1 Defining Big Data

Small data are slow and sampled. Big Data are quick and n=all.

Kitchin & McArdle (2016)

This chapter searches for defining properties of big data, focusing on characteristics with possible implications for cartographic practice. Review of related works outlines the main attitudes towards grasping the concept.

1.1 Ontological characteristics

Despite the lively interest triggered by the subject, the explanation of the term *big data*[1] remains hazy and there is no widely accepted definition to the date. Perhaps the most systematic effort in this matter by Kitchin (2014) (refined in Kitchin & McArdle (2016)) summarizes the key properties attributed to big data. Kitchin critically evaluates these properties and goes on to assign them a relative importance in distinguishing big from "small" data. He also takes care to separate the concept in itself from accompanying social phenomena, hence he speaks of *ontological* characteristics.

Kitchin's taxonomy provides a useful starting point for our thinking of big data from the cartographic standpoint, so let

1 Throughout the text we will treat the term as plural, without capitalization. Although there are strong arguments for "data" as singular (Widman (2014), Nunberg (2013), for counterargument emphasizing the plurality of big data see Wilson, Thompson, Watson, Drew, & Doyle (2017)) and some authors do capitalize, we chose to match with the majority of big data related literature. This does not apply to direct citations where we preserve the original author's formulation.

us list the ontological characteristics including some of the Kitchin's comments:

Volume — can be measured in storage requirements (terabytes or petabytes) or in number of records **Velocity** — data generation happens in real-time either constantly (e.g. CCTV) or sporadically (e.g. web search); we can distinguish the frequency of generation from the frequency of data *handling*, *recording*, and *publishing*, where all three can be delayed from the time of generation **Variety** — data are heterogeneous in nature, though this property is rather weak as various levels of organization are allowed (*structured*, *semi-structured* or *unstructured*) **Exhaustivity** — an entire system is captured (*n=all*), rather than working with a subset created by sampling **Resolution and indexicality** — fine-grained (in resolution) rather than being aggregated; uniquely indexical (in identification), which enables linking to other datasets **Relationality** — containing common fields that enable the conjoining of different datasets **Extensionality and scalability** — flexibility of data generation, possibility to add or change new fields easily, possibility to rapidly expand in size

In relation to these characteristics it is important to mention two open questions that for many people make attempts to define big data vague at best, sometimes to the point of questioning the existence of the phenomenon itself.

First, there are no quantitative thresholds that would define exactly how large the "big" volume is, how fast the "big" velocity is, and so on. Some properties would even be hard to describe in quantitative terms (for example extensionality). Other properties sound too general or vague to act as a sound defining parameter (scalability). What is more, one could extend the properties ad absurdum, for example *variety* could refer to differences in structure, origin, quality, or any other property of a dataset. Such multilevel hierarchy of

parameters and sub-parameters does not add to the overall comparability of datasets, especially when we consider that data generation procedures may be unique to certain domains and not found in others. Finally, many datasets lack metadata detailed enough to allow to judge all mentioned properties. It is possible that these issues will clear out with time, but parameter thresholds may as well remain blurry and ever in flux.

The second problem is that even if we had a clearly defined set of criteria, in practice we could hardly find a dataset that would fit all of them. Therefore not all properties are deemed mandatory, which in turn leads to confusion and labeling almost anything as big data. To articulate the gist of the term, more work is needed on the relations of the parameters, some might be merged (resolution is a consequence of exhaustivity, indexicality enables relationality) or discarded (extensionality and scalability seem to describe the infrastructure rather than data).

Aware of these problems, Kitchin & McArdle (2016) argues that *velocity* and *exhaustivity* are qualities that set big data apart and distinguish them from "small" data. We can add that these two characteristics also present the most interesting challenges to cartographic presentation of such data. So even though we will continue to use the established term in the following chapters, the little too simplistic adjective "big" will be meant as a proxy for **generated continuously in real time and containing an unreduced set of elements**.

1.2 Other ways of understanding big data

In this section we briefly review the writing of authors seeking to define big data. The term itself was fist used in context of dealing with massive datasets in mid-1990s by

John Mashey (Diebold et al., 2012), but the heaviest circulation of the term in scientific and popular media takes place only in recent years. From the breadth of works, several tendencies can be identified, providing more or less illuminating interpretations of the subject.[2]

1.2.1 Vs and keywords

Kitchin's taxonomy mentioned in the previous section is based on a review of older definitions, starting with the often-cited three Vs (standing for *volume*, *velocity*, and *variety*) by Laney (2001). The notion of *exhaustivity* was added by Mayer-Schönberger & Cukier (2013), concepts of *resolution* and *indexicality* came from Dodge & Kitchin (2005), Boyd & Crawford (2012) adds *relationality*, and the qualities of *extensionality* and *scalability* were taken from Marz & Warren (2012).

Other properties attributed to big data include *veracity* (data can be messy, noisy and contain uncertainty and error) and *value* (many insights can be extracted, data can be repurposed), both brought forward by Marr (2014) referring to the messiness and trustworthiness that is usually less controllable in case of big data. One could argue that these properties are just an another aspect of variety, as data vary not only in type and structure, but also in quality. This is can be the case for small data as well, however as Marr (2014) hopes, "the volumes often make up for the lack of quality or accuracy", which is sure debatable.

Moreover, *variability* (the meaning obtainable from data is shifting in relation to the context in which they are generated) was identified by David Hopkins in relation to text analysis (Brunelli, 2011). Li et al. (2016) name also

2 for an alternative summary of definitions see Gandomi & Haider (2015), for bibliometric analysis of related scientific literature see Nobre & Tavares (2017).

visibility (efficient access to data via cloud storage and computing) and more curiously *visualistation* as big data properties.

Suthaharan (2014), dealing with a task of early recognition of big data characteristics in computer network traffic, argues that three Vs do not support such early detection in continuous data streams. Instead he proposes three Cs: *cardinality* (number of records), *continuity* (meaning both representation of data by continuous functions, and continuous growth of size with time), and *complexity* (which is again a combination of three parameters: *large varieties of data types*, *high dimensionality*, and *high speed of processing*). One might ask why authors seek to propose parameters in triples, even at the cost of occluding additional properties as sub-parameters. Possible answer might be that such triples allow to create three-dimensional parameter spaces or "cubes" where we can place datasets to create neat visualisations. Humor aside, Suthaharan's approach is interesting in observing the rate of change in parameters in real time.

Laney's 3 Vs were brought into commercial management-speak and became a slogan further powering the hype of big data. Nevertheless, it inspired a number of other authors to extend it quite creatively. For example Uprichard (2013) lists other v-words to be considered, both in positive (*versatility*, *virtuosity*, *vibrancy*...) and negative (*valueless*, *vampire-like*, *violating*...) light. Marr (2014) describes five Vs of big data, Van Rijmenam (2013) sees seven Vs, Boellstorff & Maurer (2015) propose three Rs and Lupton (2015) even uses thirteen p-words to describe the subject. But as Kitchin & McArdle (2016) notes, "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 data themselves".

1.2.2 A challenge for technical infrastructure

Several authors understand big data mainly as a management issue, which is probably due to the fact that handling large datasets is challenging. Hence, the computational difficulties of storing and processing a dataset on a single machine often act as a defining measure. Consider for instance Storm (2012) quoting Hillary Mason: "Big Data usually refers to a dataset that is too big to fit into your available memory, or too big to store on your own hard drive, or too big to fit into an Excel spreadsheet." Or similarly Shekhar, Gunturi, Evans, & Yang (2012) state that "the size, variety and update rate of datasets exceed the capacity of commonly used spatial computing and spatial database technologies to learn, manage, and process the data with reasonable effort".

The problem with such definitions is determining exactly what size is "too big to fit" and what is the "reasonable effort". The computational power of hardware accessible for personal use is constantly increasing,[3] not to mention the technical infrastructure accessible to large enterprises and governmental organizations — datacenter construction is steadily growing and is expected to almost double the current capacity in 2021 (Networking, 2018; statista.com, 2018).

