Original Research Article

BIG DATA & SOCIETY

Big Data & Society
January–June 2016: 1–10

DOI: 10.1177/2053951716631130
bds.sagepub.com
SAGE

What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets

Rob Kitchin[1] and Gavin McArdle[2]

Abstract

Big Data has been variously defined in the literature. In the main, definitions suggest that Big Data possess a suite of key traits: volume, velocity and variety (the 3Vs), but also exhaustivity, resolution, indexicality, relationality, extensionality and scalability. However, these definitions lack ontological clarity, with the term acting as an amorphous, catch-all label for a wide selection of data. In this paper, we consider the question 'what makes Big Data, Big Data?', applying Kitchin's taxonomy of seven Big Data traits to 26 datasets drawn from seven domains, each of which is considered in the literature to constitute Big Data. The results demonstrate that only a handful of datasets possess all seven traits, and some do not possess either volume and/or variety. Instead, there are multiple forms of Big Data. Our analysis reveals that the key definitional boundary markers are the traits of velocity and exhaustivity. We contend that Big Data as an analytical category needs to be unpacked, with the genus of Big Data further delineated and its various species identified. It is only through such ontological work that we will gain conceptual clarity about what constitutes Big Data, formulate how best to make sense of it, and identify how it might be best used to make sense of the world.

Keywords

Big Data, ontology, taxonomy, types, characteristics

Introduction

The etymology of 'Big Data' has been traced to the mid-1990s, first used by John Mashey, retired former Chief Scientist at Silicon Graphics, to refer to handling and analysis of massive datasets (Diebold, 2012). In 2001, Doug Laney detailed that Big Data were characterised by three traits:

- *volume* (consisting of enormous quantities of data);
- *velocity* (created in real-time) and;
- *variety* (being structured, semi-structured and unstructured).

Since then, others have attributed other qualities to Big Data, including:

- *exhaustivity* (an entire system is captured, $n = $ all, rather than being sampled) (Mayer-Schonberger and Cukier, 2013);
- *fine-grained* (in resolution) and *uniquely indexical* (in identification) (Dodge and Kitchin, 2005);
- *relationality* (containing common fields that enable the conjoining of different datasets) (Boyd and Crawford, 2012);
- *extensionality* (can add/change new fields easily) and *scaleability* (can expand in size rapidly) (Marz and Warren, 2012);
- *veracity* (the data can be messy, noisy and contain uncertainty and error) (Marr, 2014);
- *value* (many insights can be extracted and the data repurposed) (Marr, 2014);
- *variability* (data whose meaning can be constantly shifting in relation to the context in which they are generated) (McNulty, 2014).

[1]NIRSA, Maynooth University, County Kildare, Ireland
[2]UCD School of Computer Science, University College Dublin, Dublin, Ireland

Corresponding author:
Rob Kitchin, NIRSA, Maynooth University, County Kildare, Ireland.
Email: rob.kitchin@nuim.ie

Uprichard (2013) notes several other v-words that have also been used to describe Big Data, including: 'versatility, volatility, virtuosity, vitality, visionary, vigour, viability, vibrancy... virility... valueless, vampire-like, venomous, vulgar, violating and very violent.' More recently, Lupton (2015) has suggested dropping v-words to adopt p-words to describe Big Data, detailing 13: portentous, perverse, personal, productive, partial, practices, predictive, political, provocative, privacy, polyvalent, polymorphous and playful. While useful entry points into thinking critically about Big Data, these additional v-words and new p-words are often descriptive of a broad set of issues associated with Big Data, rather than characterising the ontological traits of the data themselves.

Based on a review of definitions of Big Data, Kitchin (2013, 2014) contends that Big Data are qualitatively different to traditional, small data along seven axes (see Table 1). He details that, until recently, science has progressed using small data that have been produced in tightly controlled ways using sampling techniques that limit their scope, temporality and size, and are quite inflexible in their administration and generation. While some of these small datasets are very large in size, they do not possess the other characteristics of Big Data. For example, national censuses are typically generated once every 10 years, asking just c.30 structured questions, and once they are in the process of being administered it is impossible to tweak or add/remove questions. In contrast, Big Data are generated continuously and are more flexible and scalable in their production. For example, in 2014, Facebook was processing 10 billion messages, 4.5 billion 'Like' actions, and 350 million photo uploads per day (Marr, 2014), and they were constantly refining and tweaking their underlying algorithms and terms and conditions, changing what and how data were generated (Bucher, 2012).

