2. Theoretical Framework

2.1 Big Data

2.1.1 Etymological Origin

The pursuit of understanding big data first requires exploring the term *data*. It subsumes a variety of meanings and ideas. It is loaded with contextual meaning and depends on the beholder's point of view; there are different perspectives of data. That aside, the term can be derived etymologically as follows:

> "English *data* is derived from Latin, where it is the plural of *datum*, which is in turn the past participle of the verb *dare*, "to give," generally translated into English as "something given." Sanskrit *dadāmi* and ancient Greek *δίδωμι* are related forms. While *data* (piece of information) and *datum* (calendar date) are separate lexemes in contemporary English, their association is not accidental; medieval manuscripts frequently closed with the phrase *datum die* (given on …), effectively time-stamping the preceding text" (Puschmann & Burgess 2014: 1691).

In addition to its variety of context-determined meanings, the denotation of the term has shifted over time. In the 18th century, it represented a rather quantitative understanding as it "was most commonly used to refer to facts in evidence determined by experiment, experience, or collection" (Rosenberg 2013: 33). Rosenberg himself specifies this point of view by claiming that "facts are ontological, evidence is epistemological, data is rhetorical. A datum may also be a fact, just as a fact may be evidence … When a fact is proven false, it ceases to be a fact. False data is data nonetheless" (2013:18). Nowadays, however, any mention of data is likely to refer to their digital sense. They are perceived as a common resource generated without any effort and without any loss of information. This is a precise description of today's ubiquitous generation of data. It is for this reason that data are often referred to as new oil (Thorp 2012, Helbing 2015) or lead to a new gold rush (Peters 2012). The latest conception of *data* can be outlined as "anything recordable in a relational database in a semantically and pragmatically sound way" (Frické 2015: 652).

Some researchers (e.g. Kitchin 2014a), however, claim that the term *data* fails to precisely capture the described phenomenon in modern contexts. They suggest the use of *capta* (from the Latin word *capere* which means to take) instead. Data in the modern sense are the extraction of elements through observation, recording and other means (Borgmann 2007), *data* or *capta* are taken from all potential data (Kitchin & Dodge 2011). This etymological permutation is described the following way:

> "It is an unfortunate accident of history that the term datum … rather than captum … should have come to symbolize the unit-phenomenon in science. For science deals, not with 'that which has been given' by nature to the scientist, but with 'that which has been taken' or selected from nature by the scientist in accordance with his purpose" (Jensen 1952: ix).

Although, the term *capta*, therefore, bears more precision than the term *data*, the term data has become generally accepted.

The logical next step is analysis of the term *big*. Heuristically speaking, *big* describes something large in size, large in number, or involving many people or things. Applying this to data allows for the inclusion of huge data sets and correlates to the challenge of dealing with an "information explosion" (Marron & de Maine 1967: 711) and, subsequently, the belief that this kind of "information overload" (Eppler & Mengis 2004: 325) leads to a "data avalanche" (Miller 2010: 181) or "data deluge" (Bell et al. 2009: 1297), and that we are "facing the waves of big data" (Marder 2015: 2). There is, however, more than meets the eye in the simple term *big data*. In order to draw a more precise picture of the term, it is essential to review its chronological history.

One of the earliest examples of big data analysis is attributed to John Snow in 1854 (Khoury & Ioannidis 2014). London had been struck by an outbreak of cholera, and Snow collected all available data about the deaths and was able to locate their origin to the area around Broad Street. He hypothesized a connection between the outbreak and a specific water pump. Shutting down the pump led to a significant reduction in the number of new infections (e.g. McLeod 2000, Koch 2004, Johnson 2007). Snow collected data, used it to develop a hypothesis and derived an action from it (Khoury & Ioannidis 2014) and this can be described as data-driven science. His results can be seen in Figure 1. There are several other examples that can potentially be retrospectively attributed to the use of big data. Snow's example, however, is exceptionally well documented, ultimately led to the beginnings of geographical epidemiology (Newsom 2006), and is a prominent example of the early visualization of information (Friendly 2008).

Figure 1: Original Map Used by Snow (1854)

Even though Snow's analytical effort is seen as an example of using big data, the term itself is still relatively young. Its origin, however, is currently under debate. Diebold (2012) attributes the first use of big data to a work by Tilly (1984), and to the use of data analysis for historians. Diebold explains that 'big data' was used in the context of computer science by Weiss and Indurkhya (1998), and by himself in econometrics in 2000. Others (e.g. O'Leary 2013) claim Cox and Ellsworth (1997) and related follow-up research (Bryson et al. 1999) to be the earliest contributions to the term development as used today. Recent research dates the first academic use of 'big data' to 1969 (Scholz 2015a). Even though this early reference uses the term 'big data', the connection to its present conception is vague. Nonetheless, the term 'big data' was used frequently in the 1960s and 1970s. However, it may just be a coincidence that big was combined with data. Table 1 cites several occurrences of the term big data, which clearly foreshadow modern terminology.

