1

Defining Big Data

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

Kitchin & McArdle (2016)

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

1.1 Ontological characteristics

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

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.

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”

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 Mayer (2013): “Data was no longer regarded as static or stale, whose usefulness was finished once he purpose for which it was collected was achieved [...]. Rather, data became a raw material of business, a vital economic input, used to create a new form of economic value. Every single dataset is likely to have some intristic, hidden, not yet un-

earthed 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) argue 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 autmated algorithmic analysis supplants domain expertise".

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, discuss general ways how cartography can support understanding the world trough the lens of big data. We will also discuss some objections

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

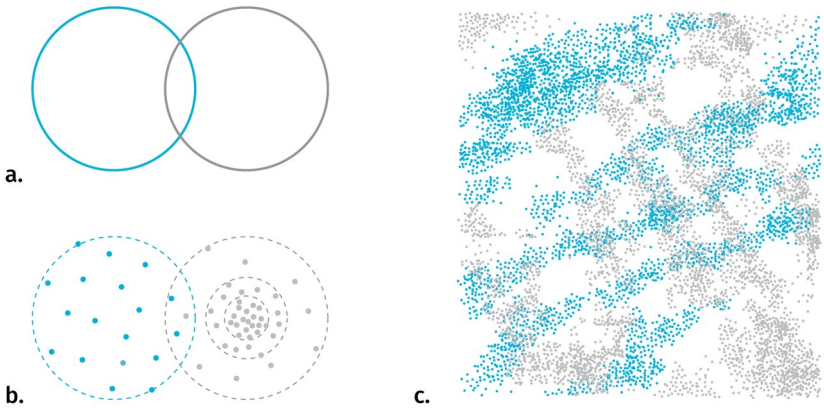

Fig.2 *With polygonal features it is straightforward to identify the type of spatial relationship (a). When replacing point clouds with polygon representations to apply set logic, the problem of meaningful boundary delineation arises (b). For several complex layers it is hard to say anything revealing about their spatio-temporal relationship (c)*

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 collection of any entity can be either continuous or discrete in time, it is therefore useful to distinguish between *time point* and *time interval* when investigating temporal relations (see figures). Linear conceptualization of time can be supported with cyclical and branching time, there can be discrepancies between the temporarily of base map and the thematic overlay,

or between the time interval of existence and representations. We'll untangle these complexities in chapter 5.