Similarly, Florescu et al. (2014), in a study examining the potential for Big Data to be used to generate new official statistics, details how Big Data differs from small data generated through state-administered surveys and administrative data. Kitchin (2015) extended their original table, adding three further fields to their 14 points of comparison (see Table 2). Table 2 makes it clear that Big Data have a very different set of characteristics to more traditional forms of small data across a range of attributes which extend beyond the data's essential qualities (including methods, sampling, data quality, repurposing, management).

Table 1. Comparing small and Big Data.

	Small data	Big Data
Volume	Limited to large	Very large
Velocity	Slow, freeze-framed/ bundled	Fast, continuous
Variety	Limited to wide	Wide
Exhaustivity	Samples	Entire populations
Resolution and indexicality	Course and weak to tight and strong	Tight and strong
Relationality	Weak to strong	Strong
Extensionality and scalability	Low to middling	High

In contrast, rather than focusing on the ontological characteristics of what constitutes the nature of Big Data, some define Big Data with respect to the computational difficulties of processing and analyzing it, or in storing it on a single machine (Strom, 2012). For example, Batty (2015) contends that Big Data challenges conventional statistical and visualization techniques, and push the limits of computational power to analyze them. He thus contends that we have always had Big Data, with the massive datasets presently being produced merely the latest form of Big Data, which require new technique to process, store and make sense of them. Murthy et al. (2014) categorises Big Data using a six-fold taxonomy that likewise focuses on its handling and processing rather than key traits: (1) data ((a) temporal latency for analysis: real-time, near real-time, batch; and (b) structure: structured, semi-structured, unstructured); (2) compute infrastructure (batch or streaming); (3) storage infrastructure (SQL, NoSQL, NewSQL); (4) analysis (supervised, semi-supervised, unsupervised or re-enforcement machine learning; data mining; statistical techniques); (5) visualisation (maps, abstract, interactive, real-time); and (6) privacy and security (data privacy, management, security).

Regardless of how Big Data have been defined it is clear that, despite widespread use, the term is still rather loose in its ontological framing and definition, and it is being used as a catch-all label for a wide selection of data. The result is that these data are characterised as holding similar traits to each other and the term 'Big Data' is treated like an amorphous entity that lacks conceptual clarity. However, for those who work with and analyze datasets that have been labelled as Big Data it is apparent that, although they undoubtedly share many traits, they also vary in their characteristics and nature. Not all of the data types that have been declared as constituting Big Data have volume, velocity or variety, let alone the other characteristics noted above. Nor do they all overly challenge conventional statistical techniques or computational power in making sense of them. In other words, there are multiple forms of Big Data. However, while there has been some rudimentary work to identify the 'genus' of Big Data, as detailed above, there has been no attempt to separate out its various 'species' and their defining attributes.

Table 2. Characteristics of survey, administrative and Big Data.

	Survey data	Administrative data	Big Data
Specification	Statistical products specified ex-ante	Statistical products specified ex-post	Statistical products specified ex-post
Purpose	Designed for statistical purposes	Designed to deliver/monitor a service or program	Organic (not designed) or designed for other purposes
Byproducts	Lower potential for by-products	Higher potential for by-products	Higher potential for by-products
Methods	Classical statistical methods available	Classical statistical methods available, usually depending on the specific data	Classical statistical methods not always available
Structure	Structured	A certain level of data structure, depending on the objective of data collection	A certain level of data structure, depending on the source of information
Comparability	Weaker comparability between countries	Weaker comparability between countries	Potentially greater comparability between countries
Representativeness	Representativeness and coverage known by design	Representativeness and coverage often known	Representativeness and coverage difficult to assess
Bias	Not biased	Possibly biased	Unknown and possibly biased
Error	Typical types of errors (sampling and non-sampling errors)	Typical types of errors (non-sampling errors, e.g., missing data, reporting errors and outliers)	Both sampling and non-sampling errors (e.g., missing data, reporting errors and outliers) although possibly less frequently occurring, and new types of errors
Persistence	Persistent	Possibly less persistent	Less persistent
Volume	Manageable volume	Manageable volume	Huge volume
Timeliness	Slower	Potentially faster	Potentially much faster
Cost	Expensive	Inexpensive	Potentially inexpensive
Burden	High burden	No incremental burden	No incremental burden
Geography	National, defined	National or extent of program and service	National, international, potentially spatially uneven
Demographics	All or targeted	Service users or program recipients	Consumers who use a service, pass a sensor, contribute to a project, etc.
Intellectual Property	State	State	State/Private sector/ User-created rights.