Table 1: The Term "Big Data" in the Years 1961–1979

Source	**Quotes with the Term "Big Data"**
U.S. Congress (1961: 197)	"So I think it is quite important that we do not end up doing a *big data*-collecting job, with a quick, casual look at it and that being the end of it."
Kates (1969: 50)	"Most geographers are for *big data* banks, most support an expanded range of census questions, most accept in some vague general way the notion that the more we know about people the better off we are."
DPMA (1970: 8)	"Instead of a *big data* dump where all information collected by government agencies on all Americans would be gathered, he proposes the following…"
Exemplary Miller (1971: 253)	"Eventually, the governance of data centers may fall into the hands of those we now jokingly refer to as 'computerniks', creating a danger that policy will be formulated by information managers who are so entranced with operating sophisticated machines and manipulating *large* masses of *data* that they will not be sufficiently sensitive to privacy considerations."
U.S. Senate (1972: 1270)	"In actual fact, the practice has spawned *big data* center bureaucracies at taxpayer expense. Industry claims millions of dollars are wasted each year – as each federal agency tries to build its own data empire."
Merriam (1974: 40)	"In the future, *big data* storage and retrieval systems will be put into use."
Bassler and Joslin (1976: 300)	"A *big data* center may handle several thousand tapes a day. In addition to tracking the use of tapes and disks, the librarian must be an expert in the care and preservation of the tape and disk media."

Source	Quotes with the Term "Big Data"
Patrick (1977: 35)	"More and more it is becoming apparent that a *big data* processing system requires careful design attention to be given both to the computer processing and the manual processes such as data capture, balancing, error correction, reports distribution that support the computer system."
Müller (1979: 11)	"The dreams of *big data* banks – that would even work – of course raised public fears against the uncontrolled circulation of personal information."

Many issues mentioned in early sources are still current: concerns about data analysis, data accumulation by the government, increasing complexity, and the essential need for the precise design of big data systems. There is also the question of privacy (e.g. Müller 1979). In the book "The assault on privacy", Miller (1971) tackles several aspects of potential surveillance by means of big data (he calls it large data), one of which is the individual loss of control over personal information. Another element is the general tendency to quantify people based on their data. Interestingly, Miller already mentions the delicate challenge this entails for humanity: "Perhaps the single most imperative need at this point in time is a substantial input from human resources to help solve the difficult problem of balancing privacy and efficiency" (1971: 259). While Miller himself was a lawyer, he stressed the crucial need for any discipline to deal with the subject. "There will be no one to blame but ourselves if we then discover that the mantle of policymaking is being worn by those specially trained technicians who have found the time to master the machine and have put it to use for their own purposes" (Miller 1971: 260).

Although a certain interest in big data can be seen, and various people discussed relevant questions (that have yet to be answered), the term 'big data' was used only sporadically in the following years. Apart from Tilly (1984), Cox and Ellsworth (1997), Weiss and Indurkhya (1998), Bryson et al. (1999), and Diebold (2000), no substantial contributions to big data research followed at first. In 2001, however, the paper "3D data management: Controlling data volume, velocity, and variety" by Gartner analyst Douglas Laney moved big data into the focus of business and academia. Laney's article can be understood as the foundation of numerous studies of big data.

2.1.2 Epistemological Conceptualization and Hermeneutical Observations

In order to approximate big data from a hermeneutical perspective, it is necessary to identify the embodiments of data, the first one being its incompleteness:

> "Data harvested through measurement are always a selection from the total sum of all possible data available – what we have chosen to take from all that could potentially be given. As such, data are inherently partial, selective and representative, and the distinguishing criteria used in their capture has consequence" (Kitchin 2014a: 3).

