Table of Contents

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 ("big" standing for "fast" and "exhaustive")

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

[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.

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 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]

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

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 *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

[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)

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 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*

[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.

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 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:

[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.

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 lifecycle management, and big data security.[7]

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

[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.

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

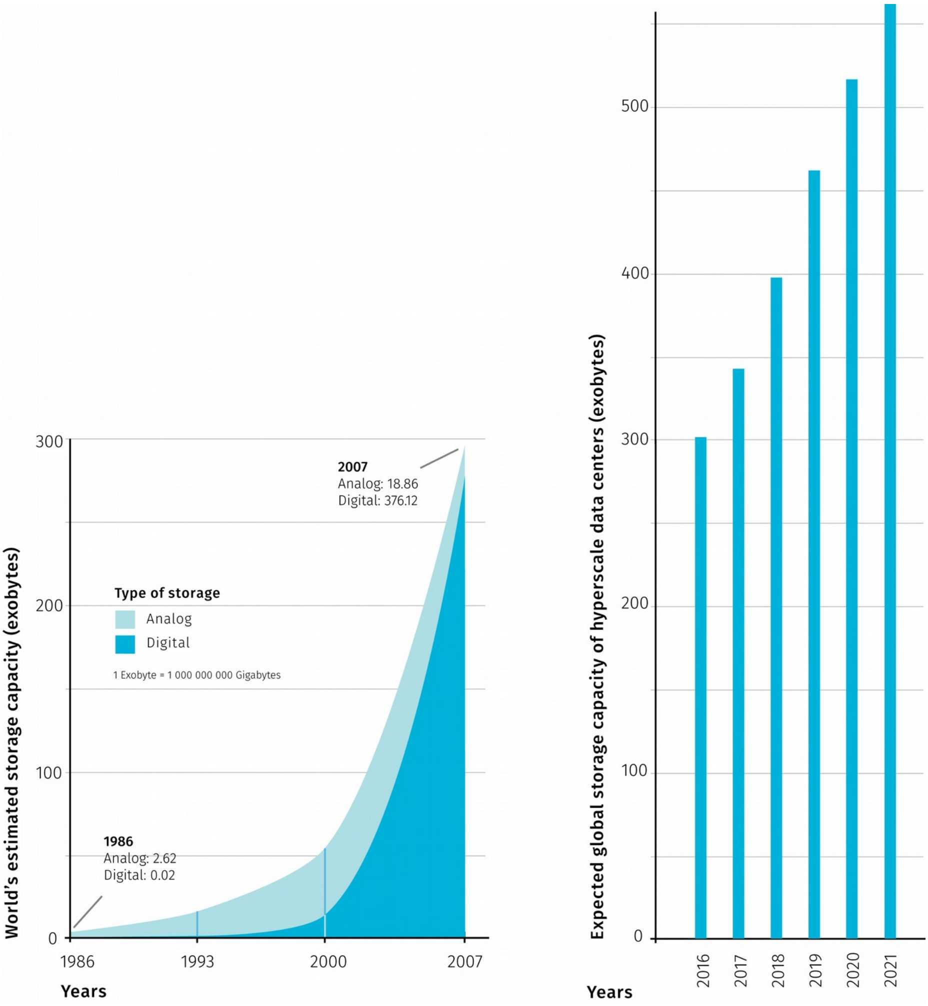


***Fig.1** Comparison of the World's estimated data storage capacity between years 1968 and 2007 (modified after Hilbert & López (2011)) and the expected storage capacity of large scale data centers in the period from 2016 to 2021 (modified after Networking (2018))*

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 potential 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

TODO – newer forcasts from 2020 (More than 59 zettabytes (ZB)), and how COVID-19 contributes to it https://www.idc.com/getdoc.jsp?containerId=prUS46286020

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

[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)

sources tend to be volatile and unstructured, therefore their representativeness is harder (if possible) to assess.

1.2.4 Metaphoric accounts

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 econimic input, used to create a new form of econimic 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 streaming
- *storage infrastructure* – sql, nosql or newsql
- *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 adjustability 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 approaches, that is definition by *keywords*, *constraints*, *examples*, *metaphors* or a *holistic* definition.

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

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

additional value, if only as a platform for joining otherwise un-joinable datasets. Much of data in the world is or can be geo-referenced, which indicates the importance of geospatial big data handling.

Cartography and geographic information science are the disciplines closest to the specialisation of this thesis, both having 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 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 rows 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.

[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.

- 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 (for more on efforts in automated generalisation see for example Burghardt, Duchêne, & Mackaness (2016)). The same phenomenon can be studied at various levels of detail even without changing the 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 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 **big data -> spatial big data**, where in the second case it is **spatial data -> 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 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 contents 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[13] 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.

To Li et al. (2016), main sources of spatial big data are in volunteered geographic information (VGI)[14] and in geo-sensor

[13] 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.

[14] VGI is defined as "the harnessing of tools to create, assemble, and disseminate geographic data provided voluntarily by individuals" (Goodchild, 2007). This well describes contributions to the Open Street

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[15].

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:

Map project, 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.

[15] 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*[16] – 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.

[16] 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).

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 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 possibly tie to them:

- Extensionality & Indexicality – spatial reference in itself is a unifying platform for combinining data from various sources, space is a natural concept to extend our understanding of a dataset 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 coocurrence in dense displays. Indexicality is a natural prerequisite for thematic mapping.

- Volume – from cartographic standpoint, the number of records is the most interesting measure of volume (compared to storage size or attribute lengt). Extensive volume does not need to present a problem for effective visualisation, especially if it plays out in the attribute space and the spatial reference is relatively static. Using right visualisation methods, map allows for information compression and clarification (TODO lepsie slovo).
- 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 cartograpy.
- 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 thoug a gap in incorporating unstructured data to digital mapping, for example in adjusting metadata profiles (e.g. move from linnean hierarchical classification to messier but more flexible methods like tagging), or in determining data quality from spatial context. Cartography is in position to search for ways to combine structured and unstructured data in meaningful way.
- Velocity & Exhaustivity – these parameters will be dealt with in chapers 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 depicted topic. Cartography is ideal for depicting timespace regularities and relationships within and between datasets. Exhaustivity then projects into longtime problem of graphic fill and tailoring cartographic visualisation to human congnitive 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 geoinformation science may be impacted by big data. The whole project of GIS might need to to be rethinked again, but this is not unprecedented. rom 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 GIS might come as a response to big data.

1.4 Assessing impacts, threats and opportunities rather than seeking definitions

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 new technologies that became part of the notorios hype cycle.[17] 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. But we should not forget that even though the technological industry is largely fashion-driven, its societal impacts are real, though maybe unevenly distributed.

It is beyond the scope of this thesis to consult all 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 discussions on consequences in *scientific inference* and *knowledge-based decision making*, the areas cartography always aimed to support. Closely related is the issues of *surveilience* trough big data and *emerging digital divides*.

Swan (2015)

a) correlation vs causation or more broadly the need of theory vs purely data driven inference
b) bias-free interpretation of big data
c) how should big data abuses be addressed and emerging digital divides

[17] 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 cycle differs per technology, and some technologies may never reach productivity in the foreseeable future. Hype cycles are a construction of the Gartner consultancy that issues regular reports, see for example Gartner (2018b)

d) The scientific reflection on big data revolves mainly around the question if automating reasearch changes the definition of knowledge. The anticipated mindset changes voiced mayer2013big can be summarized into the following points:

- Reduced need for sampling with accessibility of n=all datasets
- Loosened requirements for exacticude as minimizing sampling errors would leave room for more relaxed standard for measurement error (i.e. will to sacrifice a bit of accuracy in return for knowing the general trend faster)
- Move away from the search of causality: “big data is about *what* not *why*.” Multi factor correlation with large data enables decision making even without understanding the mechanisms behind the relationsip. 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 but to obtain some instuction to base action, correlation might be enough to provide some backing. For the optimistic commentators, this abandoning of theoretizing can open door to iterative experimentation and building of useful heuristics that are independet of preconceptions and biases of our thought processes. To others, this shounds scary at best, as such naive data appreciation can dangerously rationalize incopetent 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 corelations. So it is more honest to claim that data to *support* decision making, and there always is a decision maker. Data-driven rethorics is suspicious as it allows decision makers to evade responsibility or to ignore alternative decisions. Furntermore, in decision making under opacity, over-reliance to historical records can catch us ill-prepared for unprecedented large scale events. 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 powerfull exploratory potential of big data to opportunistically exploit new avenues as they appear sounds promising, though there is no need to discard the hypotheses whatsoever as these can be generated and 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, 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 (episthemology). Any mode of interacting with big data is representation and not necessarily reality, and the reality gap may be so big that data howerver 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 critcal 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 to the right direction. Big data can then be aimed to support an optimistic goal, for example in view of West (2013): *overreaching predictive mathematical frameworks for complex systems*. Big global issues in ecology, pandemics or financial markets tend to show signs of a complex system[18]. “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)). One final quote on optimistic agenda: “[…] 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)).

The afrmentioned discussions point to lock-step evolution of science and technology, and most importanlty, to strong reflection and self-correcting mechanisms inherent to science that usually set in motion when innovation is accompanied with some troubling signals[19]. In broader society we also need such a reflection of new realities created by big data and the accompanying ethical issues.

First set of issues revolves around collecting data when users have no choice to opt out and do not give explicit or informed

[18] 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)

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

consent. Even with informed consent, there is no visibility into the secondary uses that the collected date will cater to, to what additional sources it will be combine with, what analytical engines will be applied, and what third parties will the data be resold to. At the time of writing, the legislation to address these issues is catching up[20], 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.