Source: Florescu et al. (2014: 2–3) and Kitchin (2015)

In this paper, we examine the ontology of Big Data and its definitional boundaries, exploring the question 'what makes Big Data, Big Data?' We employ Kitchin's (2013) taxonomy of the characteristics of Big Data (Table 1) to examine the nature of 26 specific types of data, drawn from seven domains (mobile communication; websites; social media/crowdsourcing; sensors; cameras/lasers; transaction process generated data; and administrative), that have been labelled in the literature as Big Data (see Table 3). These 26 types of data are by no means exhaustive of all types of Big Data, for example there are a multitude of Big Data generated within scientific experiments, science computing, and industrial manufacturing. Rather, these 26 datasets are used for illustrative purposes and were selected due to our familiarity with them. We start by examining each of the parameters detailed by Kitchin with respect to the 26 different data types, in effect working down the columns in Table 3. We then examine the rows to consider how these parameters are combined with respect to the data types to produce multiple forms of Big Data.

Our aim in performing this analysis is not to determine a tightly constrained definition of Big Data – to definitively set out precisely the nature of Big Data and their essential qualities – but rather to explore the parameters, limits, and 'species' of Big Data. The analysis is thus an exercise in boundary work designed to test

Table 3. Ontological traits of Big Data.

Data type		Volume (number of records)	Volume per record	Volume (TBs, PBs, etc.)	Velocity frequency of generation	Velocity frequency of handling, recording, publishing	Variety	Exhaustivity	Resolution	Indexical	Relational	Extensionality	Scalable
Mobile communication	Mobile phone data	High	Low	High	Real-time constant (bkgrd comms), real-time sporadic (at use)	At time of generation	Structured	$n = all$	Fine-grained	Yes	Yes	No	Yes
	App data	High	Low	High	Real-time constant (bkgrd comms), real-time sporadic (at use)	At time of generation	Structured & unstructured	n = all	Fine-grained	Yes	Yes	Yes	Yes
Websites	Web searches	High	Low	High	Real-time sporadic	At time of generation	Structured & unstructured	$n = all$	Fine-grained	Yes	Yes	Yes	Yes
	Scraped websites	High	Medium	High	Real-time sporadic	At time of generation	Semi-structured	$n = all$	Fine-grained	Yes	Yes	Yes	Yes
	Clickstream	High	Low	High	Real-time sporadic	At time of generation	Structured	$n = all$	Fine-grained	Yes	Yes	Yes	Yes
Social media/ Crowdsourcing	Social media (full pipe) (e.g. Twitter)	High	Medium	High	Real-time sporadic	At time of generation	Structured & unstructured	$n = all$	Fine-grained	Yes	Yes	Yes	Yes
	Social media (spritzer) (e.g. twitter)	Low	Medium	Medium	Real-time sporadic	At time of generation	Structured & unstructured	Sampled	Fine-grained	Yes	Yes	Yes	Yes
	Picture sharing/ social media (flickr, Panoramio, Instagram)	High	High	High	Real-time sporadic	At time of generation	Structured & unstructured	$n = all$	Fine-grained	Yes	Yes	Yes	Yes
	Collaborative mapping platforms (OpenStreetMap, Wikimapia)	Low	Low	Low	Real-time sporadic	At time of generation (open to editing)	Structured & semi-structured	$n = all$	Fine-grained	Yes	Yes	Yes	Yes
	Citizen science (wunderground)	High	Low	Medium	Real-time constant or real-time sporadic	At time of generation	Structured	$n = all$	Fine-grained	Yes	Yes	No	Yes
Sensors	Traffic loops	Medium	Low	Low	Real-time constant	At time of generation	Structured	$n = all$	Aggregated	Yes	Yes	No	No

(continued)