At the same time, new technologies emerge to address the issue — virtualization of storage, networking, and memory make it possible to rent computational infrastructure from "cloud" providers, or to delegate workloads previously carried

3 Gordon Moore's 1965 paper (reprint Moore, 2006) stated that the number of transistors on integrated circuits will double every two years. The prediction has proven accurate for several decades and became known as *Moore's law*. The pace has slowed down with smaller transistors suggesting that the prediction is reaching its technological limit, though the opinions here vary. The overuse of the idea as a synonym of progress has been criticized as too simplistic for example by Kreye (2015)

out by the operating system to remote platforms.[4] Other innovations take place in data processing algorithms, analytic engines, and in database design (a whole range of No-SQL databases as well as enablement of distributed processing in traditional databases).[5] Some attempts to summarize technical solutions for big data can be found in Pääkkönen & Pakkala (2015), or Jin, Wah, Cheng, & Wang (2015).

As we can see, the "too big to fit" definitions are highly dependent on the resources currently available, plus we need to take into account future improvements that are hard to predict. That being said, understanding the subject as *data that prevent local offline processing on common desktop in reasonable time* is a useful shorthand for judging big from "small" data. The border between local (offline) and remote (cloud-dependent) processing exists even though it is a blurry and a dynamic one. As the remote processing may be more widely accessible in the future, it can be best advised to consider the scalability of any data-processing workflows early on. In other words, any workflow designed as a potential big data process will likely have an advantage, as design limitations may prove to be overcome harder than the technical ones.

One point of confusion for readers of big data related literature that often reoccurs is mixing the characteristics of the subject (stored information) with properties of

4 *Cloud computing* enables companies to consume a compute resource, such as a virtual machine, storage or an application, as a utility rather than having to build and maintain computing infrastructures in house (Rouse, 2018). The cloud models include providing infrastructure, platform or application as a service; main vendors of public cloud solutions are Amazon Web Services, Google Cloud Platform or Microsoft Azure.

5 Processing and analytical frameworks designed for big data include Apache Hadoop, Apache Spark, or Apache Flink. No-SQL databases use a column, graph, document, key-value, or multi-model solution as an alternative to traditional relational database design.

technologies used to process it (storage, analytics, visualisation, etc.). It is debatable if this is a fallacy, depending on to what degree we consider digital data independent from the technical infrastructure around it[6]. To illustrate the difference, compare the following two definitions. Fist by Gartner (2018a):

Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

The second by Gantz & Reinsel (2011) defines big data as:

A new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis.

The understanding of big data as an asset prevails, though the second type portraying big data as an ecosystem is not uncommon (e.g. Demchenko, De Laat, & Membrey (2014) or Olshannikova, Ometov, Koucheryavy, & Olsson (2015)). Eventually, this division may lead to dual understanding of big data in narrow sense as a fuel or raw material and in broad sense as an ecosystem, architecture, or framework. A good example of broader thinking is Demchenko et al. (2014) that proposes a "Big Data Architecture Framework" comprised of big data infrastructure, big data analytics, data structures and models, big data life cycle management, and big data security.[7]

---

6 Real world analogies may not be helpful here: for example the properties of gold are independent of the tools used to mine it. On the other hand, many forms of interaction with digital data are inseparable from the technical infrastructure.

7 This is close to holistic definitions discussed later in this chapter, though these tend to be less confined in technology realm and mixing in procedural aspects and wider societal implications.

1.2.3 Showing example sources and quantities

A very common description of big data goes along the lines of "I will give you some numbers and you will get what I mean". Such writing may not provide an exact understanding of the concept, but can put us into context about the scales we are moving at. Doubtlessly the mass of retained data is growing, as McNulty (2014) put it, "90% of all data ever created was generated in the past 2 years" (that was in 2014). In a notable attempt to estimate the World's overall data generation between 1986 and 2007, Hilbert & López (2011) claim that more then 300 exabytes[8] of stored data existed in 2007 (for the methodology of reckoning see Hilbert & López (2012)). The key insight is the growing domination of digital technologies accounting for the majority of the annual growth after year 2000. More recent accounts report on machines potentially capable of processing brontobytes[9] of data (Bort, 2014).

Increasing the storage capacity itself does not speak of any qualitative change in what is stored, therefore some archives could indeed be described as big piles of small data. Under certain circumstances, new quality can arise from increased quantity, for example as Norvig (2011) points out, an array of static images projected at a sufficient frame rate creates an illusion of movement, and hence the new medium also known as film. Multiplication of an old medium creates a new one. The remaining question is under what conditions this change of essence arises, and if such thing occurs or will occur in case of big data. To fast forward a bit, the cartographic version of this question would be: *will a digtal map based on big data (fast and n=all) be essentially different from web maps based on static and sampled data sources?*.

---

8 1 exabyte = 1 000 000 000 gigabytes

9 1 brontobyte = 1 000 000 000 exabytes

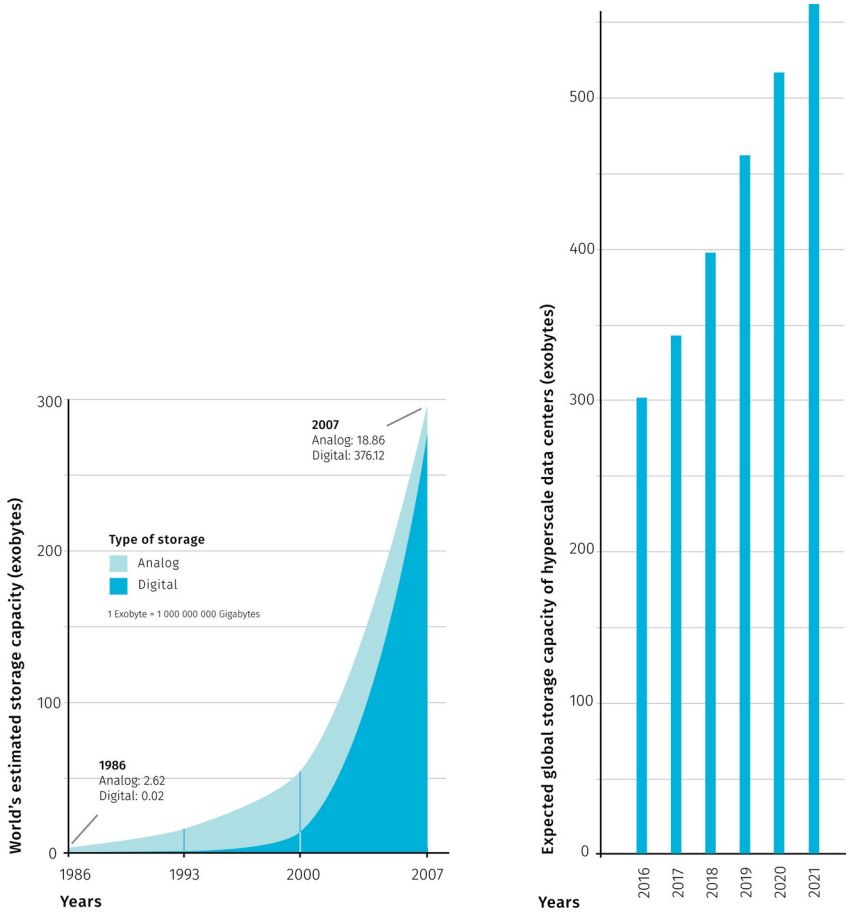

Fig.1 Comparison of the World's estimated data storage capacity between years 1968 and 2007 (modified after Hilbert & López (2011)) and the expected storage capacity of large scale data centers in the period from 2016 to 2021 (modified after Networking (2018))

Rather than putting up to a gargantuan task of counting the mass of all existing data items, authors use the available statistics related to operations of large companies (Kambatla, Kollias, Kumar, & Grama (2014), McNulty (2014), Marr (2014) and others). For example, Facebook was said to process 10 billion messages, 4.5 billion button clicks and 350 million picture uploads each day (Marr, 2014). It goes without saying these numbers are outdated and certainly outgrown today. Other companies prominently mentioned in context

of big data are Google, Wallmart, or Amazon. This connection is justified, as these companies have put user (or customer) data analytics to the core of their businesses, thus supporting the progress in the field. Social media, web search and browsing data, online or offline shopping patterns, but also mobile devices, sensors and large scientific projects are mostly named as generators of big data.

Another quantity tying to big data that is surely of interest is, according to estimates potentially huge, market value. For example Kayyali, Knott, & Van Kuiken (2013) reports on promise in reduced health care costs of 12 to 17 percent thanks to emerging big data related initiatives in USA health care. On the other hand, the use of poor data is also estimated to have vast impacts on businesses, mainly in form of unrealized opportunities (McNulty (2014)). Another financial aspect is the costs incured by creating and maintaining big data itself, it is sound to remind that apart from all the promise, big data also has the potential to cost unlimited amounts of money Fischer (2015).