Anonnymization may no longer work as combining digital traces from several sources allows for re-identification of an individual. Anoter topic is the ability of user to access the data collected about him, either to use it for his own self-analysis, or to issue its removal (tough how to verify this has actually happened?). In an alternative vision of big data economics, individuals may gain power to sell their data themselves of through intermediaries and sell them.

Penalties based on propensities – that is a short description of a concern that with increased surveilience 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 the Minority report) (Mayer-Schönberger & Cukier (2013)). It is a fact that the technical infrastructure for close personal

[20] 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.

scrutiny and behaviour enforcing has been already implemented at the scale of a plant (Head (2014)) as well as country (most (in)famously in China). At the time of this writing, the global pandemics of COVID-19 provided consent for public scrutiny at unprecedented levels, and it is yet to see for us how things will evolve afterwards.

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 malicions 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."

Looking at the peak BD excitement around 2014 from the perspective, the hype now thransfered to deep learning that gets inflated nowdays. Leaving the shiniest spotlight does not mean the end of existence, the item is just deflated to its just place in the state of things. BD and ML are can not really function without each other, processing algorithm quality does not withstand poor data quality and same goes for visual presetnation.

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". On the other hand, the number of researchers and practitioners willing to invest their time in big data related endeavours is relatively high[21], which sheds some positive light on the future vitality of the discipline.

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

[21] *Jounal 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.

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 buzz-word, the fact is that more digital information gets created and can be linked more easily, which has many 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 some economic, societal and scientific impacts of big data, as they can provide a motivation for cartography to take part in addressing the related issues. We will also offer some speculation on how the roles of cartography and GIS may be transformed by the data deluge.

Todo incorporate this to the conclusion:Press (2014)

some definitions:

(11) The belief that the more data you have the more insights and answers will rise automatically from the pool of ones and zeros.

(12) A new attitude by businesses, non-profits, government agencies, and individuals that combining data from multiple sources could lead to better decisions.

I like the last two. #11 is a warning against blindly collecting more data for the sake of collecting more data (see NSA). #12 is an acknowledgment that storing data in “data silos” has been the key obstacle to getting the data to work for us, to improve our work and lives. It’s all about attitude, not technologies or quantities.

2 Making sense of spatial big data

> *Technology is the answer, but what was the question?*

> Cedric Price

This chapter looks more closely on the properties of data *with point spatial reference that count for the majority of spatial big data. Then we will outline the tendencies in spatio-temporal knowledge discovery, and we will discuss general ways how cartography can support understanding the world trough the lens of big data. We will also discuss some objections to the idea of insight generation (or rather of certain naive ways in which data is interpreted) and speculate on how cartographic practice could overcome such risks.*

2.1 Spatial big data classification: stations, events, and agents

The vast majority of what is presently understood as spatial big data has point spatial reference. This prevalence comes naturally if we realize that the “data point” location is described basically as a coordinate pair – two digits that can be easily stored in standard database systems without the need to observe topological rules and other constraints that GIS vector data model enforces on line and polygon geometries. Point data are spatial data that are easily created and handled by non-spatial (meaning not GIS-enabled) systems that account for majority of data production. For this reason, and due to the scope limits of this thesis, we will almost exclusively focus on visualisation issues related to point data[22].

Point spatial data are not a homogeneous group. We can describe three classes representing three kinds o objects differentiated by their behaviour in space and time, more precisely by the by how dynamic their *existence*, *location*

[22] We’ll use the term *point data* as a shorthand for “data with point spatial reference”.

and *attributes* are over the course of observation. These properties are determined largely by the source of data, so for convenience we can nickname the three types as *stations*, *agents* and *events*:

- stationary objects (*stations*) have static position and existence, meaning that they don't move or disappear during observation. What is dynamic is the set of attributes attached to the object – in big data world these attributes can come as a continuously updated streams. Basic examples include weather stations, traffic cameras, and any kind of stationary sensors. - moving objects (*agents*) move around, so their position changes during observation, also their existence can be dynamic, meaning they can enter or exit the area of interest. Various kinds of dynamic attributes can be attached. We can reconstruct the history of movement of these objects, which invites conversion to linear representation. Examples are vehicles or pedestrians carrying GPS devices and sensors. - episodic objects (*events*) have existence limited to a specific point in space and time. As they are short-lived, we can say that position and associated attributes are static. Prime example are data collected from social networks.

Think of this distinction as a convenience model fit for the majority of big data related use cases that expect short time frame for data utilization. Technically speaking, the difference between stations and events is dependent on the frame of reference, as objects seen as stationary in shorter observation periods can become mere events if the observation time frame is significantly extended. The existence of a building usually spans over a long time period, though if we stretch the perspective to a century or a millennium, most buildings will become mere glimpses existing a tiny fraction of time[23]. Geographers would note that also the location of seemingly static environmental features doesn't hold over time (think of a meandering riverbed or a volcanic landscape). So again, longer time frame changes our assumptions of static location.

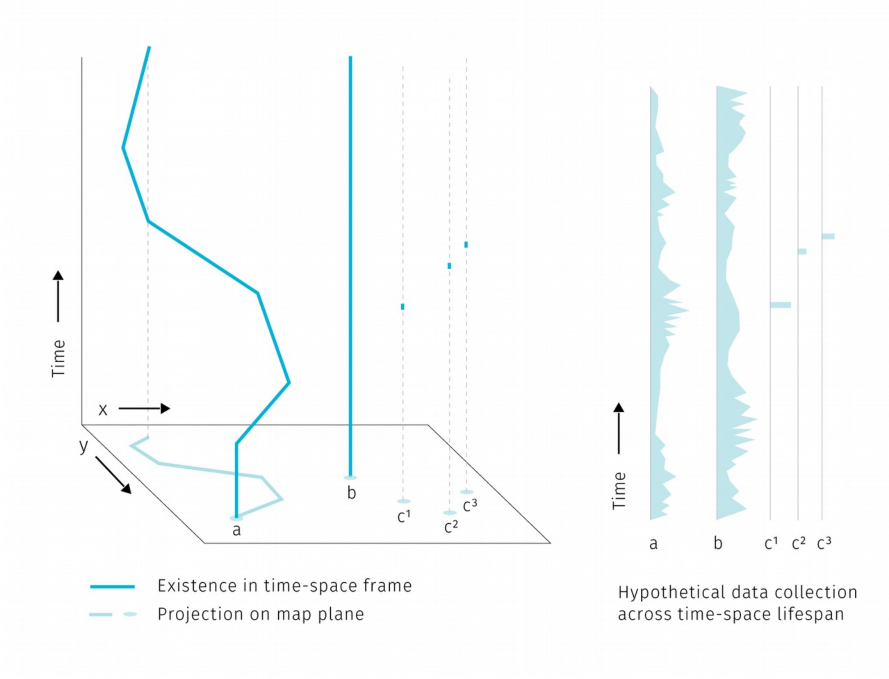
Furthermore, the spatial extent of the observed area and hence the scale of the map influences the distinction between moving and stationary objects – if the movement is too limited to be recognized at a given scale, we can model it as a stationary object. Also, some events can be reimagined as moving objects with discrete presence across observation time frame, for example if social media events dislocated in space and time are traced back to a single moving source device[24]. But let us not problematize any further, with most big data sources being temporally and spatially limited (to near real time and mostly urban environment)[25], the distinction to stations, agents and events would suffice. Judging by the real data samples we can say that stations are usually physically present in the environment while events are mainly records of something that happened at given location, either physically

[23] see https://waitbutwhy.com/2013/08/putting-time-in-perspective.html for an excelent visualisation of perspectives changing with time frame

[24] See examples of such practice here https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html

[25] Again, as an exception from the rule, there are some impressive global-scale data projects: global fishing watch..., tracking ships...

expressed in the environment and observable by onlookers ("I was at a restaurant") or not ("I was shopping online while waiting at the bus stop").


Time
x
y
a
b
c1
c2
c3
Existence in time-space frame
Projection on map plane
Time
a
b
c1
c2
c3
Hypothetical data collection
across time-space lifespan

Tab1 Properties of point spatial objects

type of object	existence (records of spatial and temporal reference)	attribute collection	location
station	continuous	continuous or discrete	static
agent	continuous or discrete (can reappear)	continuous or discrete	dynamic
event	discrete	discrete	static

In the above image and a table we assume that the attribute collection is happening continuously for stations and agents. This does not mean that the attributes have to be collected continuously at all times. Some sensors can record at a regular time interval or only in case of an event. The data output can then contain several “no data” records or even no records at all if the event did not happen. It then depends on the goal of the analysis how such data are conceptualized. For example a traffic camera is a stationary object but some part of its data collection is episodic – a photo is taken just when a speeding vehicle drives by. The classification introduced above differentiates between the existence of an object and the act of recording data by the object. We assume that the sensor’s presence without recording has also some analytical potential as it proves the absence of event, while with no sensor in place we cannot say if the event did take place or not.

Compared to stations and agents, events with episodic presence seem to be the least data-rich, but their analytic potential grows when lot of them is accumulated. Clusters of georeferenced point events, a.k.a. point clouds are at the core of spatial analysis based on data from mobile devices.

The gaps in data collection and the absence of abrupt changes hints how to optimize data storage from such sources. Even though storage optimization techniques are not within the scope of this thesis, they can pose a certain lesson for cartographic visual analysis. For cartographers, the utilized resource is a space within the map plane, that can only hold a certain amount of graphic elements to remain useful. The graphic fill reduction (or better optimizaiton) is an aspect that can enhance the knowledge discovery at the end of the visualisation pipeline (more about it in the practical part of thesis on aggregation).