Table 3. Continued

Data type		Volume (number of records)	Volume per record	Volume (TBs, PBs, etc.)	Velocity frequency of generation	Velocity frequency of handling, recording, publishing	Variety	Exhaustivity	Resolution	Indexical	Relational	Extensionality	Scalable
	Automatic Number Plate Readers (ANPR)	Medium	Low	Medium	Real-time constant	At time of generation	Structured	n = all	Fine-grained	Yes	Yes	No	Yes
	Real-time passenger info (RTPI)	Medium	Low	Low	Real-time constant	At time of generation	Structured	n = all	Fine-grained	Yes	Yes	No	No
	Smart meters	High	Low	Medium	Real-time constant	At time of generation	Structured	n = all	Fine-grained	Yes	Yes	No	No
	Pollution and sound sensors	Medium	Low	Low	Real-time constant	At time of generation	Structured	n = all	Fine-grained	Yes	Yes	No	No
	Satellite images	Medium	High	High	Real-time constant	At time of generation	Unstructured	n = all, delayed repeat of coverage	Fine-grained	Yes	Yes	No	No
Cameras/Lasers	Digital CCTV	High	High	High	Real-time constant	At time of generation	Unstructured	n = all	Fine-grained	Yes	Yes	No	No
	Lidar mapping (by HERE)	High	High	High	Real-time constant (when in use)	Delayed and consolidated (daily)	Structured	n = all, but no or infrequent repeat coverage	Fine-grained	Yes	Yes	No	No
Transactions of process generated data	Supermarket scanner and sales data	High	Low	High	Real-time sporadic	At time of generation	Structured	n = all	Fine-grained	Yes	Yes	No	Yes
	Immigration (inc. photo, fingerprint scan)	Low	High	High	Real-time sporadic	At time of generation	Structured	n = all, infrequent repeat coverage	Fine-grained	Yes	Yes	No	Yes
	Flight movements	High	Low	High	Real-time constant	At time of generation	Structured	n = all	Fine-grained	Yes	Yes	No	Yes
	Credit card data	High	Low	High	Real-time sporadic	At time of generation	Structured	n = all	Fine-grained	Yes	Yes	No	Yes
	Stock market trades	High	Low	High	Real-time sporadic	At time of generation	Structured	n = all	Fine-grained	Yes	Yes	No	Yes
Administrative	House price register	Low	Low	Low	Real-time sporadic	Delayed and consolidated (monthly)	Structured	n = all	Fine-grained	Yes	Yes	No	Yes

(continued)

Table 3. Continued

Data type	Volume (number of records)	Volume per record	Volume (TBs, PBs, etc.)	Velocity frequency of generation	Velocity frequency of handling, recording, publishing	Variety	Exhaustivity	Resolution	Indexical	Relational	Extensionality	Scalable
Planning permissions	Low	Low	Low	Real-time sporadic	Delayed and consolidated (weekly)	Structured	$n =$ all, but no or infrequent repeat coverage	Fine-grained	Yes	Yes	No	Yes
Employment register (at release)	Low	Low	Low	Real-time sporadic	Delayed and consolidated (monthly)	Structured	$n =$ all	Aggregated	Yes	Yes	No	Yes

bkgrd comms: constant background passive monitoring.

the edges of what might be considered Big Data and to internally tease apart what is presently an amorphous concept to reveal its inner diversity – its multiple forms. In other words, we consider in much more detail than previous studies the ontology of Big Data. This is an important exercise, we believe, as it enables the production of much more conceptual clarity about what constitutes Big Data, especially given the ongoing confusion over its traits and its amorphous description. In turn, acknowledging and detailing the various types of Big Data facilitates a much more nuanced understanding of its forms, its value, and how they might be analyzed and for what purposes.

The parameters of Big Data

In Table 3 we have mapped 26 sources of data, defined as Big Data within the literature, against the traits identified by Kitchin (2014) in Table 1. Through the process of evaluating each dataset against each characteristic it quickly became apparent that the categories of volume and velocity needed to be further teased apart. Similarly, while resolution and indexicality, and extensionality and scalability, are combined into two characteristics in Table 1, we consider them separately in Table 3 given that they are not synonymous traits.

