). Given these tensions, normalizing processes have to constantly and recursively be reaffirmed through implementation, management and system governance (Star and Lampland 2009).

As Dourish and Bell (2007) contend, databases and infrastructures then cannot be considered in purely instrumental terms as they are thoroughly cultural, economic and cognitive in nature and steeped in social significance. They thus suggest two lenses through which to understand data infrastructures. The first is a sociopolitical reading which examines them as "crystallizations of institutional relations" (p.416). The second perspective is an experiential reading that examines "how they shape individual actions and experience" (p.417). In both cases, data infrastructures are understood as relational entities. This relationality reshapes the world contingently around it, as it in turn is shaped by the world. As we come to use and rely on databases and data infrastructures to make sense of and do work in the world, our discursive and material practices adapt and mutate in response to them (Star and Ruhleder 1996). The world is not just reflected in data, it is changed by them; "the work of producing, preserving, and sharing data reshapes the organizational, technological, and cultural worlds around them" (Ribes and Jackson 2013:147).

In other words, databases and data infrastructures do not simply support research, they fundamentally change the practices and organization of research -- the questions asked, how they are asked, how they are answered, how the answers are deployed, and who is conducting the research and how are they operate as researchers. For example, in her study of the evolution of the Canada Census and the Atlas of Canada, Lauriault (2012) details how each has developed recursively and iteratively based on models of the world which construct ways to imagine and produce Canada. She argues that the data infrastructures and the data themselves constitute an institutional "extrasomatic memory system that allows for the telling of stories about the nature of Canada ... [through] maps, graphs, models and statistics which rely on sensors, data, interoperability and web mapping standards, portals, metadata and models, science, and open architectures" (p.27). In turn, these stories modulate the underlying models and thus the data infrastructure mutates, inflecting the means through which the stories are created.

Making sense of data, databases and data infrastructures then requires carefully unpacking and deconstructing their always emerging, contingent, relational and contextual nature (Star and Ruhleder 1996). As Lauriault (2012) argues this also requires a genealogical analysis that documents how databases and data infrastructures develop over time and space. This kind of in-depth, historically rich deconstruction of the processes, practices and political economy of data infrastructures has been largely neglected to date, despite the fact that data and how they are handled underpins and explicitly shapes scientific endeavour, large components of governance, and the work of institutions and companies.

Conclusion

We are presently witnessing a fast changing landscape with respect to data. Not only are we witnessing the roll-out of a new form of data in the guise of big data, but traditional, small data are evolving through new data infrastructures that enable them to be scaled and analyzed in new ways. In this paper we have compared small and big data before going on to examine how small data are being scaled, combined with big data, and being made amenable to big data analytics. Our argument has been fourfold.

First, despite the rapid growth of big data and associated new analytics, small data will continue to be a vital part of the research landscape. There will not be a paradigm shift in the near future in which studies using big data replace those employing small data, rather small and big data will complement one another; mining narrow seams of high quality data will continue alongside open pit mining because it enables much more control of the research design and to answer specific, targeted questions. As such, rather than directing research funding to projects that have access to vast quantities of data in the hope that they will inherently produce useful insights, funding needs to be focused on answering critical questions, whether they are tackled using small or big data (Sawyer 2008).

Second, the small data landscape is changing through the development of data infrastructures. Small data gain value and utility when made accessible for re-use and are combined with other data sets. As a consequence, much effort is being directed at building such infrastructures and in trying to harmonize small data, with respect to data standards, formats, metadata, and documentation, to ensure their compatibility with systems, maximize discoverability, and facilitate the linking together of data sets. The pressure to harmonize, share and re-use small data will continue to grow as research funders seek to gain the maximum return on their investment through new knowledge and innovations.

Third, the scaling of small data into data infrastructures has three consequences. One: by pooling and linking small data to create larger, interconnected data sets, small data are opened up to analysis by big data analytics. Small data are thus exposed to the new epistemologies of data science, fostering the growth of new approaches such as the digital humanities and computational social sciences. Two: small data are more easily conjoined with big data to produce more diverse derived data that enables more wide-ranging and extensive analysis. This reconfiguration of the data landscape is facilitating the rapid growth of data brokers and new data products, including detailed profiling. Three: the scaling of small data, and their combination with big data and exposure to big data analytics, produces a set of potential pernicious effects such as dataveillance, social sorting, control creep, and anticipatory governance that impinge on privacy, social freedoms and have structural consequences for individual lives. As such, the scaling of small data raises normative questions concerning how data should be managed and utilized. We have barely begun to examine these consequences, with developments running ahead of critical and normative reflection and political, policy and legal reaction.