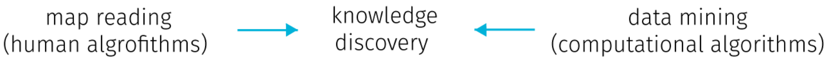
TODO – revisit when case studies are done and connect it better with the actual data used

2.2 Spatio-temporal knowledge discovery and visual analytics

In this section we will briefly discuss techniques for exploring spatio-temporal data, with emphasis on practices that would benefit from enhanced cartographic visualisation.

People engaged in data-related practices are motivated by an expectation that their work can help to provide some insight into how the world works, that there is some knowledge that can be unlocked, mined, or distilled from otherwise untelling piles of data. Such insight seeking is the crux of the concepts such as *data mining*, *spatio-temporal knowledge discovery* and *visual analytycs* that we will explore furhter.

Data mining is exploring databases using low-level algorithms to find patterns. *Knowledge discovery* is then a higher-level extension of data-mining techniques that requires human-level intelligence and domain knowledge to guide the process and interpret the results (Miller (2015)). In the knowledge discovery process, computation is seen as an extension of human force rather than its replacement, therefore the goal is to marry the best of both worlds. This is reconciled with the (current) capabilities of information technologies: there are tasks that are very simple for computers and very hard for humans (e.g. calculate the square root of 567789898) and vice-versa (basically any task requiring improvisation). *Visual analytics*, the science of analytical reasoning supported by interactive visual interfaces (Thomas & Cook (2005)), then zooms in at the interaction frontier between human and computer in order to find the best tools for visual interaction between the two. If we imagine a continuum ranging from “work done purely in human brain” towards “work done by machines”, knowledge discovery places itsefl somewhere in the middle.



We can draw the humman-machine continuum in the field of digital cartography as well. Here, the human cognitive abilities are applied to seek patterns, explore spatial context or make decisions, while computational aspects include data management and processing. Now, the computation heavy algoritms like optimal route calculation already step in to unburden humans from some decision-making so the destinction shouldn’t be taken as something rigid (TODO see the disucssion at the end of the chapter on how the continuum might evolve and how it could affect cartography). For now, just note that cartography operates as an interface at the human side. Some authors go on to define *visual analytics for spatio-temporal data* as interlinked techniques in interfaces with map as a central metaphor (Guo, Chen, MacEachren, & Liao (2006)). We can think of it as map reading with robot assistants.

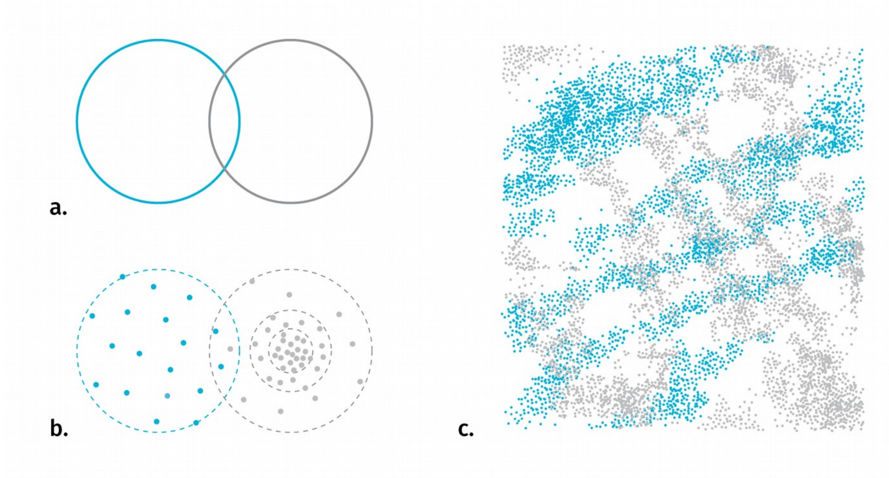
2.2.1 Spatio-temporal relations

To develop further on the kinds of interaction with spatial data, we can explore the concept of *spatial* and *temporal* queries. On the general level we can search for spatial and temporal relations in all theree types of point objects mentioned in the first section. In addintion, moving agents can genrate specific relations not innate to stations and events.

Spatial relations are at the very basis of map reading for orientation clues, but are also vital for interpreting thematic information. We percieve these relations between the dominant themes (e.g. in weather maps of precipitation and atmospheric pressure zones) or between the theme and the topographical base map. The major classes of spatial realtions are: *set-oriented* (union, difference, intersecton, complement, etc.), *topological* (connectivity, interior, exterior, boundary), *directional* (cardinal, object-centered, ego-centered directons) and *metric* (e.g. Euclidean or network-based distance) (Worboys & Duckham (2004)).

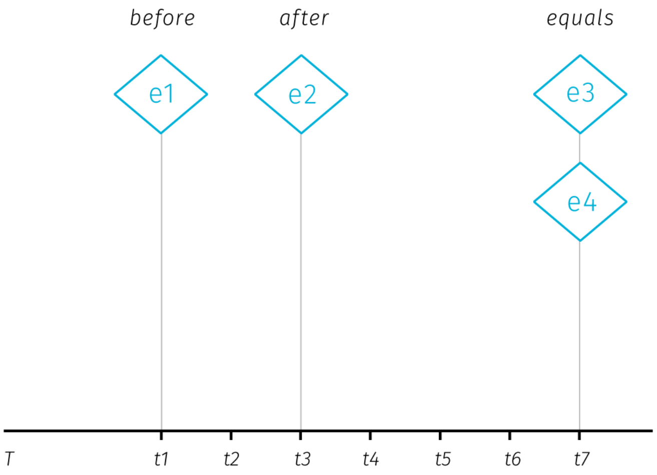
Point spatial data of large extent complicate observng such relations. We rarely ask about a single specific point from the set, more often we seek to extract some tendency of the whole point cloud. The nature of some data sources can dictate some spatial relationships (such as vehicles being spatially bound the road network), but in many cases the density of point cloud obscures the base map and precludes reading of attribute variablity within the set.

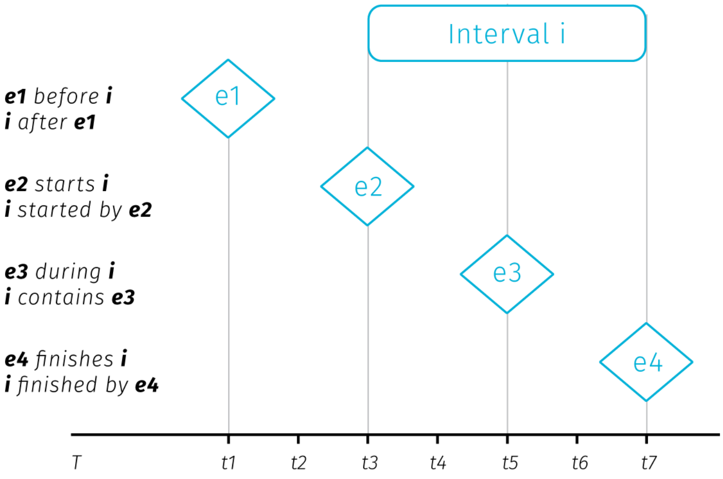
In this thesis we are mosty considering point data clusters in two dimensional space, so it is worth to say that spatial relations between such sets are harder to conceptualize than it is with polygonal features. Egenhofer & Franzosa (1991) describe 16 actual relations (9 if reduced to spatial regions relevant in GIS) in two dimensional space. Hovewer, in their approach Egenhofer & Franzosa (1991) define the point sets by their exterior boundry and then effectively treat them as polygons. But delineating the exterior boundry is a challenge in itself, for example when dealing with smooth transitions in point density at the border, or with outliers. Any line would be in a sense an inaccurate approximation. Spatial relations between point clouds in three dimensions are a subject of extensive research in the fields of computer vision and indoor navigation (e.g tran2017extracting or chen2019deep). However, the motivation here is object identificaton. In these lines of research the point cloud is hiding distinct solid objects in the real space that need to be extractd, so the point cloud itself is not an object of research. For cartography, the point sets already come with some assigned attributes, so there is usually no need to label them algorithmically (but maybe there could be some inspiration). Large point sets tend get unruly in the wild, and saying anything meaninful about spatiotemporal relations of multiple such clouds is increasingly hard using the basic set theory (see Fig).

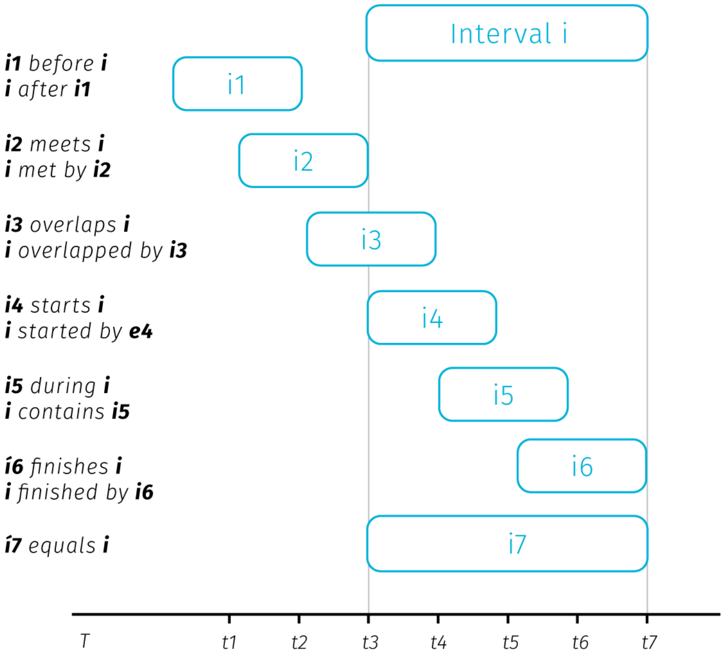


TODO maybe some cluster shape measures from geostatistics?.