The type of data source is another classification property. Authors distinct “traditional” ways of collecting data from the new, technology-powered sources. The definition of big data then comes as simple as data coming from these new sources. The United Nations Economic Commission for Europe proposed a taxonomy that recognizes three main sources of big data (UNECE (2013)):

- *Social Networks (human-sourced information)* — this information is the record of human experiences
- *Traditional Business systems (process-mediated data)* — these processes record and monitor business events of interest

- *IoT (machine-generated data)*[10] — information is derived from sensors and machines used to measure and record the events and situations in the physical world

Data sources labeled as big differ from traditional sources such as surveys and official administrative statistics — Florescu et al. (2014) and Kitchin (2015) closely examine those differences as well as the potential for big data to extend the official statistics. Interesting point is that volume is not actually distinctive as governmental offices tend to store large amounts as well. What makes the difference is that classical data sources have statistical products and by-products specified beforehand, big data tend to be reused beyond the original intent. On the other hand, big data sources tend to be volatile and unstructured, therefore their representativeness is harder (if possible) to assess.

The estimation in the fig1 couldn't have predicted the spread of COVID-19 pandemic. According to International Data Corporation (IDC), more than 59 zettabytes (ZB) were to be created, captured, copied, and consumed around the world in 2020. The COVID-19 pandemic contributed to this figure by causing an abrupt increase in the number of work from home employees and changing the mix of data being created to a richer set of data that includes video communication and a tangible increase in the consumption of downloaded and streamed video. IDC also measures the amount of data created and consumed in the world each year. The ratio of unique data (created and captured) to replicated data (copied and consumed) is roughly 1:9, and it is expected to move to 1:10 by 2024. This trend is also fuelled by increased consumption of replicated data due to COVID-19 pandemic. (International Data Corporation, n.d.)

---

10 Internet of Things (IoT) can be described as a vision of a network of devices, vehicles and home appliances that can connect, interact and exchange data. Similarly to big data, there are manifold definitions of the concept, for overview see Atzori, Iera, & Morabito (2010)

1.2.4 Metaphors

Metaphors rely on a notion of analogy between two dissimilar things, but can also become independent verbal objects, aesthetically appealing but not overly revealing. Despite that, we should not ignore metaphoric accounts as they contribute to the mythology surrounding big data that reflects what many people expect.

Puschmann & Burgess (2014) identified two prevailing ways of imagining the subject: big data seen as a *natural force* to be controlled and as a *resource* to be consumed.

The utilitarian mindset comparing digital world to excavation of valuable minerals in far from new (think of "data mining" or more recently "cryptocurrency mining") but it is tempting pursue to this analogy further. For example, how to estimate the ratio of valuable information to "debris", and shouldn't such estimation be done before any data "mining" endeavour? The value of real-world analogies may be in provoking some common-sense reasoning often missing in wannabe-visionary proclamations.

For example mayer2013big: "Data was no longer regarded as static or stale, whose usefulness was finished once he purpose for which it was collected was achieved [...]. Rather, data became a raw material of business, a vital economic input, used to create a new form of economic value. Every single dataset is likely to have some intristic, hidden, not yet unearthed value...". So what is yet to be unearthed is not the data itself but new way of using it.

As Lupton (2013) notes, by far the most commonly employed rhetorical descriptions of big data are those related to water or liquidity, suggesting both positive and negative connotations. For example Manyika et al. (2013) argues for unlocking data sources to become "liquid" in a sense of open and free-flowing, at the same time keeping privacy concerns

in mind — what is liquid is also susceptible to unwanted leaks.

Big data has also been described as a *meme* (a unit of cultural transmission) and as a *paradigm* (a set of thought patterns), in both cases not without certain concerns. Gorman (2013) explores big data as a technologic meme: "[t]he reductionist methods of understanding reality in big data produce new knowledge and methods for the control of reality. Yet it is not a reality that reflects the larger society but instead the small minority contributing content." To Graham & Shelton (2013) "big data could be defined as representing a broader computational paradigm in research and practice, in which automated algorithmic analysis supplants domain expertise".

Of course, big data descriptions are not limited to verbal form, visual means can be much more expressive and informative — not a surprising claim to be found in a thesis on visual analytics. We will discuss cartographic tools later, here we can mention artistic renderings that employ more free-form visual analogies. We should distinguish pursuits like *information visualisation* that are close to graphic design (for good overview see Klanten, Ehmann, Bourquin, & Tissot (2010) or Lima (2011)) from artistic projects that use data as a raw material and don't aim to convey information or comfort to general user's cognitive expectations (like some projects at Network (2018)). From the cartographer's standpoint, aspects of visual art can be inspiring (graphic quality, employment of computation and rendering software, creative uses of interaction and animation), though artistic means are often too different to be transposed. Without referring back to the source phenomenon, data-driven art becomes unrecognizable from the generative art that uses artificially generated data rather than any existing information.

1.2.5 Holistic accounts

Multifaceted phenomena tend to provoke descriptions that narrowly focus on specific components, ignoring other parts as well as relationships between them. Experts of different specializations notice aspects of phenomena that are close to their research interests and priorities, cross-disciplinary definitions then try to combine these views to paint the full picture. Naturally, listing holistic accounts will include topics already mentioned, therefore pardon some repetition in this section.

For instance Murthy, Bharadwaj, Subrahmanyam, Roy, & Rajan (2014) prepared a taxonomy of big data comprised of:

- *data* — with various levels of temporal latency and structure
- *compute infrastructure* — batch or stream processing
- *storage infrastructure* — distributed, sql or nosql databases
- *analysis* — supervised, semisupervised, unsupervised or reenforcement machine learning
- *visualisation* — maps, abstract, interactive, real-time
- *privacy and security* — data privacy, management, security

As another example, Boyd & Crawford (2012) define big data as a "cultural, technological, and scholarly phenomenon that rests on the interplay of":

- *technology* — maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets
- *analysis* — drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims

- *mythology* — the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy

As the two taxonomies above illustrate, there are many ways to slice a cake. The fate of overreaching definitions is that they are often too intricate to explain the phenomena crisply, yet they are never complete as there is always a point of view that hasn't been included yet. So here we arrive at a trade-off between preciseness of a definition and its practicality. One way out of this is simply rejecting the view of big data as a singular phenomenon. Big data is then a non-specific covering term that could mean different things to different people. As Helles & Jensen (2013) observes, "[d]ata are made in a process involving multiple social agents — communicators, service providers, communication researchers, commercial stakeholders, government authorities, international regulators, and more. Data are made for a variety of scholarly and applied purposes [...]. And data are processed and employed in a whole range of everyday and institutional contexts." The process, the actor, the purpose and the context then determine what big data "is" in that given constellation.

We can conclude the section on holistic approaches with a historical view that is rarely taken in commentaries on the nature of big data, probably because the perceived novelty of the concept. For Barnes (2013) "[b]ig data has been made possible because of the particular conjuncture of different elements, each with their own history, coming together at this our present moment. But precisely because these different elements have a history, the issues, problems and questions that were there in their earlier incarnation can remain even in the new form". We can add that some issues can get worse in the new incarnation and totally new set of problems can arise. For example, as Mayer-Schönberger &

Cukier (2013) note, current anonymization techniques can be rendered ineffective as combining several "data traces" of online activity can still identify the person. Or, as Taleb (2012) realizes, if big data come with too many variables but with too little data per variable, it becomes nearly impossible not to find high but spurious correlations, which can tempt researchers to cherry-pick the results that "support" their hypothesis. Considering wider implications of technology can potentially make such unintended effects less surprising, which is certainly a virtue of holistic thinking.