In his seminal work on big data, Kitchin derived various types of data. He categorized them as follows (Kitchin 2014a: 4):

- Form: Qualitative or quantitative
- Structure: Structured, semi-structured, or unstructured
- Source: Captured, derived, exhaust or transient
- Producer: Primary, secondary, or tertiary
- Type: Indexical, attribute, or metadata

Data, in general, are the basis of information and generate knowledge. In the field of knowledge management this process is seen as following a hierarchy (Alavi & Leidner 2001). Data by themselves are informative but do not give insights that are usable in decision making or planning. Only through context (Kidwell et al. 2000), and supplied with meaning and by understanding relationships (Alavi & Leidner 2001), do data become information. Information transforms into knowledge when combined with experience, cognition, and competence (Zins 2007). Knowledge is, therefore, necessary in order to deal with given data and information (Kebede 2010). Other researchers (e.g. Adler 1986, Weinberger 2011) describe the process as a pyramid and an inherent process of distillation that moves up the pyramid, thus reducing complexity, organizing information, interpreting, and finally applying processed data to decisions (McCandless 2010). Weinberger illustrates the process as follows: "Information is to data what wine is to the vineyard: the delicious extract and distillate" (2011: 2). Consequently, data can be processed into something useful, but, thereby, data will be transformed.

As Kitchin notes, "data are never simply just data; how data are conceived and used varies between those who capture, analyse and draw conclusions from them" (2014a: 4). Consequently, data may be sufficiently defined; nowadays however, given their omnipresence data are proverbially multiplying, thus prompting the need for a different framing.

Although the interest in big data coincided with the paper written by Laney (2001), the main reason for the exponential growth in big data can be attributed to the turning point at which storing digital data became cheaper and more cost-effective than storing data on paper (Morris & Truskowski 2003). More and more data are generated digitally and digitization allows them to be shared. Interestingly, many issues affecting the growth of big data follow Moore's law (Schaller 1997) of exponential growth. Some researchers (e.g. Dinov et al. 2014) suggest that although computational capabilities still follow Moore's law, data acquisition behaves according to Kryder's law (Walter 2005), which suggests that data volume is growing at an even higher pace. Computational power, however, remains a main driver for the success of big data, and nowadays it is possible to conduct elaborate big data research on an average computer (Murthy & Bowman 2014).

All things considered, we truly are using *big* data, and the technological perspective suggests that it is growing exponentially. Even though data are ubiquitous in society, however, there is no unified definition of what big data really are. An initial point in the discussion of big data is its classification in dimensions.

This perspective stems from the original paper by Laney (2001), which categorized big data into three dimensions: volume, variety, and velocity. *Volume* denotes the amount of data that is collected. Big data volumes are currently measured in petabytes (1,000 terabytes), however, the amount of data collected is rapidly increasing (McAfee & Brynjolfsson 2012). The dimension of *variety* marks the types and forms in which data are collected. Data can be structured or unstructured, and there are numerous forms: numbers, text, audio, and video (Aakster & Keur 2012), to name only a few. *Velocity* refers to the pace at which data are generated and analyzed. The issue of speed can be dealt with by focusing on data collection, or the challenge of parsing data in real-time (Hendler 2013). Within the course of the following years, a variety of new dimensions were added as depicted in table 2.

Table 2: Dimensions of Big Data

Dimension	Definition
Volume	"E-commerce channels increase the depth/breadth of data available about a transaction (or any point of interaction)" (Laney 2001: 1).
Variety	"Through 2003/04, no greater barrier to effective data management will exist than the variety of incompatible data formats, non-aligned data structures, and inconsistent data semantics" (Laney 2001: 2).
Velocity	"E-commerce has also increased point-of-interaction (POI) speed and, consequently, the pace data used to support interactions and generated by interactions" (Laney 2001: 2).
Veracity	"Data uncertainty. Veracity refers to the level of reliability associated with certain types of data. […] The need to acknowledge and plan for uncertainty is a dimension of big data that has been introduced as executives seek to better understand the uncertain world around them" (Schroeck et al. 2012: 5).
Variability	"In addition to the speed at which data comes your way, the data flows can be highly variable – with daily, seasonal and event-triggered peak loads that can be challenging to manage" (Troester 2012: 3).
Complexity	"Difficulties dealing with data increase with the expanding universe of data sources and are compounded by the need to link, match and transform data across business entities and systems. Organizations need to understand relationships, such as complex hierarchies and data linkages, among all data" (Troester 2012: 3).

Dimension	Definition
Value	"The economic value of different data varies significantly. Typically there is good information hidden amongst a larger body of non-traditional data; the challenge is identifying what is valuable and then transforming and extracting that data for analysis" (Dijcks 2013: 4).
Viability	"Our first task is to assess the viability of that data because, with so many varieties of data and variables to consider in building an effective predictive model, we want to quickly and cost-effectively test and confirm a particular variable's relevance before investing in the creation of a fully featured model" (Biehn 2013).