Temporal relations are measures of coincidence. There are thirteen possible relations between two temporal records described in Allen (1984). As we have seen with stations, agents and events, the existence and data colection of any entity can be either continous or discrete in time, it is therefore useful to distinguish between *time point* and *time interval* when investigating temporal relations (see figures). Linear conceptialization of time can be supported with cyclical and branching time, there can be discrepances between the temporality of base map and the thematic overlay, or between the time interval of existence and representations. We'll untangle these complexities in chapter 5.


before
after
equals
e1
e2
e3
e4
T
t1
t2
t3
t4
t5
t6
t7


Interval i
e1 *before* ***i***
i *after* ***e1***
e2 *starts* ***i***
i *started by* ***e2***
e3 *during* ***i***
i *contains* ***e3***
e4 *finishes* ***i***
i *finished by* ***e4***
e1
e2
e3
e4
T
t1
t2
t3
t4
t5
t6
t7



Relations specific to moving objects – moving objects have a specific set of properties comming from the combination of their spatiotemporal circumstances. These can be *instantenious* (actual position and speed), *interval-based* (e.g. travel distance from departure), *episodic* (related to extenal event) or *total* (related to entire trajectory). (Laube, Dennis, Forer, & Walker (2007), andrienko2008basic).

2.2.2 From data mining to visualisation for human interpretaton

Havig described the fundamental spatio-temporal relations in big data sets, we can briefly describe some of the methods to uncover them. Recalling the human-machine continuum at Fig., we will start at the machine side with methods from the data mining group to eventually move towards the human side.

Several data mining concepts are of interest. *Association rule mining* is searning in databases for conditions ocurring together frequently:

x => *y (s%,c%)*

Where *x*,*y* are conditions, together forming an *itemset* and *s*,*c* are levels of support and confidence. Suport and confidence are basic rule performance measures, support being the measure of how often the itemset occurs in the whole database and confidence being the proportion of x being a memeber of an itemset x => y. For example: *park* => *school (4%, 55%)* means that 55 percent of parks are near schools, for 4% of items in the database (Han, Pei, & Kamber (2011)). The measures of support and cofidence allow us to set tresholds for significantly frequent co-ocurrence.

Spatio-temporal association rules extend associtation rules to describe how objects move among a set of regions over time (Verhein & Chawla (2008)). Inocorporation of spatiality into assotiation rules takes form of a simple binary condition telling if the items coocurred in the same predefined sets of regions or not.

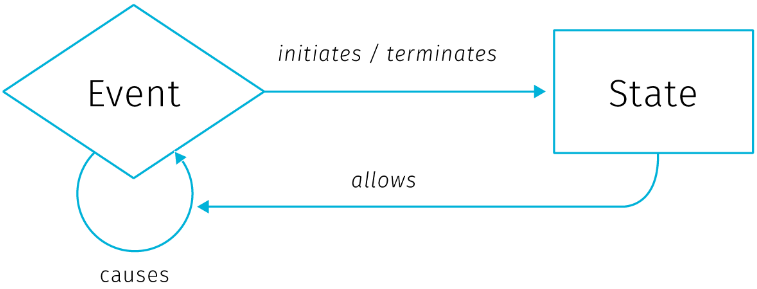
Sequence mining is seraching for patterns in time and other sequences. Similarly to association rules, we search for events occuring frequently together by considering three parameters: the *duration* of the whole sequence, the *event window* (time-horizon for considering events as temporally coincident) and the *time interval* between events (Miller (2015)). These parameters allow us to turn the temporal relations between two items into binary parameter telling if the items co-ocurred (that is when the time interval fits into the the event window).

Periodic pattern mining is a type of sequence mining that searches for recurrent patterns in time sequences. Such patterns can be: *full periodic patterns*, *partial periodic patterns* (e.g. just on mondays), and *cyclic or periodic association rules* that associate events that occur periodically together (Han et al. (2011)).

Considering the breadth of posible spatial and temporal relations described earlier, the conceptualization of spatial and temporal coocurrence in the association rules may seem rather simplistic. Basically it is reduced to a yes/no parameter. Moreover, moving from the level of individual database entries towards assessing relations between compound entities such as spatial point clusters seems to be out of the scope of these methods. Of course, the way spatiality is inscribed into association rules could be made more sophisticated, though with inevitable implicatons for mining performance. With large datasets, mining even the simple rules forces us to consider performance. (TODO maybe develop to say here simple visual comparison is cheaper, at least for clusters).

At this point we can step back from mining algorithms to invite some human interpretation and to consider what conclusions we can actually draw from spatial and temporal co-ocurrence of events. The usual assumption is that such coocurrence can point to some form of causality.

Drawing from approaches by Allen, Edwards, & Bédard (1995) and Galton (2012); Bleisch, Duckham, Galton, Laube, & Lyon (2014) distinguish between the trigger that apparently causes the event and the environmental conditions that have to be fulfilled for the effect to occur.

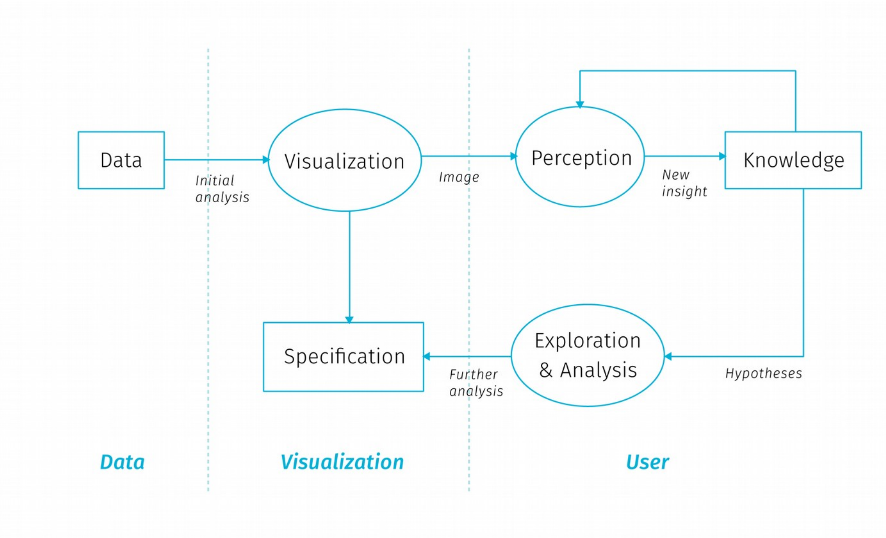


In this model, *state* is an environmental condition and *event* is a change of state. Events are caused only by other events, while states only affect causation by allowing events to cause other events. Events *initiate* and *treminate* states, while states *allow* causation. The *initiate*, *terminate* and *allow* relationships are then dubbed *causal-like* to distinguish them from the event-to-event causation. TODO example

In conceptual framework for finding *candidate* causal relationships in movement patterns Bleisch et al. (2014) distinguish between three kinds of granularity at which the phenomena can be described: *spatial*, *temporal*, and *causal*. While the first two are defined by the the smallest spatial and temporal units, causal granularity is given by the kinds of events observed. Spatial and tempral granularities can be easily reduced to “see the bigger picture” (by changing the spatial scale, or extending the time range of observation), but causal granularity is more firmly defined by the data collection design.

El-Geresy, Abdelmot, & Jones (2002) note that alghough the general expectation would be that the effect occurs immediately after the cause, some delay between the effect and the cause can occur, possibly because the cause must attain some intensity treshold to trigger the event or because the effect and cause are spatially separated and it takes time until the influence of the cause reaches the location where it takes effect. Bleisch et al. (2014) suggest that these apparent delays result from lower causal granularity of observation, i.e. there is some intermediary chain of effect and cause that happens during the delay but it is not recorded by the observation. (TODO: Example, fish, signall, make up something). Wheter we accept the effect delays as real or illusionary might be more of an academic question, tracing down the potential causal link between start and end events can yield predictive potential even when the intermediary causal chain remains undiscovered.

Discussing the interpretation of spatiotemporal co-ocurence we have moved on the human-machine continuum towards the human end. At this point, visualisation becomes important as an interface between the user and the data. One of the general models describing how knowledge discovery proceeds via inference and interaction is the sense-making loop (fig).



Visual analytics extends the concept of visualisaton: not only it provides a visual interface to the database, but also makes the data processing piplines transparent for an analytic discourse. Keim2008visual in their introductory paper say the goal of visual analytics is the creation of tools and techniques to enable people to:

– Synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data – Detect the expected and discover the unexpected – Provide timely, defensible, and understandable assessments – Communicate assessment effectively for action

This is turly a long way from the low-level search for coocurrences, though it is not clear how should these grand goals materialize in practice. Keim2008visual call for broad inter-disciplinary collaboration between related fields (Visualisation, Data Management, Data Analysis, Perception and Coginition, Human-Computer interaction) and idetnify a range of application and technical challenges.

The breif tour we just went trough lets us appreciate the prospect of gaining the best of both worlds, that is to support human analytical efforts with algorithmic power doing the heavy lifting around data manipulation. We have seen that inscribing spatiality a and temporality to data mining processes can be both cumbersome and simplistic. Furthermore, the coocurence we want to search for needs to be defined beforehand, so in many cases the data mining and transformation is insufficient to provide the required insight. Search algorithms can be performace heavy so some coorination with human observer that is able to easily gain an overview of clusters beyond individual database entities. Visualization and visual analytics provide this exploratory potential, especially for big data in situation where we don't yet know what questions we want to ask. Visualisation as a sense-making tool gives us a way to find things that we had no theory about and no statistical models to identify and to explore the space of models in more expansive ways (Bollier & Firestone (2010)). Many possible data transformations may be applicable to a particular problem, but it is not necessarily clear which ones will be of most value in facilitating insight. Also, because visual analytics is qualitative as well as quantitative, there are no assumptions of exact parameters and well-defined boundaries between what is interesting and what is not. A priori criteria of significance may be manipulated based on the judgment of the analyst (Thomas & Cook (2005)).