1.3 Spatial big data

Apart from the general definitions mentioned above, there have also been field-specific efforts to contextualize big data. The fields include governance (Crampton, 2015), journalism (Lewis & Westlund, 2015), ecology (Shin & Choi, 2015), social sciences (Ovadia, 2013), business administration (Wamba, Akter, Edwards, Chopin, & Gnanzou, 2015), urban studies (Thakuriah, Tilahun, & Zellner, 2017), learning analytics (Wilson et al., 2017), education (Kabakchieva, Stefanova, & others, 2015), health informatics (Herland, Khoshgoftaar, & Wald, 2014) and doubtlessly many others. Authors here consider existing data processing and analytical practices in their respective disciplines in light of possibilities created by big data. Some expect forthcoming changes such as enrichment in available methods (e.g. analysing social networks in epidemiology), others analyze the adaptability of currently used processes to conditions of higher data load. With some generalization, the overall mood of these works seems to be welcoming towards big data as a possible toolbox extension, though doubting that the core scientific methods could be deeply altered by it. When it comes to defining big data, field-specific accounts use one or more of

the aforementioned definitions by *keywords*, *constraints*, *examples*, *metaphors* or combination of all in a*holistic* description.

Within geography, Kitchin (2013) highlights possible opportunities, challenges and risks posed by big data, encouraging geographers to engage in big data related case studies. He also lays some groundwork for definitions, he later developed into ontological characteristics cited at the beginning of this chapter. González-Bailón (2013) understands big data predominantly as a rich set of observations of intricate and nested social life that can improve theories of human geography, for example by exposing diversity that would otherwise go unnoticed in scientific models. Barnes (2013) reminds us of the so called *quantitative revolution* in geography (starting from 1950's) that besides bringing many good to the discipline has also been criticized on various levels. Some of this critique, Barnes argues, "continue[s] to apply to the *über* version of the quantitative revolution that is big data". For Goodchild (2013) geography provides a distinct context for discussion about what kinds of science might be supported by big data. He is also concerned with the potential for building rigorous quality control and generalizability into big data operations, because so far "instead of relying on the data producer to clean and synthesize, in the world of big data these functions are largely passed to the user". We could go on much further with how geographic thought internalizes big data, those interested in the topic may refer to Thatcher, Shears, & Eckert (2018).

Cartographers and GIS practitioners like to say that 80% of all data is geographic, and even though such claim is hard to prove[11], few would doubt that spatial reference can unlock additional value, if only as a platform for joining otherwise

11 see Morais (2012) for discussion and Hahmann, Burghardt, & Weber (2011) for a validation attempt

un-joinable datasets. Much of data in the world is or can be georeferenced, which underlines the importance of geospatial big data handling.

Cartography and geographic information science have both developed distinct and elaborate notions of data in general. Scientists and practitioners from these fields are in good position to contribute to the way big data is understood and utilized, given their focus on the space as a unifying factor and with visual analysis being at the core of their practice. For these reasons, we will first take an aside to briefly outline how cartography and geoinformatics conceptualize spatial data, before moving on to how the disciplines contended with the adjective big. We consider the following points important:

- Data describing spatial phenomena used in GIS are traditionally divided into *spatial* and *non-spatial* (thematic, attribute) components. Spatial component holds information on location and geographic extent of an entity and can be thought of as a geometry that is visualized on a map or used for spatial analysis (spatial querying, overlay algebra, network analysis, etc.). Attribute information can be used to set visual parameters of geometries on a map as well as in spatial analysis. Visualising attributes lets us observe the variability of a phenomenon across the area of interest. Andrienko & Andrienko (2006) offer more general view of data as a correspondence between referential and characteristic components. Referential components (or referrers) are described as independent variables — mostly employed referrers are *location*, *time* and *population*. Referrer or a combination of referrers provides context and unique identification for dependent variables — attributes.

- Literature distinguishes two approaches to representing the spatial component of data in GIS: *object-based* and *location-based* (Peuquet, 1994). The object-based approach arranges spatial and non-spatial information into discrete geographic objects (features). In the location-based approach, attribute information is stored relative to specific locations. With this approach, a territory is divided into same-size elements that represent locations to assign attributes to. Object-based approach is manifested in *vector data model*, location-based approach corresponds to *raster data model*. In vector data model objects have either point, line or polygon representation. Objects are usually grouped into layers of same theme and geometry type. In raster data model, representation is defined by the size of the element (almost always being a rectangular pixel). Raster model suits better for displaying spatially continuous phenomena, whereas vector model tends to be more appropriate for discrete objects, though reverse situation is not uncommon and transformation between models is a frequent practice.

- Attributes are typically distinguished according to the levels of measurement introduced by Stevens (1946): *nominal* (named variables), *ordinal* (allow ordering), *interval* (allow measuring difference), and *ratio* (having natural zero). Jung (1995) proposed an alternative classification more tailored to spatial data handling: *amounts* (absolute quantities), *measurements* (quantities requiring units of measurement), *aggregated values* (amounts or measurements summarized by area), *proportional values* (normalised by a fixed value), *densities* (divided by corresponding area), *coordinates* (position in some coordinate system).

- The temporal aspect of a phenomenon includes the existence of various objects at different moments, and

changes in their properties (spatial and thematic) and relationships over time (Andrienko & Andrienko, 2006). Including the temporal aspect into the data model is problematic as it is treated separately from spatial and attribute components despite having influence on both. For the attribute part, the time changes can be stored by adding table columns with new values. However, changes in the spatial component are not easily stored, which complicates linking the past forms of geometries with corresponding past values of attributes[12]. Incorporating flexible time changes into GIS data model remains a challenge for spatialization of big data.

- Spatial component of data may be displayed at various scales. The scale along with the purpose of the map influences the level of comprehensible detail in displayed geometry. Cartographic generalisation is the process of adjusting the map geometry to the spatial scale in which the area is displayed. This goes beyond mere simplification, as factors as *highlighting the important*, *maintaining the object relationships* and *preserving the aesthetic quality* come to play. The dynamic change of scale comes naturally to users of digital interfaces, the generalization is however hard to automate as it involves complex reasoning and considerations of object relationships that span through the strict topic-based separation of layers common in spatial datasets[13]. The same phenomenon can be studied at various levels of detail even without changing the

12 This is most pressing when handling spatial data in discrete files (e.g. in Shapefile or GeoJSON formats). Using versioning systems like Git, which has become incredibly popular for handling software source code and text files, is not suitable for spatial data files as these often exceed repository size limits (though there is a project attempting to solve this called *geogig* http://geogig.org/). Handling spatial data within relational database provides more options for spatial data versioning, also there is a range of database project specialized on storing time series like InfluxDB or TimescaleDB.

scale of the map. Some spatial datasets, such as administrative units, exhibit the nesting property that allows to vary the granularity of the displayed spatial pattern.

The above summary is inevitably simplistic as there are many other research areas in cartography and GIS that are relevant to big data efforts. Some will be touched on later in the thesis, others are unfortunately out of its scope. One such case for all is spatial imagery that is an example of truly big data source that is inherently spatial. "Big" in this case means unprecedented spatial, temporal and spectral resolutions brought about by improvements in global monitoring systems.

In light of big data advent, authors form spatial fields consider what difference does it make to conceptualize a specifically *spatial* big data as opposed to big data per se. Is spatial big data a subset or an extension of big data? From the GIS point of view of view there are two ways of understanding spatial big data: either as *adding a spatial reference to big data* or as *adjusting the current spatial data models and processes to higher data load*. We can say that these two approaches arrive at the concept of spatial big data form the opposite sides, in the first case the path is *from big data to spatial big data*, whereas in the second case it is *from spatial data to spatial big data*.

Authors from the first group use some of the previously mentioned definition styles. For example to Jiang & Shekhar (2017), spatial big data refer to "georeferenced data whose volume, velocity, and variety exceed the capacity of current spatial computing platforms". This combines definitions by V-words and computational difficulties. Lee & Kang (2015), on the other hand, combines definition by constraints and

13 for more on efforts in automated generalisation see for example Burghardt, Duchêne, & Mackaness (2016)

by example. In this context we can mention some early critique that condemned narrow understanding of big data, aiming mainly at analyzing geotagged social media content (labeled as “burger cartographies” by Crampton et al. (2013) and Shelton (2017)). As Leszczynski & Crampton (2016) note, social media content covers just a limited facet of the data productions, presences, and practices that fall under spatial big data.

Representing the second group, Yao & Li (2018) recognizes five categories of spatial big data (while admitting some intersections): *remote sensing data*, *large data from surveying*, *location-based data from mobile devices*, *social network data*, and *Internet of Things (IoT) data*. Yao and Li then focus on a subgroup they name *big spatial vector data* (BSVD), and provide a comprehensive survey of techniques applicable for managing such data. In short, adjusting the vector spatial data model for distributed storage impacts how the data is indexed[14] and queried for processing and application. Yao & Li (2018) also provide an overview of other authors’ approaches to thinking about GIS in the era of big data.