In the context of Big Data, volume generally refers to the storage space required to record and store data. Big Data, it is commonly stated, typically require terabytes (2^{40} bytes) or petabytes (2^{50} bytes) of storage space (The Economist, 2010), far more than an average desktop computer can provide, with the data often stored in the cloud across several servers and locations. However, when we examine our 26 datasets it is clear that some of them, for example pollution and sound sensors, require very little storage space, maybe producing a gigabyte (2^{30} bytes) of data per annum (easily storable on a datastick). Although each sensor might be producing a steady stream of readings, say once per minute, each record is very small, consisting of just a few kilobytes (2^{10} bytes). Even summed over the course of a year, the sensor dataset would be relatively small in stored volume, in fact much smaller than many 'small datasets' such as a Census. As detailed in Table 3, we have thus teased apart volume into three dimensions: (1) the number of records (which is reflective of velocity and the number of generating devices), (2) the storage required per record, and (3) the total storage required (effectively the sum of the first two).

Using this threefold classification of volume it is clear that the 26 Big Data sets have differing volume characteristics. Automated forms of Big Data generation, where records are created on a continual basis every few seconds or minutes, often across multiple sites or individuals, produce very large numbers of

records. Human-mediated forms, such as creating administrative records (immigration, unemployment registration), might have a steady stream of new records, usually generated from a constrained number of sites (a small number of entry points to a country, unemployment offices), that produce much lower volumes than automated systems. Likewise, while each sensor record is generally very small in file size, imagery data (such as streaming video, photographs and satellite images) are typically quite large in file size, meaning that relatively low numbers of records soon scale into huge storage requirements. In many cases, although the volume per record is low, the sheer number of devices generating data produce very large storage volumes. For example, the million customers flowing through thousands of Walmart stores every hour generate 2.5 petabytes of transaction data (Open Data Center Alliance, 2012).

Velocity is considered a key attribute of Big Data. Rather than data being occasionally sampled (either on a one-off basis or with a large temporal gap between samples), Big Data are produced much more continually. When we examined our datasets, however, it became apparent that there are two kinds of velocity with respect to Big Data: (1) frequency of generation; (2) frequency of handling, recording, and publishing; and that the 26 datasets varied with respect to these two traits. In terms of frequency of generation, data can be generated in real-time constantly, for example recording a reading every 30 seconds or verifying location every 4 minutes (as many mobile phone apps do), or in real-time sporadically, for example at the point of use, such as clickstream data being generated in real-time but only while a user is clicking through websites, or an immigration system recording only when someone is scanning their documents.

In some cases, as the data are recorded, the system is updated in real-time and the new data are also published in real-time (with only a fraction of delay between the two). For example, as a tweet is tweeted it is recorded in Twitter's data architecture and microseconds later it is published into user timelines. Here, even though the data generation is sporadic at the point of generation (each user might only produce a couple of tweets a day), it is far from the case at the point of recording by the company (the millions of Twitter users collectively generate thousands of tweets per second, meaning that the company databases and servers are constantly handling a data deluge). In other cases, the data are recorded in real-time, but their transmission to central servers and/or their processing or publication is delayed. For example, the HERE LIDAR scanning project involves 200 cars driving around cities taking a LIDAR scan every second to produce high definition mapping data (Nokia, 2015). A single LIDAR scan generally produces a million plus points of data (Cahalane et al., 2012). At the end of every day the local storage device is removed from the vehicle performing the scan and its data transferred to a data centre. Similarly, unemployment data are recorded at the time a person updates their status on the system, but the overall unemployment rate is published monthly and in an aggregated form. In some cases, even once the data are generated they are open to further editing, as with crowdsourced data within Wikipedia or OpenStreetMap, with the edits also recorded in real-time and becoming part of the dataset.

Perhaps not unsurprisingly, there is a fair range of variety in the data form across our 26 datasets, including structured, semi-structured and unstructured data types. Of all the characteristics attributed to Big Data this seems to us to be the weakest attribute. Indeed, small data are also highly heterogeneous in nature, especially datasets common to humanities and social sciences where the handling and analyzing of qualitative data (text, images, etc.) is normal. Our suspicion is that this characteristic was attributed to Big Data because those scientists who first coined the term were used to handling structured data exclusively but were starting to encounter semi-structured and unstructured data as new data generation and collection systems were deployed.

As noted, small datasets consist of samples of representative data harvested from the total sum of potentially available data. Sampling is typically used because it is unfeasible in terms of time and resources to harvest a full dataset. In contrast, Big Data seeks to capture the entire population ($n = \text{all}$) within a system, rather than a sample. In other words, Twitter captures all tweets made by all tweeters, plus their associated 32 fields of metadata, not a sample of tweets or tweeters. Similarly, a set of pollution sensors is seeking to create a continuous, longitudinal record of readings, captured every few seconds, from a fixed network of sensors. Likewise, a credit card company or the stock market seeks to record every single transaction and alter credit balances accordingly.