Back to the machine to make addtional queries and transformations to support a hypotesis or model outcomes.

The hybrid system is more powerful than either the machine or the analyst working alone.

Digital cartography has the potential to dynamically support cognitive tasks in described fashion, be it with simple visual higlihting, or with algorithmic assistance yet to be designed.

2.3 The role of cartography

Cartography has a long tradition of making data comprehensible to our visual minds. Beautiful and authoritative maps in school atlases explaining the formation of air masses or the positions of ocean streams give off and impression of definitiveness but were build upon a generalization of data from loads of observatios. These data had to be collected, brushed and analyzed for the presence of meaningful patterns, and than visualised in a way that would appeal to human comprehension. The process for creating such maps was nowhere near “real-time” but allowed for fine tuning of all aspects of a map: from carefully shading the outlines of water bodies to making the street connections visually pleasing. Map making allowed for perfectionism, and the resulting maps remain beloved by collectors long after their ‘utilitarian’ function is gone.

For digital cartography^(Here and further we will use the term digital cartography as a shorthand refering to dynamic maps allowing user interacation, consumed almost exclusively throuhgh the web, viewed on screens of various sizes) it took a long time to come any closer to the visual quality of the best works in cartography in print. Arguably, there is still some unfulfilled potential in getting towards graphic excellence in web mapping, though recent improvements in available tools open possibilities for improvment, but also risks of uniformity. Digital maps have the obvious advatage of allowing interaction – user can zoom, pan, change, filter and combine the displayed data. The second big advatage is the possibility to update the displayed data real-time as the data source is updated. Sure, many digital maps are not dynamically updated, simply because the topic does not require it (e.g. medieval monasteries in France or 1991 election results in Yugoslavia). But interactive maps based on dynamically updated data are interesting as they pose a whole new set of challenges on authors. Ensuring cartographic quality in the map field now means desitgning for yet unseen changes in data also with user-induced modifications in mind.

School atlases served for presentation of knowledge, were confirmatory. Digital cartography allowed for exploratory mode of map reading to emerge, or more precisely, moved the exploration part down the process pipeline from *before* to *after* map publication, and from the cartographer/author to the map user. Visual analytics based on spatial data provide interfaces to manipulate and visualize data, or better to say to pick from the predesigned visualisation modes. This has implications for both the cartographer and the user.

The ability to interact with the map view can surely be empowering for the user, being passing on the sensation of exploration. On the other hand, things can go the wrong as it is very hard to create an imersive experience that would be immediately understandable to the newcommer. Exploratory map applications indtend for general public can leave users overwhelmed with the amount of possible interaction points. Left to her own devices, without any stated framework for interpretation users need to create her own narration about what is displayed. Visual interfaces are prone to be terrifyingly cluttered, untorubled with the dangers of fostering misinterpration. Lack of guidenance on where to start results in poor engagement with the application, that is quickly aboandoned. With applications for specialized audience, this can be mitigated by learning as users are forced to work with the applicaton as part of their job. Simmilar problems occur in business analytic dashboards proliferating in enterprises, which fail to make sense to users, or worse, fake insight with vaguely understood and hardly interpretable metrics. All these caveats pose a big responsibility on application designers.

(TODO mostík)

# Map reading and interpretation

Interactive map as a data manipulation interface is useful for those who know what questions they want to ask, but also for those who want to find out what they might be asking. So what kind of inference should an interactive map support? We should start simple, with basic quantitative questions. A big advantage of interactive maps over print is that we can display the exact quantities on demand (e.g. with some pop-up window bound to cursor hover action) and not rely on the viewer's ability to infer quantities form the legend (especially if categorized to some interval scale). The ability to answer simple quantitative queries shouldn't be left in vain, because as Tufte, McKay, Christian, & Matey (1998) warns: "when scietifiec images become dequantified, the language of analysis may drift toward credulous descriptions of form, pattern and configuration […] rather than answer to questions *How many? How often? Where? How much? At what rate?*".

We can say that these queations are at the basic level of map reading. bertin1983semiology distinguishes three reading levels for the thematic map, and at each level, different sorts of questions can be asked: - *elementary level* – questions introduced by a single element of the visualisatin (What is the level of unemployment in this district?) - *intermediate level* – questions introduced by a group of elements or categories in the visualisation (What are the five most populous disticts in the region?) - *overall or global level* – questions introduced by the whole visualisation (What are the spatiotemporal trends of traffic in this city?)

It is obvious that even a simple map has a potential to introduce countless possible combinations of questions at various levels. Although the importance of allowing for elementary-level questions, in thematic cartography we are often interested mainly in the global level of reading. Often times, just to *see* the overall level is a relevation, the kind of overreaching macroscope perspective unique to maps. But what else we can do with the overall patterns?

As we'll see in the next chapter, showing the basic quantities with cartographic means becomes more challenging with multiparametric visualisation, especially if we want to support both elementary and global levels of reading for individual topics.

Are there any examples of cartographic visualisation succesfully supporting the analytical reasoning? Maybe the most frequent answer of this question would be the celebrated map of the cholera outbreak in London 1855 by John Snow, that helped to identify the source of the epidemy in a polluted water pump. This feat is lauded for launching spatial eipdemiology and for bringing the thematic cartography to the fore (Clarke & Pickles (2015)). But what exactly made the Snow's method worth following? Tufte et al. (1998) notes four key points:

1. Placing data in appropriate context for assesing cause and effect 2. Making quantitative comparisons 3. Considering alternative explanations and contrary cases 4. Assesment of possible errors in the numbers reported in graphics

TODO: elaborate ^ We can make this a moto also for current exploratory maps.

Hypothesis formation support that these tools aim to provide is however a property that is hard to measure: it is hard to prove the interface works, it is hard to compare solution to establish which interface is better.

^ TODO brush up the line of thought, more about the overal patterns (from paper Illuminating the path…)

TODO - other visualisation-related issues: - don't know what questions I want to ask - looking for ouliers

TODO – proti interpretácii (niekam zahrnúť??) – https://www.kinecko.com/proti-interpretacii/?fbclid=IwAR2Ms83cZyRofIbDOqhucPID9Qt9aOQyRKFExdxcb9_zhcKndrgPYNN53LQ (Susan Sontag) Sontag & others (1994) – niečo v tom zmysle aj o pozorovaní reprezentácie (mapy) spôsobom akým pozorujeme prírodu – bez automatických hľadaní príčinných súvislostí medzi javmi. (možno odkaz na Kahneman kapitola o statistical thinking? – Jak tu zaváži kartografia?) Kahneman (2011)

TODO – building engagement – uloha kartografa – najrv vzbudit zaujem – nejaka inspiraci av good practices dizajnových studii orientovaných na komunikáciu. Jak komunikovať jasne a pritom nie prvoplánovo (na druhej strane objem inoformácii v marketingu je neporovnatelne nizsi a cielom je skor vyvolat emocie – ale netreba tieto postupy podcenovat) – k engagementu diagramy z algorithms in art TODO – procedurálne diadramy z designu (dual diamond diagram) – to je vlastne zoom in na návrhovú zložku v kolaboracných diagramoch nižsie (zdôrazniť cyklickosť procesu). – iné príklady – user roles, user stories. Nezačínať od dát ale od používateľov – protiargument: weaponization of design

Problems with pattern interpretation around big data.

# BD discussions from bollier2010promise (Bollier & Firestone (2010) – For Joi Ito, the Chief Executive Officer of Creative Commons, the search for correlations is a trap to be avoided, at least in his capacity of a computer security expert and a venture capitalist. Ito says he is “always looking for unpredictable things that you can use

opportunistically.” As a venture capitalist, he is looking for the “subversive outlier” whose ideas could have a big upside. From a security perspective, Ito says he wants to be alert to the unexpected forms of intrusion and deceit, not to the ones whose correlations can be easily discovered using computers When you do that kind of analysis on, say, terrorist networks, you have to understand that Hezbollah is actively trying to continuously come up with patterns that they think you won’t predict.” “Remember,” said Ito, “the same technology that we’re using to analyze Big Data enables these other actors to become more actively random. The people who are outliers, who used to sort of behave randomly, now have access to the same tools as the rest of us and are looking at the same data.

“Big Data is about exactly right now , with no historical context that is predictive,” said Ito. “It’s predictive of a linear thing—but you can use data collection to discover non-linearity as well. … It’s important not to be obsessed with the old models that come from the old data. It’s more important to be ignorant enough to come up with a new model of the future.” (Bollier & Firestone (2010) – Many innovative uses of Big Data could be called “now-casting,” said Varian. This term refers to the use of real-time data to describe contemporaneous activities before official data sources are available. “We’ve got a real-time variable, Google search queries, which are pretty much continuous,” said Varian. “Even if all you’ve got is a contemporaneous correlation, you’ve still got a six-week lead on the reported values” for certain types of data.