In context of transportation, Shekhar et al. (2012) distinguish between *traditional* and *emerging* spatial big data. Traditional stands for topological vector data representing transportation infrastructure, emerging represents sensor and positional data from large number of vehicles — termed as *spatio-temporal engine measurement data*. Shekhar, Evans, Gunturi, Yang, & Cugler (2014) call for performance testing of existing and new algorithms to assess proper comparison between spatial big data processing techniques.

14 Spatial indices are used to optimize retrieval of spatial data from database. They decrease the time it takes to locate features that match a spatial query.

To Li et al. (2016), main sources of spatial big data are in *volunteered geographic information (VGI)*[15] and in *geo-sensor networks* (with extended understanding of sensor including CCTV and mobile devices). Li et al. (2016) also touches on a wide range of topics, ranging from quality assessment (big data properties challenge the current error propagation methods) to the importance of parallel processing of data streams (where the advantages of functional programming languages are recognized). Zee & Scholten (2014) mentions the *Internet of Things* concept as a main future source of big data — here understood as a sum of sources from "smart" devices. Geospatial technologies are considered a binding principle that would eventually help to meaningfully combine data from devices to facilitate the rise of smart city[16].

In relation to big spatial data processing, we should mention the work of Bin Jiang that is somewhat isolated from the categories mentioned above, but provides interesting thought on how the current GIS processes could be altered. Jiang (2018) recognizes the following dichotomies and potential paradigm shifts:

15 VGI is defined as "the harnessing of tools to create, assemble, and disseminate geographic data provided voluntarily by individuals" (Goodchild, 2007). This description fits for example the contributions to the Open Street Map project very well, but is less applicable to social media, where users are more likely indifferent to their data being collected, rather than contributing data as a primary goal.

16 Smart city is a concept of urban area that uses digital information to make more efficient use of physical infrastructure, engage effectively with people in local governance, and respond promptly to changing circumstances. For more information see McLaren & Agyeman (2015)

- *Gaussian* vs *Paretian statistics*[17] — the first suits better for sets with elements of more or less similar size and expects normal distribution, the latter is based on the notion of far more "smalls" than "larges" and expects Poisson or other fat-tailed distribution.
- *Tobler law* vs *scaling law* — complementary concepts, where the first expects inverted proportionality between the distance and similarity of objects, which is often justified locally but does not attribute to abrupt spatial heterogeneity brought about by fat-tailed distributions. Scaling law, as Jiang formulates it, accounts for uneven distributions across scales.
- *Euclidean* vs *fractal (natural) geometry* — the first is needed "to measure things", the second can help us to "develop new insights into structure and dynamics of geographic features". (Jiang & Brandt (2016))
- *data quality* vs *data character* — Jiang defines data character mainly as topological relationships between meaningful geographic objects (e.g. connectivity of street network), which for many purposes can be more important than the precision of geometric primitives.
- *mechanistic thinking* vs *organic thinking* — the latter promotes the understanding of geographic space as a living structure shaped by the interaction of elements at various scales.

Though some of Jiang's distinctions may seem unclear and he is silent about how to incorporate organic approaches to GIS data models, he recognises that big data would be vital in changed GIS practices. For example in his notion of natural cities, social media data are used to define the "natural" extent of the city, so a city is understood more as a

17 Named after Vilfredo Pareto who more than century ago noticed that in 20% of people in Italy owned 80% of land. The ratio of 20% of causes leading to 80% of consequences has been observed in many systems, though the distributions can be far more uneven, like that 99% of Internet traffic is attributable to 1% of sites (Taleb, 2012).

bottom-up emergence rather than a top-down administrative demarcation.

As we have seen in this section, geospatial authors rarely diverge from general definitions of big data, but when it comes to spatial big data, they consider the topic from the standpoint of pre-existing theory generated in the field. This conscious assessing of current data models and processes and possible creation of new ones can bring interesting developments in the future.

The potential role of cartography will be examined in more detail later in the thesis, here let us briefly go over the big data properties listed at the beginning of the chapter to see the most obvious cartographic concepts and challenges that could possibly tie to them:

- *Extensionality & Indexicality* — spatial reference in itself is a unifying platform to combine data from various sources and map is a proven tool to explore spatial interrelations. From the perspective of data processing workflows spatial extensionality poses a challenge for geocoding services to spatialize previously unchartable data. From the map design perspective the task is to support recognition of spatial co-ocurrence in dense displays. Indexicality is a natural prerequisite for thematic mapping.
- *Volume* — from the cartographic standpoint, the number of records is the most interesting measure of volume (compared to storage size or attribute length). Extensive volume does not necessarily present a problem for effective visualisation, especially if it plays out in the attribute space and the spatial reference is static. Maps that use the right visualisation methods naturally support information compression and clarification.

- *Scalability & Resolution* — adjusting visualisation to different scales both in terms of spatial extent and in terms of data load is a domain of cartographic generalization. Effects of varying time, space, and attribute resolution on displayed information has long been studied within cartography.
- *Variety* — digital mapping requires some structure in data, though it is not a requirement for attributes as long as the spatial reference is valid. There is though a gap in incorporating unstructured data to digital mapping, for example in adjusting metadata profiles (e.g. move from hierarchical classification to messier but more flexible methods like tagging), or in determining data quality from spatial context. Cartography is in a good position to search for ways to combine structured and unstructured data in meaningful way.
- *Velocity & Exhaustivity* — these parameters will be dealt with in chapters 4 and 5, they relate to a large set of topics internal to cartography. Velocity is mainly concerned with rate of visualization update and time span of the depicted theme. Cartography is ideal for depicting time-space regularities and relationships within and between datasets. Exhaustivity then projects into the longtime problem of graphic fill and tailoring cartographic visualisation to human cognitive capabilities.

It is not within the scope of this thesis (and within the author's powers) to consider all directions and areas where cartography and geographic information science may be impacted by big data. The whole project of GIS might need to to be rethinked again, but this is not unprecedented. From the desktop GIS (1960s) to the web GIS (1980s), and the distributed GIS (1990s), to the cloud GIS (2010s), it is well known that the development of GIS is greatly influenced by

computer science technology (Yang, Raskin, Goodchild, & Gahegan (2010)). Another turn in might come as a response to big data.

1.4 Assessing impacts, threats and opportunities

Often times big data are described indirectly by the impacts (real or imagined) they have on the society. For some authors, the debate on the definition of big data may be dismissed as unproductive. The popularity of the term itself may diminish like many other buzzwords that went through the technology hype cycle.[18] Many ideas in the IT industry exist under changing or concurrent names, and big data have indeed a lot in common with concepts such as *data mining*, *business intelligence* or *visual analytics* to name just a few. For many the term is just too underdefined and overused. But we should not forget that even though the technological industry is largely fashion-driven, its societal impacts are real, event though at times unevenly distributed.

It is beyond the scope of this thesis to consult all of these impacts in detail (for such discussions see Bollier & Firestone (2010), Swan (2015), or Mayer-Schönberger & Cukier (2013)), though the puzzle of big data definitions would miss an important piece without touching on some of the consequences in *scientific inference* and *knowledge-based decision making* — the areas cartography aims to support. Closely related are the issues of *surveillance* trough big data and the *emerging digital divides*.

18 Hype cycles describe how expectations from emerging technologies evolve with time. Stages in the cycle are: *innovation trigger*, *peak of inflated expectations*, *trough of disillusionment*, *slope of enlightenment*, and *plateau of productivity*. The expected duration of the cycle differs per technology, and some technologies may not reach productivity in the foreseeable future. Hype cycles are a construction of the Gartner consultancy that issues regular reports, see for example Gartner (2018b)

The scientific reflection on big data revolves mainly around the question if the advances in data acquisition change the definition of knowledge. The anticipated mindset changes voiced in mayer2013big can be summarized into the following points:

- Reduced need for sampling with accessibility of n=all datasets
- Loosened requirements for exactitude as minimizing sampling errors would leave room for more relaxed standard for measurement error (will to sacrifice a bit of accuracy in return for knowing the general trend faster)
- Departure from the search for causality: “big data is about *what* not *why*.” Multi factor correlation with large data enables decision making even without understanding the mechanisms behind the relationship. In words of Anderson (2008): “Who knows why people do what they do? The point is they do, and we can track it and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.”