All our 26 datasets hold the characteristic of $n = \text{all}$, except for the spritzer of Twitter; this is the sample of tweets harvested from the full fire hose that Twitter shares with some researchers. It is important to note, however, that the temporality of $n = \text{all}$ can vary. For example, an immigration system at an airport aims to capture details about all passengers passing through it, but a passenger might only pass through that system infrequently. In the case of a satellite, it might capture imagery of the whole planet, but it only flies over the same portion of the Earth every set number of days. Likewise, the HERE LIDAR project aims to scan every road in every country, but each street is only surveyed

once and is unlikely to be rescanned for several years. In other words, Big Data systems seek to capture $n = $all, but capturing $n = $all varies with respect to what is being measured and their spatial coverage and temporal register.

As with exhaustivity, all 26 datasets hold the traits of fine-grained resolution (with the exception of employment data, which is fine-grained in the database but is published in aggregated form), indexicality and relationality. In each case, the data are accompanied by metadata that uniquely identifies the device, site and time/date of generation, along with other characteristics such as device settings. These metadata inherently produces relationality, enabling data from the same and related devices but generated at different times/locales to be linked, but also entirely different datasets that share some common fields to be tied together and relationships between datasets to be identified. However, the data themselves might not provide unambiguous relationality or be easily machine-readable. For example, a tweet is composed of text and/or an image which requires either data analytics or human interpretation to identify the content and meaning of the tweet. Similarly, a CCTV feed will be indexical to a camera and be time, date, and place stamped, but the content of the feed will either require image recognition to identify content (e.g., using facial recognition software) or operator recognition to make the image content indexical.

Extensionality and scaleability refer to the flexibility of data generation. A system that is highly adaptable in terms of what data are generated is said to possess strong extensionality (Marz and Warren, 2012). For example, web-based and mobile apps are constantly tweaking their designs and underlying algorithms, performing on-the-fly adaptive testing and rollout, as well as altering their terms and conditions and the metadata they capture. The result is the data they generate are changeable, with new fields being added and removed as required. However, this is not a trait common to all big datasets. For example, many systems, such as smart meters, credit card readers and sensor-networks, are seeking rigid continuity in what data are generated to produce robust, comparable longitudinal datasets. Scaleability refers to the extent to which a system can cope with varying data flow. Social media platforms such as Twitter need to be able to cope with ebbs and surges in data generation, scaling from managing a few thousand tweets at certain times of the day to tens of thousands during popular live events. Such rapid scaling is not required in systems that have a constant flow of data, such as a sensor network that produces data at set intervals (the timing can be altered, but the flow remains constant rather than surging). As such, some of the 26 datasets are generated and stored within rapidly scaleable systems, but not others.

The forms and boundaries of Big Data

What is clear from examining each Big Data parameter with respect to the 26 datasets is that there is no one characteristic profile that all Big Data fit. Big Data does not possess all of the seven traits detailed by Kitchin (2013, 2014). Indeed, not all data termed Big Data in the literature possess the 3Vs of volume, velocity and variety. If one looks across the rows in Table 3 then the diversity of Big Data becomes clear, with datasets possessing differing profiles, especially with regard to volume, velocity, variety, extensionality and scalability. Big Data are clearly then not an amorphous category and there are certainly different 'species' of Big Data.

Examining these profiles starts to suggest the boundary markers of what constitutes Big Data. Indeed, it may be the case that some of our 26 datasets might not be considered Big Data by some. Or it might be that some consider certain datasets to constitute Big Data that we would not, for example, national censuses (which have volume, exhaustivity, resolution, indexicality and relationality, but no velocity (generated once every 10 years and taking 1–2 years to process), no extensionality or scaleability, and are published in aggregated form). It seems to us, based on the datasets that we have examined, that the key boundary characteristics of Big Data, which together differentiate it from small data, are velocity (both frequency of generation, and frequency of handling, recording, and publishing) and exhaustivity. Small data are slow and sampled. Big Data are quick and $n = $all. Small data can hold all of the other characteristics (volume, resolution, indexicality, relationality, extensionality and flexibility) and still be considered small in nature. It is the qualities of velocity and exhaustivity which set Big Data apart and are responsible for so much recent attention and investment in Big Data ventures. While some datasets have possessed these two qualities for some time, such as stock market and weather data, it is only in the past 15 years that these characteristics have become much more common and routine.