Fourth, whilst much theoretical attention has been focused on the derivatives of data, information and knowledge, data themselves have been relatively neglected from a conceptual and philosophical standpoint. Instead, attention has largely been technical and how best to generate and analyze data to leverage insight, rather than to consider their nature. Given the rapidly changing data landscape, the growing importance of evidence-based management and governance, and rise of data-driven science it is important to think critically about the nature of data and data infrastructures. Our suggestion, drawing from the nascent theoretical work in the literature, is to conceive of data as being situated, contingent, relational and contextual, and to understand data infrastructures as socio-technical assemblages composed of many elements and shaped by governmentalities, political economy and other processes.

Small data are set to continue being an important component of research endeavours. However, they are in the process of taking on new forms that have consequences for how we think about and utilize such data. We have made an initial attempt to detail some of these transformations, but further critical reflection and normative thinking is required to make sense of the changes taking place and their implications.

Acknowledgements

The research for this paper was funded by a European Research Council Advanced Investigator award (ERC-2012-AdG-323636-SOFTCITY).

References

Amin, A. and Thrift, N. (2002) *Cities: Reimagining the Urban*. Polity, London.

Anderson, B. (1991) *Imagined Communities*. Revised Edition, New York: Verso.Anderson, C. (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired*, June 23, 2008, http://www.wired.com/science/discoveries/magazine/16-07/pb_theory (last accessed October 12th 2012)

Batty, M. (2013) *The New Science of Cities*. MIT Press, Cambridge, MA.

Berry, D. (2011) The computational turn: Thinking about the digital humanities. *Culture Machine* 12. http://www.culturemachine.net/index.php/cm/article/view/440/470 (last accessed 3rd December 2012)

Bollier, D. (2010) *The Promise and Peril of Big Data*. The Aspen Institute. http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf (last accessed, 1st October 2012)

Borgman, C.L. (2007) *Scholarship in the Digital Age*. MIT Press, Cambridge, MA.

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

boyd, D. and Crawford, K. (2012) Critical questions for big data. *Information, Communication and Society* 15(5): 662-679

Brooks, D. (2013b) What data can't do. *New York Times,* 18 February 2013http://www.nytimes.com/2013/02/19/opinion/brooks-what-data-cant-do.html (last accessed 18 February 2013)

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

Clarke, R. (1988) Information Technology and Dataveillance. *Commun. ACM* 31,5 (May 1988) 498-512

Cohen, D. (2008) Contribution to: The Promise of Digital History (roundtable discussion), *Journal of American History* 95(2): 452-491

Constine, J. (2012) How Big Is Facebook's Data? 2.5 Billion Pieces Of Content And 500+ Terabytes Ingested Every Day, 22 August 2012, 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 28 January 2013)

Crampton, J., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M.W. and Zook, M. (2012) *Beyond the Geotag? Deconstructing "Big Data" and leveraging the Potential of the Geoweb.* http://www.uky.edu/~tmute2/geography_methods/readingPDFs/2012-Beyond-the-Geotag-2012.10.01.pdf (last accessed 21 February 2013)

Dasish (2012) Roadmap for Preservation and Curation in the Social Sciences and Humanities. http://dasish.eu/publications/projectreports/D4.1_-_Roadmap_for_Preservation_and_Curation_in_the_SSH.pdf/ (last accessed 15 October 2013)

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 – 881

Dourish, P. and Bell, G. (2007) The infrastructure of experience and the experience of infrastructure: meaning and structure in everyday encounters with space. *Environment and Planning B* 34: 414-430

Edwards, J. (2013) Facebook Is About To Launch A Huge Play In 'Big Data' Analytics. Business Insider, May 10th http://www.businessinsider.com/facebook-is-about-to-launch-a-huge-play-in-big-data-analytics-2013-5 (last accessed 18 September 2013).

Environics Analytics (2013) Wealth$capes: Dollars and Sense, http://www.environicsanalytics.ca/environics-analytics/data/financial-data/wealthscapes (last accessed 26 November 2013)

Environics Analytics (2013) PRiZMc2 Segmentation Lifestyle Lookup tool, http://www.environicsanalytics.ca/prizm-c2-cluster-lookup. (last accessed 26 November 2013)

Foucault, M. (2010) Biopower, in Paul Rabinow, Ed. *The Foucault Reader*, New York: Vintage Books. Pp. 257-273.

Fry, J., Lockyer, S., Oppenheim, C., Houghton, J.W. and Rasmussen, B. (2008) *Identifying benefits arising from the curation and open sharing of research data produced by UK Higher Education and research institutes*, JISC, London and Bristol. http://repository.jisc.ac.uk/279/

GeoConnectins (2001) Canadian Geospatial Data Infrastructure - Architecture Description, Version 1, http://geoconnections.nrcan.gc.ca/18.

GeoConnectins (2005) Canadian Geospatial Data Infrastructure - Architecture Description, Version 2, http://geoconnections.nrcan.gc.ca/18.