– “To make money, you’ve got to predict two things—what’s

going to happen and what people think is going to happen.
```

Talking about the human interpretation we can surely adress a wide range of use cases and motivations regarding to data. User roles (TODO move to cartographic part?), weaponization, uncovering secrets… entepreneurial approach. Large scale optimizations, smart city concepts etc. Let us not bloat here.

Paying attention to the cognitive part of information processing (differences between users, influence of learining from the app,..) – Kim Taipale of the Center for Advanced Studies in Science and Technology warned that visualization design choices drive results every bit as much as traditional "data-cleaning" choices. Visualization techniques contain embedded judgments.

### 2.3.1 What next? Research challenges

Researchers in cartography and geovisualistaion see big data as an opportunity and also as a certain call to action. The research agenda for geospatial big data and cartography layed down in Robinson et al. (2017) shows the general interest of moving the field toward fullfilling its potential to make maps that "pique interest, are tacitly understandable and are relevant to our society". It is certainly reassuring that the community is aware that new sources of data "stretch the limits of what and how we map". Building on this, Robinson et al. (2017) list several large-scale and long-term research challenges to face cartography in relation to big data as well as some short-term research oportunities for more concentrated investiagation (see appendix A for the overview). Even
```

though some points seem vague and repetative, and the imprint of the individual ICA commisions is clearly wisible, the agenda states some trully inspirative problems to tackle. In relation with the scope to this thesis we can highlight the following challenges for cartography:

- *Develop visual analytical reasoning systems that can help users add meaning to and organize what theay discover form geospatial big data* – we need to move beyond naive exploration and focus attention on tools that help people reason about what they are seeing. Users need to be able to save, annotate and compare their findings as they work on complex problems. - *Develop methods that embody the volume of geospatial big data* – we need cartography that can intelligently process and display big data at a size and a format that users can realistically handle. This will require solutions that support coupled analysis and visualisation as big data often need to be analysed before they are visualised (the order is reversed in exploratory visualisation). - *Create maps and map-oriented interfaces that prompt attention to important changes in dynamic geospatial big data sources* – We will need to work with global changes, local changes and combinations across scales. In addition, if we display every possible change at once, then the graphical displays become cluttered. Creating summaries of change may be the solution, but we do not yet know how to select important patterns and generalize to something that a user can understand. - *Leverage what we know about map animation and interactive cartography to construct visual solutions for dynamic sources of geospatial big data* – Conventional

> solutions for interactive mapping, animated mapping or geovisual analytics can be used for representing big data. However, because of the high velocity characteristic of big data, it is necessary to develop solutions that can automate map design decisions to support interactive design solutions that respond (or potentially precede based on modelled outcomes) as the data changes.
>
> Thomas & Cook (2005) also provide a set of recommendations for research and development agenda for visual analytics. Particularly resonating with goals of this thesis is their account on new visual paradigms, that include: - Organizing Large Collections of Information - Reasoning about Space and Time - Abstraction – Changing to the Appropriate Representation - Integrating Powerful Analysis Tools with Visualization
>
> Thomas & Cook (2005)

An emerging discipline progresses through four stages. It starts as a craft and is practiced by skilled artisans using heuristic methods. Later, researchers formulate scientific principles and theories to gain insights about the processes. Eventually, engineers refine these principles and insights to determine production rules. Finally, the technology becomes widely available. The challenge is to move from craft to science to engineering to systems that can be widely deployed. – my commentary: Cartography, being a universisty study field had arguably crossed the four stages in the past, though with interactive mapping it could benefit from retruning to the craft stages as the tools and possibilites for mapping changed profoundly.

Cognitive scientists have studied visual representations and the larger class of external aids to cognition. An external aid to cognition is an artifact that helps us reason about the world.

A first step in developing principles for visual representations is to understand how they enable cognition [Card, 1999; Norman, 1993]. Some basic principles for devel- oping effective depictions include the following (adapted from [Norman, 1993]):

• Appropriateness Principle – The visual representation should provide neither more nor less information than that needed for the task at hand. Additional information may be distracting and makes the task more difficult. • Naturalness Principle – Experiential cognition is most effective when the properties of the visual representation most closely match the information being represented. This principle supports the idea that new visual metaphors are only useful for representing information when they match the user's cognitive model of the information. Purely artificial visual metaphors can actually hinder understanding. • Matching Principle – Representations of information are most effective when they match the task to be performed by the user. Effective visual repre- sentations should present affordances suggestive of the appropriate action. Another prominent cognitive scientist has suggested the following two basic prin- ciples [Tversky et al., 2002]: • Principle of Congruence – The structure and content of a visualization should correspond to the structure and content of the desired mental repre- sentation. In other words, the visual representation should represent the important concepts in the domain of interest. • Principle of Apprehension - The structure and content of a

The subjects of mental representations and reasoning are the main focus of cog- nitive science, so the principles for depicting information must be based on research in cognitive science. The apprehension principle underlies the importance of research in perception. These meta-principles underscore that the biggest challenge in choos- ing a visual representation is to find the right one (not just any one) for the reasoning task at hand. – naive scientism? lecturing birds how to fly...

TODO – science of interaction just as a preview for next chapters: Too often in the visual analytic process, researchers tend to focus on visual representations of the data but interaction design is not given equal priority. We need to develop a "science of interaction" rooted in a deep understanding of the different forms of interaction and their respective benefits. Then, R&D should be focused on expanding the repertoire of interaction techniques that can fill those gaps in the design space.

role of cartoman: Creating effective visualization representations is a labor-intensive process that requires a solid understanding of the visualization pipeline, characteristics of the data to be displayed, and the tasks to be performed by the analyst. Current visualiza- tion software generally has been written in environments where at least some of this necessary information was missing.

In addition to the aforementioned agendas, we conclude this section with formulating a number of low-level challenges that we feel are not widely discussed. This thesis does not have the ambition to imagine all paths cartography could take, so we subsequently pose several questions related to the practice of map making that would inform the rest of content of this thesis. A mini-agenda for adjusting mapmaking to post 2020 cirumstances, if you please. As we have seen many times in history of innovation, progress is often hampered by the mental roadblock we don't even realize we have. (here merge challenges with questions?)

1. Is cartography fully exploitng the digital medium?

Before hopping on the wagon of augmeted reality and immersive experieces (that make a tenth of the population sick) cartographers could consider if they made the most of the previous medium shift. The same graphic technologies that power on the burst of immagination in web games, don't seem to bring much revolutionary changes to design in the map field (3D cartography being a honorary exception).

- Challenges stemming from the medium shift

Desktop GIS mapping – strugging to tranfer to web (basically fron-end development, which is also always in flux), maybe in the future desktop tools will suport generation of web map interfaces, but wouldn't count on it. Interface as a part of cartographic experience.

TODO maybe look at the state of the art (opiniation/freedom) - leaflet, openlayers, mapbox, cartodb, arcgis online, self-hosted mapbox alternative...

Apart from the limitations posed by opinionated mapping frameworks there are also certain mindset limitations that come from transfering a visual artifact from one medium to the other. Such transfer is not the same as if the visualisation was designed for the new medium from scratch as there are realized or unrealized ideas of how things should be done transrered from the practices required by the old medium. This was apparent for example the grid-like organization transferred from printed newspapers to web news portats initally and still lives there though responsivity required by small-screen devices pushed its rethinking. Simmilar case in cartography is the dichotomy between the topographic base and the tematic overlay. A good mental excercise for cartographers would be imagining map interaction unattached from any medium – what would we design if anything was possible?

Ford’s quote: “If I asked people what they want, they would ask for a faster horse” (find exact.). Similarly, we can test the cognitive efficiency of the visualisation methods that already exist, and users would prefer the methods they know. Cartography’s quest (in my opinion) is to extend the arsenal of visualisation and interaction methods. As we will see further, interaction and animation pose new challenges to cartographic visualisation, with possibly multiplied opportunities for method combinations and innovations for data exploration and possibly knowledge generation. Further, plenty of tricks from the rich history of cartographic practice did not make it to web mapping toolbox.

Recent emerging technology owing much to the gaming industry promise to bring web cartography to the flexibility of the pen and paper[26] of pre-digital cartographer. Only now the shifted role of cartographer would be in enabling data to paint the picture for us. Much of the rest of this thesis will be exploring this truly exciting prospect.

Limits of old media. Danger: processes of old media are transfered to new media, leaving possibilities in new tools unexplored.

e.g.

Narration as a workaround for cartographic rules – legibility, etc. Static map must adhere to the cartographic rules. In intercative maps (both presentational and exploratory) the argument is as follows: application doesn't need to be cartographically legit in all of it's states provided that it shows a path from the messy state to the cartographically treated state.

2. *What inspiraton can interactive web cartography take from the heritage of pre-digital mapping?*

Cartogaphic quality of web maps are not yet on par with the (best) examples of static maps and atlases (regardles of the date production – even though maps and atlases age in the sense of content, cartographic metods used in them often remain inspirational and valid, so for cartographer old map

[26] or brush, engraving tool etc.

products doesn't have to be outdated). This might be a side effect of non-cartographically aware people producing maps (either amateur cartographer acessing easy to use tools, or people comming from graphic design backgrounds bringing 'creative' changes).[27] Another possible cause is that the web mapping frameworks tend to provide just an opinionated set of visualisation options (TODO: picture of the google map's point sign) and cartographers lack the skills to customize or extend these visualisation toolkits. Little to say that some of the classical cartographic techniques are quite demanding when transfered to variable-scaled enviroment. For exaple, the symbol collsions that needed to be resolved just once (as if that was not enough) have to be treated (perhaps algoritmically) in all zoomlevels. (TODO: examples of such methods in pictures – call it "craftsmen" side of cartography) Cartographers who want to venture to raising the quality of web maps are then forced to dive into frond-end web development[28].

This question is complementary to the previous one and also arises from the conditions of the medium shift. What was lost in transition to digital? Are there methods and practices from tradition that could be used but aren't (because cartographers usually aren't software developers, and

[27] This is not intended to belittle such activities, as the eyesoreness of some maps is a little tax for democratization of mapping tools. In fact, efforts to help citizen cartographers have been made by more experienced practicioners, most notably in Wood, – TODO cituj tu ich prirucku

[28] The situation is improving with rise of web-gl based mapping platforms such as Mapbox and CartoDB, which provide map design environments resembling the visualisation tiers of desktop GIS

software developers are usually unaware of old map stocks). Danger: the perceived mantinels of the new technologies limit us in imaginig what we could do (e.g. what visualisation possibilities are provided by APIs like Leaflet).