Correlation does not necessarily imply causation, though if we do not aim for understanding the phenomenon and just want to obtain some instruction for action, correlation might be enough to provide some backing. For the optimistic commentators, this abandoning of theory can open door to iterative experimentation and building of useful heuristics that are independent of preconceptions and biases of our thought processes. To others, this sounds scary at best, as such naive data appreciation can dangerously rationalize incompetent guesswork. As Silver (2012) puts it, most of the data is just noise, as most of the universe is filled with empty space.

Claims to objectivity and accuracy of big data are often criticized as misleading, numbers obviously never do the speaking, as there is always a need for human interpretation. For such interpretation bigger data are not always better. For example, multidimensionality of datasets can increase probability of spurious correlations. Data-driven rhetoric can be suspicious as it allows decision makers to evade responsibility or to ignore alternative decisions. Furthermore, in decision making under opacity, over-reliance to historical records can catch us ill-prepared for unprecedented large scale events (so called black swans). Despite the air of progress and innovation Barnes (2013) sees big data as an inherently conservative project: “By utilizing the numbers as they are given, big data is stuck with what is rather than what should be”. In both innovation and risk management, *imagination* is the vital virtue, that is something big data cannot supplant.

The proposition of theory-free science using powerful exploratory potential of big data to opportunistically exploit new avenues as they appear sounds promising to many. Though there is no need to discard hypotheses as they are generated inevitably in some form and can be modified dynamically in the research process. In words of P. Gross: “In practice, the theory and the data reinforce each other. It’s not a question of data correlations versus theory. The use of data for correlations allows one to test theories and refine them.” (Bollier & Firestone (2010))

Apart from possible fallacies (like *more is better* or *big data = smart data*), there is a philosophical concern of *representational authenticity* (Swan (2015)) — the degree to which the representation (in this case big data) corresponds to the represented (onthology) as well as how to measure this correspondence (epistemology). Any mode of interacting with big data is representation and not necessarily reality, and the reality gap may be so big that

data however big might not be relevant (Siegfried (2013)). In words of uprichard2013focus: “If we are creating a mess by generating so many haystacks of big data that we are losing all the needles, then we need to figure out a different kind of way of doing things, as we cannot sew new cloth without any needles. Whatever else we make of the ‘big data’ hype, it cannot and must not be the path we take to answer all our big global problems. On the contrary, it is great for small questions, but may not so good for big social questions.”

The critical accounts however do not negate big data as a tool, rather they dismiss the shallow reflection of its usage. As a good outcome, such discussions can strip bare our conceptual gaps and turn our attention the right direction. Big data can then be leveraged to support an optimistic goal, for to create *overreaching predictive mathematical frameworks for complex systems* (West (2013)). Big global issues in ecology, pandemics or financial markets exhibit traits of complex systems[19]. “The trouble is, we don’t have a unified, conceptual framework for addressing questions of complexity. We don’t know what kind of data we need, nor how much, or what critical questions we should be asking. ‘Big data’ without a ‘big theory’ to go with it loses much of its potency and usefulness, potentially generating new unintended consequences” (West (2013)). All things considered, “[...] the arrival of Big Data should compel scientists to cope with the fact that nature itself is the ultimate Big Data database. Old style science coped with nature’s complexities by seeking the underlying simplicities in the sparse data acquired by experiments. But Big Data forces scientists to confront the entire repertoire of nature’s nuances and all their complexities” (Fan, Han, & Liu (2014)).

---

19 Complex system’s collective characteristics cannot easily be predicted from underlying components: the whole is greater than, and often significantly different from, the sum of its parts. A city is much more than its buildings and people. Our bodies are more than the totality of our cells. This quality, is called *emergent behavior*. West (2013)

The aforementioned discussions point to lock-step evolution of science and technology, and most importantly, to strong reflection and self-correcting mechanisms inherent to science that usually set in motion when innovation is accompanied with some troubling signals[20]. In broader society we also need such a reflection of new realities created by big data and the accompanying ethical issues.

One set of ethical issues revolves around data collection without giving people the choice to opt out, or without asking for explicit and informed consent. Even if consent is solicited, for users it is often impossible to audit the secondary uses that the collected data will cater to. It is hard to track what additional sources and analytical engines will be applied on collected user data and what third parties will get hold of it through reselling. At the time of writing, the legislation to address these issues is catching up[21], but it is unsurprising that it lags behind the new kinds of abuse stemming from the extending scope of personal information that can be collected. Even with legislation in place, enforceability is low and even learning about misuse is difficult without the rare help from whistleblowers.

Furthermore, the anonymization methods may no longer work as combining digital traces from several sources allows for re-identification of an individual. Another topic is the ability of user to access the collected data, either to use it for own self-analysis, or to issue its removal (tough how to verify is has actually happened?). In an alternative vision of

---

20 For other examples of such reflections see Lipton & Steinhardt (2018), Norvig (2012)

21 Legislation varies around the world, for European Union, the General Data Protection Regulation (GDPR), which governs how personal data of individuals in the EU may be processed and transferred came into being in 2018. For overview of digital privacy rules see https://europa.eu/youreurope/citizens/consumers/internet-telecoms/data-protection-online-privacy/index_en.htm.

big data economics, individuals may gain power to sell their data themselves of through intermediaries.

Penalties based on propensities — that is a short description of a concern that with increased surveillance and predictive analytics there will be a possibility to issue preventive penalties for offences that did not happen yet solely based on individual's observed tendencies (similarly to the movie Minority report) (Mayer-Schönberger & Cukier (2013)). It is a fact that the technical infrastructure for close personal scrutiny and behaviour enforcing has been already implemented at the scale of a warehouse (Amazon Head (2014)) as well as a country (most (in)famously in China), with little room for individuals to object. At the time of this writing, the global pandemics of COVID-19 created a justification for public scrutiny at unprecedented levels, on the other hand laid bare the inability of some state apparatuses to recast their data stacks into meaningful action.

Social media has created a new platform that apart from all good created unexpected avenues for illicit actions, sometimes at a scale that can shake up a state. Fake news, troll farms, data breaches used to manipulated election results are all examples of the weaponization of the platform. Data literacy is then one of the prerequisites for defence against malicious effects on one side and to make the most of the data availability on the other. In words of D'Ignazio (2017): "[...] although there is an explosion of data, there is a significant lag in data literacy at the scale of communities and individuals. This creates a situation of data-haves and have-nots. But there are emerging technocultural practices that combine participation, creativity, and context to connect data to everyday life. These include citizen science, data journalism, novel public engagement in government processes, and participatory data art."

The definition of big data is elusive perhaps also because the majority of involved actors, being positioned in the business world, is more focused on building productive big data ventures without much conceptual attention to the subject in itself. Then of course, the underlying technologies become a subject of marketing which often uses inflated overstatements based on expectations rather than reality. So far there is no settled consensus around big data definition in the academia either, but as Kitchin & McArdle (2016) predict, the "genus" of big data will probably be further delineated and its various "species" identified. The question is if then such an umbrella term will be necessary. Anyways, the lack of common ground in understanding what big data is (illustrated by this chapter) may be a good predictor of the term's future relevance. Problems with definition is exactly what leads Davenport (2014) to predict "a relatively short life span for this unfortunate term". Indeed, looking at the peak of big data excitement in publications that took place around 2014 from the current perspective, the hype moved towards machine learning that gets inflated nowadays. On the other hand, the number of researchers and practitioners willing to invest their time in big data related endeavours is relatively high[22], which sheds some positive light on the future vitality of the concept.

To Mayer-Schönberger & Cukier (2013) big data stand for "the ability of society to harness information in novel ways to produce useful insights or goods and services of significant value". Here, more than an exact definition, the importance lies in the real-life impacts that are likely to stay even when the big data hype is over. Even if we dismiss the term as a buzzword, the fact that more digital information gets created and can be linked more easily has many

22 *Journal of Big Data*, *Big Data Research*, *International Journal of Data Science and Analytics*, *Big Data & Society*, *Big Data Analytics*, *Big Data* are examples of scientific journals tracking cross-disciplinary efforts in the field.

implications on the way we live. Together with that, there are changing attitudes to putting data to work. In the next chapter, we will look at how we can derive insight from big data as well as on the possible role cartography can take in these endeavours.