These two traits, we believe, act as key Big Data boundary markers. In our own analysis of Table 3 it was the administrative datasets of the house price register, planning permissions and unemployment, as well as the satellite and LIDAR imagery that provoked the most discussion (we quite quickly rejected Census data, which we had initially included, due to its very long temporal gap in data generation). In the case of the administrative data, they are produced in real-time as entries are made into the system (as house sales are

completed, planning permissions sought, and unemployed people sign-on). However, the publishing of the data is either weekly or monthly, and in the case of unemployment released in an aggregated form. Do data that are generated in real-time, but released monthly and in an aggregated form constitute Big Data? Certainly they are at the point of collection, but what about at the point of publishing where they lack velocity? For some, such administrative data are Big Data (Economic and Social Research Council (ESRC), 2013), for others they are more marginal, and the key element in doubt is temporality. One month's delay is still much quicker than most administrative data that are published quarterly or annually, and the dataset still holds most of the other characteristics of Big Data such as exhaustivity (the data refers to all houses sold, all planning permissions sought, and all unemployed people), but it is nonetheless far slower than data published in real-time.

Our discussion of satellite imagery and LIDAR focused in particular on coverage and repetition of gaze. In other forms of Big Data, what is being measured remains quite constant, with the gaze and the object under surveillance relatively fixed. In social media it is the contributions of every user, for credit cards it is the transactions of every card holder, for supermarkets it is the purchases of every shopper. However, the gaze of the satellite imagery moves, only returning to capture the same terrain after a set number of days. Nonetheless the surface of the entire planet is being repeatedly generated and data are processed constantly. In the case of LIDAR, that repetition is missing. The aim is to scan every road on the planet, but to do so only once. The data are generated in real-time, and are voluminous, indexical, relational, and they produce exhaustive spatial coverage (the aim is to create a 3D model of the whole road network and the architecture bordering this network) though no longitudinal data of the same places. In both cases, most would agree that satellite imagery and LIDAR scans constitute Big Data, but they are exhaustive in a particular way which distinguishes them from other types of Big Data. The same would also be the case with respect to large scientific experiments such as data generated by the Large Hadron Collider.

Interestingly, given the meme of the 3Vs of Big Data, having examined 26 types of Big Data, our conclusion is that two of those Vs – volume and variety – are not key defining characteristics of Big Data. It is certainly the case that Big Data often consists of very large numbers of records and the storage volume required to store them is significant, however, this is not a necessary condition of Big Data. Rather volume is a by-product of velocity and exhaustivity: the real-time flow of data across a whole system can produce a deluge of data, especially if each record is large in size. In some cases, however, the flow can be generated in real-time (e.g., every 30 seconds), but because the system is small (e.g., 30 sound sensors across a city) and each record is small in size, the storage volume is relatively small. The data generated by each sensor are also highly structured. Despite the lack of volume and variety, such sensor data are widely considered Big Data. Likewise, variety is not a distinguishing characteristic because small data possesses just as much variety as Big Data.

Conclusion

To date, there has been very little work that has sought to examine in detail the ontology of Big Data, other than to suggest that they are data that possess certain broad characteristics (volume, velocity, variety, exhaustivity, etc.). Indeed, most studies that discuss Big Data treat the term as a catch-all, amorphous phrase that assumes that all Big Data share a set of general traits. Through an analysis that applied Kitchin's (2013, 2014) typology of Big Data traits to 26 datasets our study reveals that Big Data do not all share the same characteristics and that there are multiple forms of Big Data. Indeed, our analysis demonstrates that only a handful of the 26 datasets we examined held all seven traits identified by Kitchin. That said, it is the case that for Big Data to be classified as Big Data they do need to possess the majority of the traits set out in Table 1, of which velocity and exhaustivity are the most important. Volume and variety, we contend, are not necessary conditions of Big Data and without velocity and exhaustivity are not qualifying criteria. In other words, the 3Vs meme is actually false and misleading and along with the term itself is partially to blame for the confusion over the definitional boundaries of Big Data.