Overlays are solved lazyly: point clusters vs. offsets and insets dynamic data examples: Global fishing watch - https://globalfishingwatch.org/map/

** 3. Cartography and UX, designing an interactive map is also designing ways to interact with data** Saving space: coupling legends with controls. Problematizing UX research, problems of cognitive testing (verification crisis in psychology) vs. A/B testing, (data on tile usage from big providers) Inspirations from product design (praktika lens, beat machines), photography (Muybridge and Marey), cinematography, video (youtube, other video services :P) - Also dealing with dynamic data more of a problem of intracomposition but extracomposition affected too (different scales, etc.)

amount of hand-holding? (move to one of the questions) Here we need probably some inspiration from other fields. Aim is moving somewhere inbetween the presentation and exploratory interfaces, possibly to get the best of the both. Exploratory interfaces could do some hinting, notifying which findings make sense and which not. Designers of exploratory interfaces could give greater thought to what questions users might want to ask about the portrayed data. (But careful on generalization with user personas – see below on weaponization of design).

TODO: - discussion of presentational vs exploratory cartography – or better on building interfaces to support one of them. – exploaratory interfaces seen loftier, most comercial assigmnents are presentational – in fact the threshold between presentatonal and exploratory capabilites is something that needs to be considered.

4. How to extend the spatio-temporal analysis faculty in digital maps.

The range of possible future states of the application is only to be guessed. In other words, the task is to design the map well for previously unseen data.

- challenges in finding causal relationships
- challenges in processing big data

Causation-related questions for cartography (TDOO) - finding spatio-temporal co-location that would suport causation hypotheseis is in currently realized by comparing spatial patterns. The causal delays may hamper such comparison, one approach is extend the time range of records (e.g. comparing cummulative data within two choroplets can smooth the volatilites in favor of the overall tendency). - Another approach is in looking for some general similarities between two sets of shapshots (spatial patterns) – if there is some similarity ocurring at some interval then we have identified the delay interval. This is spatial but not temporal collocation. Problem: this assumes causal relationships across the whole area of pattern – how to search for delay in just a sub area? - Temporal but not spatial collocation – is map a good tool for displaying this (rather a bar chart? Yes e.g moving air masses – we infer the future state in place from the state in past elsewhere) - What amount of apparent spatio-temporal collocation allows to rule out epiphenomena? Can map alone rule out a hidden common variable? - How to map causal-like relationships, e.g. potential for causation to happen via variations of state across the area? - overall, the ability of dynamic maps to find these collocations and link them to causation is to be assesd, but how? :)

2.3.2 Challenges in collaborative practice

Having described the ontological models of causation ans sensemaking as well as visions for the future of cartographic research, we can now take an aside to dwell a bit on the nitty-gritty realities of map making in practice. Practical aspects of the profession are often overlooked in literature, as well as the fact that cartographer often needs to operate within a greater team. The smoothness of collaboration within a team is then a determining factor of the team's productivity.

There is a (somewhat mythical) notion of "full-stack" visualization designer-developer capable of conducting the full broad range of tasks needed for a visulaisation project (Gray, Chambers, & Bounegru (2012)). Though some such individuals do exist (possibly working on small applications for PhD projects or small customers), it is clear that cartographer can take only so much of additional roles (data analyst, UX designer, front-end developer, database administrator…) before getting on thin ice. Real-life visulaisation projects often include a range of team members or even teams with dissjoint skillsets. The question then arises on how to modularize the work. One possible model of decomosition is the information visualisation reference model (**Fig**).

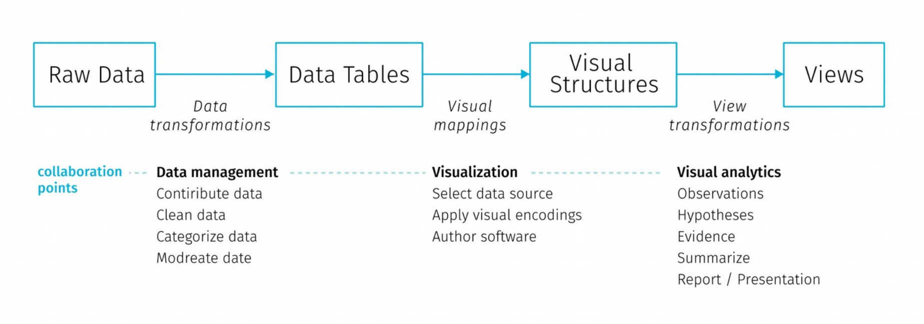


Fig *Information visualization reference model. Adopted from Heer & Agrawala (2008)*

In this model the collaboration points lie at the transitions between the stages and involve decsions on data management, visualisation and analytical capabilities (Heer & Agrawala (2008)). Physical and temporal separation of teams and institutional and disciplinary divides lead to early-stage partitioning of tasks both in the *design* (data profilation, ideation, mockup creation and prototyping) and *development* (implementation, testing, deployment and maintenance) phase (Walny et al. (2019)). Such compartmentalizaton is not unique to dataviz projects, it could match any web development project.

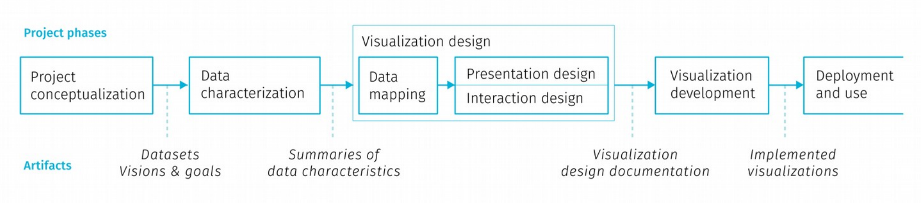


Fig *Stages of data visualization development process. Adopted from Walny et al. (2019)*

Walny et al. (2019) formalize stages of data visualisation process based on experience with several assignments (Fig.). This is an itrative process where the division of labor gives rise to *handoff* events, when one team passes work products and requirements to the next team. Particularly the handoff between the design and development team is where issues can arise to affect the end result. Speaking from the postion of design team Walny et al. (2019) articulate several key challenges that affect the success of the handoff and in turn the smoothness of the whole project:

- *Adapting to data changes* – changes in input data can have cascading effects throughout the stages of the process. Some breakages are inevitable, for example API alterations, and fixing them is a part of project maintenance. It is advisable to have data transformations automated to the largest extent possible, as it is highly likely there will be a need to reiterate them. In this sense, the scipts and the processing toolchain developed during the project can be more valuable to creators than its outputs.
- *Anticipating edge cases* – though this is incredibly hard for real-time data inputs, best effort should be made to

forsee at least the main application states resutling from the user intercations, such as filtering, changes of scale, etc.

- *Understanding technical challenges* – knowledge of technical constraints helps to produce feasible design ideas. Development team's concerns differ form the design team, they include cross-browser compatibility or future code maintainablity. In some areas the goals can overlap, for example in accessability considerations or performance optimization
- *Articulating data-dependent interactions* – prototyping interactions such as linking and brushing using conventional graphic tools is challenging, not to speak about articulaiting of animations and transitions between views. There are wireframing tools that try to address this, though misunderstandings still occur.
- *Communicating data mappings* – this is a concern when delivering static mockups for the development team. The mapping between data and the interface controls may not be obvious, especially when the chomplexity of data does not allow to exemplify all possible views in mockups. Annotations within mockups are a way to mitigate this.
- *Preserving data mapping integrity across iterations* – tracking implementation adherence to the design, finding errors, as well as chcking if change requests from previous iterations got implemented is solely a matter of visual inspection and therefore prone to error. This can be fixed by automated testing, though it is not

feasible for all types of projects, and even if implemented, the test coverage can rarely reach 100%.

These challenges were formulated based on project experience with relatively static historical data inputs, which underlines why dynamic geovisualistaion of real-time data is hard: much of the advice is almost impossible to follow for volatile real-time data inflow. TODO – something more

3 Recent objections to geospatial knowledge discovery

What is the role of cartography if:

1. It is not humans that make the decision (harrari, mayer-zukier)
2. Mapping complex systems? Beyond the rational naivety about the models (taleb) – prediction is not a goal, unpredictable events… risk mgmt… spatial modelling (naive rationalism) - risky inference, risky prediction, harmful intrusions to complex Pseudo-insights? Embracing what we canno't know (via negativa)

Non-human decsion makers

Add 1: - If human is dropped out of the equation, will we need visual analytics - man has checking, and qa function - man can direct search to speed up computations (interface needed) - algorithms can have biases - at least a proven communication tool (to pass on the results of computation to human, tailored to human cognitive capabilites)

We would mostly welcome automatization of many tedious tasks, and in realm of decision-making and orientation by the map we already do. AI Inference could go around some well known limitations of human analysts such as information overload in complex situations, inherent and unrecognized biases or tendency to settle for convenient anwers (satisficing) (Thomas & Cook (2005))

. Techniques are needed to help analysts simplify their cognitive load without compromising the analyst's effectiveness and to help compensate for faulty memory.

• *Overcoming biases*. Biases affect the way data are interpreted. Biases about the reliability of different sources may lead people to discount information from sources that aren't considered reliable. People often see what they expect to see and tend to ignore evidence that is contradictory to a preferred theory. If they form a preliminary judgment too early in the analytical process, they may hold firm to it long after the evidence invalidates it [Heuer, 1999].