Sources

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, *16*(7), 16–07.

Andrienko, N., & Andrienko, G. (2006). *Exploratory analysis of spatial and temporal data: A systematic approach*. Springer Science & Business Media.

Atzori, L., Iera, A., & Morabito, G. (2010). The internet of things: A survey. *Computer networks*, *54*(15), 2787–2805. Elsevier.

Barnes, T. J. (2013). Big data, little history. *Dialogues in Human Geography*, *3*(3), 297–302. SAGE Publications Sage UK: London, England.

Boellstorff, T., & Maurer, W. (2015). Introduction. In *Data, now bigger and better!* (pp. 1–6). Prickly Paradigm Press.

Bollier, D., & Firestone, C. M. (2010). *The promise and peril of big data*. Aspen Institute, Communications; Society Program Washington, DC.

Bort, J. (2014). There's a new word being used in the computer industry: 'Brontobytes'. *Available online at http://www.businessinsider.com/new-big-data-word-brontobytes-2014-6 (last accessed May 30, 2018)*.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, *15*(5), 662–679. Taylor & Francis.

Brunelli, M. (2011). Will your organization benefit from 'big data' processing technology? *Available online at searchdatamanagement*.

techtarget. com/news/2240036228/Will-your-organization-benefit-from-big-data-processing-technology (last accessed December 29, 2016).

Burghardt, D., Duchêne, C., & Mackaness, W. (2016). *Abstracting geographic information in a data rich world*. Springer.

Crampton, J. W. (2015). Collect it all: National security, big data and governance. *GeoJournal*, *80*(4), 519–531. Springer.

Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., & Zook, M. (2013). Beyond the geotag: Situating 'big data'and leveraging the potential of the geoweb. *Cartography and geographic information science*, *40*(2), 130–139. Taylor & Francis.

Davenport, T. (2014). *Big data at work: Dispelling the myths, uncovering the opportunities*. Harvard Business Review Press.

Demchenko, Y., De Laat, C., & Membrey, P. (2014). Defining architecture components of the big data ecosystem. In *Collaboration technologies and systems (cts), 2014 international conference on* (pp. 104–112). IEEE.

Diebold, F. X., Cheng, X., Diebold, S., Foster, D., Halperin, M., Lohr, S., Mashey, J., et al. (2012). A personal perspective on the origin (s) and development of "big data": The phenomenon, the term, and the discipline∗. Citeseer.

D'Ignazio, C. (2017). Creative data literacy: Bridging the gap between the data-haves and data-have nots. *Information Design Journal*, *23*(1), 6–18. John Benjamins.

Dodge, M., & Kitchin, R. (2005). Codes of life: Identification codes and the machine-readable world. *Environment and Planning D: Society and Space*, *23*(6), 851–881. SAGE Publications.

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National science review*, *1*(2), 293–314. Oxford University Press.

Fischer, D. (2015). Why exploring big data is hard and what we can do about it. *Available online at www. youtube. com/watch? v=UP5412nU2lI (last accessed December 29, 2016)*.

Florescu, D., Karlberg, M., Reis, F., Del Castillo, P. R., Skaliotis, M., & Wirthmann, A. (2014). Will “big data” transform official statistics? In *Q2014–european conference on quality in statistics*.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. Elsevier.

Gantz, J., & Reinsel, D. (2011). Extracting value from chaos. *IDC iview*, *1142*(2011), 1–12.

Gartner. (2018a). What is big data? - gartner it glossary. *Available online at https://www.gartner.com/it-glossary/big-data/ (last accessed October 26, 2018)*.

Gartner. (2018b). Gartner special reports. *Available online at https://www.gartner.com/en/products/special-reports (last accessed August 26, 2018)*.

González-Bailón, S. (2013). Big data and the fabric of human geography. *Dialogues in Human Geography*, *3*(3), 292–296. SAGE Publications Sage UK: London, England.

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*(4), 211–221. Springer.

Goodchild, M. F. (2013). The quality of big (geo) data. *Dialogues in Human Geography*, *3*(3), 280–284. SAGE Publications Sage UK: London, England.

Gorman, S. P. (2013). The danger of a big data episteme and the need to evolve geographic information systems. *Dialogues in Human Geography*, *3*(3), 285–291. SAGE Publications Sage UK: London, England.

Graham, M., & Shelton, T. (2013). Geography and the future of big data, big data and the future of geography. *Dialogues in Human Geography*, *3*(3), 255–261. SAGE Publications Sage UK: London, England.

Hahmann, S., Burghardt, D., & Weber, B. (2011). “80% of all information is geospatially referenced”??? Towards a research

framework: Using the semantic web for (in) validating this famous geo assertion. In *Proceedings of the 14th agile conference on geographic information science*.

Head, S. (2014). Worse than wal-mart: Amazon's sick brutality and secret history of ruthlessly intimidating workers. *Salon*.

Helles, R., & Jensen, K. (2013). Making data—big data and beyond: Introduction to the special issue. *First Monday*, *18*(10).

Herland, M., Khoshgoftaar, T. M., & Wald, R. (2014). A review of data mining using big data in health informatics. *Journal of Big Data*, *1*(1), 2. Nature Publishing Group.

Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. *science*, *332*(6025), 60–65. American Association for the Advancement of Science.

Hilbert, M., & López, P. (2012). How to measure the world's technological capacity to communicate, store, and compute information part i: Results and scope. *International Journal of Communication (19328036)*, *6*.

International Data Corporation. (n.d.). IDC's Global DataSphere Forecast Shows Continued Steady Growth in the Creation and Consumption of Data. https://www.idc.com/getdoc.jsp?containerId=prUS46286020.

Jiang, B. (2018). Spatial heterogeneity, scale, data character and sustainable transport in the big data era. *ISPRS International Journal of Geo-Information*, *7*(5), 167. MDPI AG.

Jiang, B., & Brandt, S. A. (2016). A fractal perspective on scale in geography. *ISPRS International Journal of Geo-Information*, *5*(6), 95. Multidisciplinary Digital Publishing Institute.

Jiang, Z., & Shekhar, S. (2017). *Spatial big data science: Classification techniques for earth observation imagery*. Springer.

Jin, X., Wah, B. W., Cheng, X., & Wang, Y. (2015). Significance and challenges of big data research. *Big Data Research*, *2*(2), 59–64. Elsevier.

Jung, V. (1995). Knowledge-based visualization design for geographic information systems. In *Proc. Of the 3rd acm int. Workshop on advances in geographic information systems (baltimore md* (pp. 101–108).

Kabakchieva, D., Stefanova, K., & others. (2015). Big data approach and dimensions for educational industry. *Economic Alternatives*, *4*, 47–59. University of National; World Economy, Sofia, Bulgaria.

Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. *Journal of Parallel and Distributed Computing*, *74*(7), 2561–2573. Elsevier.

Kayyali, B., Knott, D., & Van Kuiken, S. (2013). The big-data revolution in us health care: Accelerating value and innovation. *Mc Kinsey & Company*, 1–13.

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in human geography*, *3*(3), 262–267. Sage Publications Sage UK: London, England.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.

Kitchin, R. (2015). The opportunities, challenges and risks of big data for official statistics. *Statistical Journal of the IAOS*, *31*(3), 471–481. IOS Press.

Kitchin, R., & McArdle, G. (2016). What makes big data, big data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, *3*(1), 2053951716631130. SAGE Publications.

Klanten, R., Ehmann, S., Bourquin, N., & Tissot, T. (2010). *Data flow: Visualising information in graphic design*. Gestalten.

Kreye, A. (2015). Moore's law. In J. Brockman (Ed.), *This Idea Must Die: Scientific Theories That Are Blocking Progress (Edge Question Series)* (pp. 303–309). Harper Perennial: New York.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note*, *6*, 70.

Lee, J.-G., & Kang, M. (2015). Geospatial big data: Challenges and opportunities. *Big Data Research*, *2*(2), 74–81. Elsevier.

Leszczynski, A., & Crampton, J. (2016). Introduction: Spatial big data and everyday life. *Big Data & Society*, *3*(2), 2053951716661366. SAGE Publications Sage UK: London, England.

Lewis, S. C., & Westlund, O. (2015). Big data and journalism: Epistemology, expertise, economics, and ethics. *Digital Journalism*, *3*(3), 447–466. Taylor & Francis.