The observation that there are multiple forms of Big Data is perhaps no surprise given the wide variety of small data, the varying nature of the systems that generate Big Data, the differing purposes for which the data are generated, and the differing forms of the data generated. Nonetheless it is an observation that needs highlighting given that it has so far been ignored or taken for granted in the literature. Our analysis has revealed that Big Data as an analytical category needs to be unpacked, with the 'genus' of Big Data further delineated and its various 'species' identified. This is important work if we are to better understand what it is that we are talking about when we discuss and analyze Big Data, and if we want to produce more nuanced insights about and from the data. It is only through such ontological work, focused on shifting from broad generalities to specific qualities, that we will

gain conceptual clarity about what constitutes Big Data and formulate how best to make sense of it and how it might be used to make sense of the world.

Declaration of conflicting interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research for this paper was funded by a European Research Council Advanced Investigator Award, 'The Programmable City' (ERC-2012-AdG-323636).

References

Batty M (2015) Data about cities: Redefining big, recasting small. Paper prepared for the Data and the City workshop, Maynooth University, 31 August–1 September 2015. Available at: http://www.spatialcomplexity.info/files/2015/08/Data-Cities-Maynooth-Paper-BATTY.pdf (accessed 4 September 2015).

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

Bucher T (2012) 'Want to be on the top?' Algorithmic power and the threat of invisibility on Facebook. *New Media and Society* 14(7): 1164–1180.

Cahalane C, McCarthy T and McElhinney CP (2012) MIMIC: Mobile mapping point density calculator. In: *Proceedings of the 3rd international conference on computing for geospatial research and applications*, 1–3 July 2012, Washington, DC, USA: ACM.

Diebold F (2012) A personal perspective on the origin(s) and development of 'big data': The phenomenon, the term, and the discipline. Available at: http://www.ssc.upenn.edu/~fdiebold/papers/paper112/Diebold_Big_Data.pdf (accessed 5 February 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.

Economic and Social Research Council (ESRC) (2013) The Big Data family is born – David Willetts MP announces the ESRC Big Data Network. In: ESRC website, 10 October. Available at: http://www.esrc.ac.uk/news-and-events/press-releases/28673/the-big-data-family-is-born-david-willetts-mp-announces-the-esrc-big-data-network.aspx (accessed 7 September 2015).

Florescu D, Karlberg M, Reis F, et al. (2014) Will 'big data' transform official statistics? Available at: http://www.q2014.at/fileadmin/user_upload/ESTAT-Q2014-BigDataOS-v1a.pdf (accessed 1 April 2015).

Kitchin R (2013) Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography* 3(3): 262–267.

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

Kitchin R (2015) The opportunities, challenges and risks of big data for official statistics. *Statistical Journal of the International Association of Official Statistics* 31(3): 471–481.

Laney D (2001) 3D data management: Controlling data volume, velocity and variety. In: Meta Group. Available at: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (accessed 16 January 2013).

Lupton D (2015) The thirteen Ps of big data. *The Sociological Life*, 13 May. Available at: https://simplysociology.wordpress.com/2015/05/11/the-thirteen-ps-of-big-data/ (accessed 17 September 2015).

Marr B (2014) Big data: The 5 vs everyone must know. March 6. Available at: https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know (accessed 4 September 2015).

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

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

McNulty E (2014) Understanding Big Data: The seven V's. 22 May. Available at: http://dataconomy.com/seven-vs-big-data/ (accessed 4 September 2015).

Murthy P, Bharadwaj A, Subrahmanyam PA, et al. (2014) *Big Data Taxonomy*. Big Data Working Group, Cloud Security Alliance. Available at: https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Taxonomy.pdf (accessed 7 September 2015).

Nokia (2015) HERE makes HD map data in US, France, Germany and Japan available for automated vehicle tests. Available at: http://company.nokia.com/en/news/press-releases/2015/07/20/here-makes-hd-map-data-in-us-france-germany-and-japan-available-for-automated-vehicle-tests (accessed 16 September 2015).

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

Strom D (2012) Big data makes things better. *Slashdot*, 3 August. Available at: http://slashdot.org/topic/bi/big-data-makes-things-better/ (accessed 24 October 2013).

The Economist (2010) All too much: Monstrous amounts of data, 25 February. Available at: http://www.economist.com/node/15557421 (accessed 12 November 2012).

Uprichard E (2013) Big data, little questions. *Discover Society*, 1 October. Available at: http://discoversociety.org/2013/10/01/focus-big-data-little-questions/ (accessed 17 September 2015).