Gitelman, L. and Jackson, V. (2013) Introduction in Gitelman, L. (ed) “Raw Data” is an Oxymoron. MIT Press, Cambridge. pp 1-14.

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

Hacking, I. (1975) *The Emergence of Probability*. Cambridge: Cambridge University Press

Hacking, I. (1982) Biopower and the Avalanche of Numbers, in *Humanities in Society: Foucault and Critical Theory: The Uses of Discourse Analysis*, Vol.5, Numbers 3&4, pp. 279-295.

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

Hacking, I. (2007) *Kinds of People, Moving Targets*, British Academy Lecture, Read at the Academy 11 April 2006, accessed September 13, 2011 from http://www.proc.britac.ac.uk/tfiles/151p285.pdf, pp. 285-318.

Han, J., Kamber, M. and Pei, (2011) *Data Mining: Concepts and Techniques*. 3rd edition. Morgan Kaufmann, Waltham, MA.

Haraway, D. (1991) *Simians, Cyborgs and Women: The Reinvention of Nature*. New York; Routledge.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition. Springer.

Hecht, Gabrielle (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.

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

Kelling, S., Hochachka, W., Fink, D., Riedewald, M., Caruana, R., Ballard, G. and Hooker, G. (2009) Data-intensive Science: A New Paradigm for Biodiversity Studies. *BioScience* 59(7): 613-620

Kitchin, R. (2013, in press) Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*

Kitchin, R. (2014, in press) The real-time city? Big data and smart urbanism. *Geojournal*

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

Lauriault, T.P. (2012) *Data, Infrastructures and Geographical Imaginations: Mapping Data Access Discourses in Canada*. PhD 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., Hackett, Y. and Kennedy, E. (2013) *Geospatial Data Preservation Primer*. Hickling, Arthurs and Low.

Latour, B. (1987) *Science in Action: How to Follow Scientists and Engineers Through Society*, Philadelphia: Open University Press.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D. and Van Alstyne, M. (2009) Computational Social Science. *Science* 323: 721-733.

Loukides, M. (2010) What is data science? *O'Reilly Radar*, 2 June 2010, http://radar.oreilly.com/2010/06/what-is-data-science.html (last accessed 28 January 2013)

Lyon, D. (2002) Everyday surveillance: Personal data and social classifications. *Information, Communication and Society* 5: 242-257

Manovich, L. (2011) Trending: The Promises and the Challenges of Big Social Data. http://www.manovich.net/DOCS/Manovich_trending_paper.pdf (last accessed 9 Nov 2012)

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.

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

Miller, H.J. (2010) The data avalanche is here. Shouldn't we be digging? *Journal of Regional Science* 50(1): 181-201

Moretti, F. (2005) *Graphs, Maps, Trees: Abstract Models for a Literary History.* Verso, London.

O'Carroll, A., Collins, S., Gallagher, D., Tang, J. and Webb, S. (2013) *Caring for Digital Content, Mapping International Approaches* NUI Maynooth, Trinity College Dublin, Royal Irish Academy and Digital Repository of Ireland, Dublin.

Open Data Center Alliance (2012) *Big Data Consumer Guide*. Open Data Center Alliance. http://www.opendatacenteralliance.org/docs/Big_Data_Consumer_Guide_Rev1.0.pdf (last accessed 11 February 2013)

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

Rameriz, E. (2013) The privacy challenges of big data: A view from the lifeguard's chair. *Technology Policy Institute Aspen Forum*, August 19th. http//ftc.gov/speeches/ramirez/130819bigdataaspen.pdf (last accessed October 11th 2013)

Ribes, D. and Jackson, S.J. (2013) Data bite man: The work of sustaining long-term study. In Gitelman, L. (ed) "Raw Data" is an Oxymoron. MIT Press, Cambridge. pp 147-166.

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: 116-136

Sawyer, S. (2008): Data Wealth, Data Poverty, Science and Cyberinfrastructure , *Prometheus: Critical Studies in Innovation*, 26:4, 355-371

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

Singer, N. (2012) You for Sale: Mapping, and Sharing, the Consumer Genome. *New York Times*, 17th June, www.nytimes.com/2012/06/17/technology/acxiom-the-quiet-giant-of-consumer-database-marketing.html (last accessed 11 October 2013)

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

Star, S.L. and Lamplard, M. (2009) Reckoning with practices. In Lamplard, M. and Star, S.L. (eds) *Standards* and *Their Stories*: How *Quantifying*, *Classifying, and Formalizing Practices Shape Everyday Life.* Cornell University Press, Ithaca. pp. 3-34.

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): 111-134.

Wilson, M. (2011) Data matter(s): legitimacy, coding, and qualifications-of-life. *Environment and Planning D: Society and Space* 29: 857-872