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., Pettit, C., et al. (2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, *115*, 119–133. Elsevier.

Lima, M. (2011). Visual complexity. Mapping patterns of information. Princeton: Princeton Architectural Press.

Lipton, Z. C., & Steinhardt, J. (2018). Troubling trends in machine learning scholarship. *arXiv preprint arXiv:1807.03341*.

Lupton, D. (2013). Swimming or drowning in the data ocean? Thoughts on the metaphors of big data. *Available online at https://simplysociology. wordpress. com/2012/10/29/swimming-or-drowning-in-the-data-ocean-thoughts-on-the-metaphors-of-big-data/ (last accessed December 29, 2016)*.

Lupton, D. (2015). The thirteen ps of big data. *Available online at https://simplysociology. wordpress. com/2015/05/11/the-thirteen-ps-of-big-data/ (last accessed December 29, 2016)*.

Manyika, J., Chui, M., Groves, P., Farrell, D., Van Kuiken, S., & Doshi, E. A. (2013). Open data: Unlocking innovation and performance with liquid information. *McKinsey Global Institute*, 21.

Marr, B. (2014). Big data: The 5 vs everyone must know. *LinkedIn. Available online at www. linkedin. com/pulse/20140306073407–64875646-bigdata-the-5-vs-everyone-must-know (last accessed December 29, 2016)*.

Marz, N., & Warren, J. (2012). *Big data: Principles and best practices of scalable realtime data systems*. MEAP Edition Manning Publications Co.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

McLaren, D., & Agyeman, J. (2015). *Sharing cities: A case for truly smart and sustainable cities*. MIT Press.

McNulty, E. (2014). Understanding big data: The seven v's. *Available online at dataconomy. com/2014/05/seven-vs-big-data/(last accessed December 29, 2016)*.

Moore, G. E. (2006). Cramming more components onto integrated circuits, reprinted from electronics, volume 38, number 8, april 19, 1965, pp. 114 ff. *IEEE Solid-State Circuits Society Newsletter*, *11*(3), 33–35. IEEE.

Morais, C. D. (2012). Where is the phrase "80% of data is geographic" from. *Available online at https://www.gislounge.com/80-percent-data-is-geographic/ (last accessed October 26, 2018)*.

Murthy, P., Bharadwaj, A., Subrahmanyam, P., Roy, A., & Rajan, S. (2014). Big data taxonomy. *Cloud Security Alliance (CSA), Tech. Rep*.

Network, C. A. (2018). Creative applications network. *Available online at http://www.creativeapplications.net/ (last accessed May 30, 2018)*.

Networking, C. V. (2018). Cisco global cloud index: Forecast and methodology, 2016-2021. White paper. *Cisco Public, San Jose*.

Nobre, G. C., & Tavares, E. (2017). Scientific literature analysis on big data and internet of things applications on circular economy: A bibliometric study. *Scientometrics*, *111*(1), 463–492. Springer.

Norvig, P. (2011). The unreasonable effectiveness of data - ubc distinguished lecture series. *Available online at*

https://www.youtube.com/watch?v=yvDCzhbjYWs (last accessed May 30, 2018).

Norvig, P. (2012). Warning signs in experimental design and interpretation. *Available online at https://norvig.com/experiment-design.html (last accessed May 30, 2020)*.

Nunberg, G. (2013). "The data are": How fetishism makes us stupid. *Available online at http://languagelog.ldc.upenn.edu/nll/?p=4396 (last accessed September 26, 2018)*.

Olshannikova, E., Ometov, A., Koucheryavy, Y., & Olsson, T. (2015). Visualizing big data with augmented and virtual reality: Challenges and research agenda. *Journal of Big Data*, *2*(1), 22. Nature Publishing Group.

Ovadia, S. (2013). The role of big data in the social sciences. *Behavioral & Social Sciences Librarian*, *32*(2), 130–134. Taylor & Francis.

Pääkkönen, P., & Pakkala, D. (2015). Reference architecture and classification of technologies, products and services for big data systems. *Big Data Research*, *2*(4), 166–186. Elsevier.

Peuquet, D. J. (1994). It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of american Geographers*, *84*(3), 441–461. Taylor & Francis.

Puschmann, C., & Burgess, J. (2014). Big data, big questions| metaphors of big data. *International Journal of Communication*, *8*, 20.

Rouse, M. (2018). Cloud computing. *Available online at https://searchcloudcomputing.techtarget.com/definition/cloud-computing (last accessed May 30, 2018)*.

Shekhar, S., Evans, M. R., Gunturi, V., Yang, K., & Cugler, D. C. (2014). Benchmarking spatial big data. In *Specifying big data benchmarks* (pp. 81–93). Springer.

Shekhar, S., Gunturi, V., Evans, M. R., & Yang, K. (2012). Spatial big-data challenges intersecting mobility and cloud computing. In *Proceedings of the eleventh acm international workshop on data engineering for wireless and mobile access* (pp. 1–6). ACM.

Shelton, T. (2017). Spatialities of data: Mapping social media "beyond the geotag". *GeoJournal*, *82*(4), 721–734. Springer.

Shin, D.-H., & Choi, M. J. (2015). Ecological views of big data: Perspectives and issues. *Telematics and Informatics*, *32*(2), 311–320. Elsevier.

Siegfried, T. (2013). Why big data is bad for science. *Science News*, *26*.

Silver, N. (2012). *The signal and the noise: Why so many predictions fail–but some don't*. Penguin.

statista.com. (2018). Data center storage capacity worldwide from 2016 to 2021, by segment (in exabytes). *Available online at https://www.statista.com/statistics/638593/worldwide-data-center-storage-capacity-cloud-vs-traditional/ (last accessed May 30, 2018)*.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*. year.

Storm, D. (2012). Big data makes things better. *Available online at insights. dice. com/2012/08/03/big-data-makes-things-better/ (last accessed December 29, 2016)*.

Suthaharan, S. (2014). Big data classification: Problems and challenges in network intrusion prediction with machine learning. *ACM SIGMETRICS Performance Evaluation Review*, *41*(4), 70–73. ACM.

Swan, M. (2015). Philosophy of big data: Expanding the human-data relation with big data science services. In *Big data computing service and applications (bigdataservice), 2015 ieee first international conference on* (pp. 468–477). IEEE.

Taleb, N. N. (2012). *Antifragile: Things that gain from disorder* (Vol. 3). Random House Incorporated.

Thakuriah, P. V., Tilahun, N. Y., & Zellner, M. (2017). Big data and urban informatics: Innovations and challenges to urban planning and knowledge discovery. In *Seeing cities through big data* (pp. 11–45). Springer.

Thatcher, J., Shears, A., & Eckert, J. (2018). *Thinking big data in geography: New regimes, new research*. U of Nebraska Press.

UNECE. (2013). UNECE - united nations economic commission for europe. *Available online at https://statswiki.unece.org/display/bigdata/Classification+of+Types+of+Big+Data (last accessed August 26, 2018)*.

Uprichard, E. (2013). Focus: Big data, little questions? *Discover Society*, (1). Social Research Publications.

Van Rijmenam, M. (2013). Why the 3v's are not sufficient to describe big data. *Available online at http://www. bigdata-startups. com/3vs-sufficient-describe-big-data (last accessed December 29, 2016)*.

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, *165*, 234–246. Elsevier.

West, G. (2013). Big data needs a big theory to go with it. *Scientific American*, *May*, *15*.

Widman, J. (2014). When new relic says "data helps," we're saying it right. *Available online at https://blog.newrelic.com/culture/data-is-vs-data-are/ (last accessed September 26, 2018)*.

Wilson, A., Thompson, T. L., Watson, C., Drew, V., & Doyle, S. (2017). Big data and learning analytics: Singular or plural? *First Monday*, *22*(4).

Yang, C., Raskin, R., Goodchild, M., & Gahegan, M. (2010). Geospatial cyberinfrastructure: Past, present and future. *Computers, Environment and Urban Systems*, *34*(4), 264–277. Elsevier.

Yao, X., & Li, G. (2018). Big spatial vector data management: A review. *Big Earth Data*, *2*(1), 108–129. Taylor & Francis.

Zee, E. van der, & Scholten, H. (2014). Spatial dimensions of big data: Application of geographical concepts and spatial technology to the internet of things. In *Big data and internet of things: A roadmap for smart environments* (pp. 137–168). Springer.