• *Satisficing*. People settle for a "good enough" answer, sometimes stopping their analytical process before they identify critical information that would lead them to a different conclusion [Heuer, 1999]. New interaction techniques are needed to support the user in evaluating evidence, challenging assumptions, and finding alternatives. Analytical environments should support the user in identifying and understanding all relevant information to reach a solid conclusion rapidly. The tools we create need to establish a correct balance between structure and intuition.

- Add human cognition: not uniform (makes sense to look at otliers rather than the general populations, – tailoring for elderly, disabled, visually impaired), evolves with media usage (some abilties strengthen, some weaken)

data mining –> machine learning –> automated actions – no human interpretation, no need for insight? role of carto?

First do we event want to get there, best of both worlds is maybe better approach (TODO cite Thiel, also on Keeping up with machines, article on not replacing human labor) Second, hard vs easy problems – ai can now do only “easy”, will “hard” ever be possible? (general AI – TODO cite Pinker) – here, are problems solvable by maps hard or easy, better: which are hard and which are easy? – here, are problems solvable by maps hard or easy, better: which are hard and which are easy? Also the difference between AI in place of cartographers (which map-making tasks can be automated – probably those tedious ones – digitalization) and AI for cartographers (what can we now do better? Tooling and process improvements)

Thompson, Greenewald, Lee, & Manso (2020): Deep learning’s recent history has been one of achievement: from triumphing over humans in the game of Go to world-leading performance in image recog- nition, voice recognition, translation, and other tasks. But this progress has come with a voracious appetite for computing power. This article reports on the computational demands of Deep Learning applications in five prominent application areas and shows that progress in all five is strongly reliant on in- creases in computing power. Extrapolating forward this reliance reveals that progress along current lines is rapidly becoming economically, technically, and environmentally unsustainable. Thus, continued progress in these applications will require dramatically more computationally-efficient methods, which will either have to come from changes to deep learning or from moving to other machine learning methods.

TODO - revisit after reading Shane (2019) Issues – human biases can get incorporated

3.2 Dangers of naive rationalism

Causal relationships vs. epiphenomena. – see Taleb 198-200. (Maps helping to tell?)

vis for humans, machines do not need it

– the societal responsibility of designers (first things first movement?, something for data visualists), and cartographers (harley, wood, crampton, – radical cartography (Paglen?))

Meadows (2008) Donnela meadows on systems with delay – good for describing systems with already known structure (man-made), maybe use something on complex systems

Excellent map provides answers, helps to ask queations and supports understanding in an engaging way. It is perfectly OK if the conveyed message is a lack of pattern, lack of reagularity or inability to rasonably identify correlation.

It is always possible to say when not to use the map. Robinson et al. (2017)

weaponized design

Diehm (2018)

Weaponised design – a process that allows for harm of users within the defined bounds of a designed system – is faciliated by designers who are oblivious to the politics of digital infrastructure or consider their design practice output to be apolitical.

This is weaponised design: electronic systems whose designs either do not account for abusive application or whose user experiences directly empower attackers.

As platforms became more commodified – especially through mobile touch mediums – UX designers have progressively become more reliant on existing work, creating a feedback loop that promotes playfulness, obviousness and assumed trust at the expense of user safety.

A user story is “a very high-level definition of a requirement, containing just enough information so that the developers can produce a reasonable estimate of the effort to implement it”. (definition from Ambler (2014))

When designing for the digital world, user stories ultimately determine what is or is not an acceptable area of human variation. The practice empowers designers and engineers to communicate via a common problem-focused language. But practicing design that views users through a politically-naive lens leaves practitioners blind to the potential weaponisation of their design. User-storied design abstracts an individual user from a person of lived experience to a collection of designer-defined generalisations.

All intentionally-created systems have a set of things the designers consider part of the scope of what the system manages, but any nontrivial system has a broader set of impacts. Often, emergence takes the form of externalities — changes that impact people or domains beyond the designed scope of the system. Henriksen, Hyltoft, Prudkovskaya, Vygotsky, & Saitta (2016)

Through inclusion, participatory design extends a design team’s focus beyond the hypothetical or ideal user, considering the interactions between users and other stakeholders over user stories.

In particular, security research and user experience design have significant practice and goal overlap and this relationship is often antagonistic. Both fields primarily focus on the systems of wide-scale interactions between users and technology, but the goals of the two fields are diametrically opposed; design is to create the best possible experience for a user, security is to create the worst possible experience for an attacker.

```

# Conclusion

BD are fast and n=all so we need to quickly make sense of it or quickly establish there is no sense present.

Van Rijmenam (2013) on visulaisation: "Visualizing might not be the most technologically difficult part; it sure is the most challenging part. Telling a complex story in a graph is very difficult but also extremely crucial. Luckily there are more and more big data startups appearing that focus on this aspect and in the end, visualizations will make the difference. One of them is future this will be the direction to go, where **visualizations help organisations answer questions they did not know to ask.**"

“Discovery consists of seeing what everybody has seen and thinking what nobody has thought.” —Albert von Szent-Gyorgyi (1893–1986)
```

Aigner, W., Miksch, S., Schumann, H., & Tominski, C. (2011). *Visualization of time-oriented data*. Springer Science & Business Media.

Allen, E., Edwards, G., & Bédard, Y. (1995). Qualitative causal modeling in temporal gis. In *International conference on spatial information theory* (pp. 397–412). Springer.

Allen, J. F. (1984). Towards a general theory of action and time. *Artificial intelligence*, *23*(2), 123–154. Elsevier.

Ambler, S. W. (2014). User stories: An agile introduction. *Retrieved from Agile Modeling*.

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.

Bleisch, S., Duckham, M., Galton, A., Laube, P., & Lyon, J. (2014). Mining candidate causal relationships in movement patterns. *International Journal of Geographical Information Science*, *28*(2), 363–382. Taylor & Francis.

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.

Clarke, V., & Pickles, R. (2015). *Map: Exploring the world*. Phaidon Press Limited.

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.

Diehm, C. (2018). On weaponised design. *Available online at https://ourdataourselves.tacticaltech.org/posts/30-on-weaponised-design/ (last accessed September 16, 2018)*.

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.

Egenhofer, M. J., & Franzosa, R. D. (1991). Point-set topological spatial relations. *International Journal of Geographical Information System*, *5*(2), 161–174. Taylor & Francis.

El-Geresy, B. A., Abdelmot, A. I., & Jones, C. B. (2002). Spatio-temporal geographic information systems: A causal perspective. In *East european conference on advances in databases and information systems* (pp. 191–203). Springer.

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*.

Galton, A. (2012). States, processes and events, and the ontology of causal relations. IOS Press.

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.

Gray, J., Chambers, L., & Bounegru, L. (2012). *The data journalism handbook: How journalists can use data to improve the news*. " O'Reilly Media, Inc.".

Guo, D., Chen, J., MacEachren, A. M., & Liao, K. (2006). A visualization system for space-time and multivariate patterns (vis-stamp). *IEEE transactions on visualization and computer graphics*, *12*(6), 1461–1474. IEEE.

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*.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. Elsevier.

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

Heer, J., & Agrawala, M. (2008). Design considerations for collaborative visual analytics. *Information visualization*, *7*(1), 49–62. SAGE Publications Sage UK: London, England.

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

Henriksen, T. D., Hyltoft, M., Prudkovskaya, O., Vygotsky, L., & Saitta, E. (2016). Infrastructural games and societal play. *Larp Politics*, 72.

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*.

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.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

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.

Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). Visual analytics: Definition, process, and challenges. In *Information visualization* (pp. 154–175). Springer.

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.

Laube, P., Dennis, T., Forer, P., & Walker, M. (2007). Movement beyond the snapshot–dynamic analysis of geospatial lifelines. *Computers, Environment and Urban Systems*, *31*(5), 481–501. Elsevier.

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)*.

Meadows, D. H. (2008). *Thinking in systems: A primer*. chelsea green publishing.

Miller, H. J. (2015). Spatio-temporal knowledge discovery. *Geocomputation: A Practical Primer. SAGE Publications Ltd, Thousand Oaks, CA*, 97–109.

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.

Press, G. (2014). 12 big data definitions: What's yours.

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

Robinson, A. C., Demšar, U., Moore, A. B., Buckley, A., Jiang, B., Field, K., Kraak, M.-J., et al. (2017). Geospatial big data and cartography: Research challenges and opportunities for making maps that matter. *International Journal of Cartography*, 1–29. Taylor & Francis.

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

Shane, J. (2019). *You look like a thing and i love you: How artificial intelligence works and why it's making the world a weirder place*. Voracious.

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.

Sontag, S., & others. (1994). *Against interpretation*. Vintage London.

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. (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.

Thomas, J., & Cook, K. A. (2005). Illuminating the path: The r&D agenda for visual analytics national visualization and analytics center. *National Visualization and Analytics Center-US Department of Homeland Security*.

Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2020). The computational limits of deep learning. *arXiv preprint arXiv:2007.05558*.

Tufte, E. R., McKay, S. R., Christian, W., & Matey, J. R. (1998). Visual explanations: Images and quantities, evidence and narrative. *Computers in Physics*, *12*(2), 146–148. AIP Publishing.

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)*.

Van Wijk, J. J. (2005). The value of visualization. In *VIS 05. IEEE visualization, 2005.* (pp. 79–86). IEEE.

Verhein, F., & Chawla, S. (2008). Mining spatio-temporal patterns in object mobility databases. *Data mining and knowledge discovery*, *16*(1), 5–38. Springer.

Walny, J., Frisson, C., West, M., Kosminsky, D., Knudsen, S., Carpendale, S., & Willett, W. (2019). Data changes everything: Challenges and opportunities in data visualization design handoff. *IEEE transactions on visualization and computer graphics*, *26*(1), 12–22. IEEE.

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).

Worboys, M. F., & Duckham, M. (2004). *GIS: A computing perspective*. CRC press.

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.