Listed here are only the most prominent dimensions used to describe big data. There are, however, several more, including visualization (van Rijmenam 2013), valorization (Özdemir et al. 2013), validity, venue, vocabulary, vagueness (Borne 2014), versatility, volatility, virtuosity, vitality, visionary, vigor, viability, vibrancy, and virility (Uprichard 2013). In addition to using Vs, researchers have recently expanded their choice of dimensions into the Ps, such as privacy (Agrawal et al. 2015), portentous, perverse, personal, productive, partial, practices, predictive, political, provocative, polyvalent, polymorphous, and playful (Lupton 2015).

Academic discourse is currently raising the question of whether to view "big data as merely a shift in scale, reach, and intensity (a quantitative shift) or as a more profound, truly qualitative shift – implying both a shift in being (ontology) and meaning (epistemology)" (Bolin & Schwarz 2015: 2). Bolin and Schwarz (2015) also point out that big data can be classified by a heuristic logic, or can foster a religiously tainted dataism (van Dijck 2014). This may cause an increase in datafied interpretations resulting in a more "anti-hermeneutical impulse (naïve empiricism)" (Bolin & Schwarz 2015: 2). The starting point of the discussion, however, is that "the world grows in complexity, overwhelming us with the data it generates" (Chakrabarti 2009: 32). This means, "it demands a systematic response" (Bowker 2014: 1797).

A great deal of literature on the topic consists of some sort of subliminal judgment of big data. Some sources praise its potential for making the world a better place (e.g. Smolan & Erwitt 2013). According to Pentland, for example, big data will help "build a society that is better at avoiding market crashes, ethnic and religious violence, political stalemates, widespread corruption, and dangerous concentrations of power" (2014: 16). Others focus on the challenges and obstacles that accompany the use of big data, one of which is the assumption that "big data continues to present blind spots and problems of representativeness, precisely because it cannot account for those who participate in the social world in ways that do not register as digital signals" (Crawford et al. 2014: 1667). The two camps both praise and criticize big data in various ways which, although the arguments are versatile, go to show that big data is a multi-faceted term.

Puschmann and Burgess (2014) take a similar approach by conceptualizing big data on the basis of two metaphors. Firstly, they claim that "big data is a force of nature to be controlled" (2014: 1698), which is often associated with the natural force of water. Society is drowning and will deal with data floods or data tsunamis in some way. The authors claim that the analogy of water fits in the sense that water is neutral and able to exist without humans. With the appropriate technology, however, both water and data can be harnessed. The second metaphor is that "big data is nourishment/fuel to be consumed" (2014: 1700) which aligns with the idea of data as "the new oil" (Helbing 2015) and especially the concept of data as a resource. Both of these "metaphors are crucial narrative tools in the popularisation of knowledge" (van Dijk 1998: 22). Nevertheless, the opposing assumptions of the two metaphors strengthen the argument that the term is still nascent.

Big data are highly debatable in terms of the way in which data are analyzed; the term itself seems to be opaque as well. Using dimensions helps approximate the concept of big data, but they only seem to tackle certain aspects, and are, therefore, not sufficient for exhaustively defining big data. Defining big data is a difficult and complex task. In addition, subjective preconceptions influence the process of definition. Statements such as "Across all disciplines, data are considered from a normative, technological viewpoint" (Kitchin 2014a: 12) reveal the obstacles of defining big data in a logical way. Similar to the is-ought problem or Hume's law (Hume 1739), the process of describing big data faces the inherent problem of researchers making statements about what ought to be without being capable of deriving any descriptive statements whatsoever. From a technological viewpoint, a computer cannot differentiate between descriptive and prescriptive. With enough data, anything becomes a standard (Helland 2011). Both human and machine contribute to a big data fallacy that leads to a variety of hermeneutic judgments about big data. From a normative perspective, both camps can be categorized as within subjective objectivity (Leahu et al. 2008) and objective subjectivity (Diller 1997).

If big data had actually marked a scientific revolution (Kuhn 1962) and led to a paradigm shift (Kitchin 2014b) in science as well as in business, former standards would no longer be applicable. Several researchers claim big data as the fourth paradigm by going beyond the experimental, theoretical and computational sciences. Gray states: "The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science as a new, *fourth paradigm* for scientific exploration" (Hey et al. 2009: xix). The question of whether big data lead to a paradigm shift or are merely more 'hype' (Kitchin 2013) or a fad (Marr 2015) may be debatable (e.g. Gandomi & Murtaza 2015). Nevertheless, big data have the ability to disrupt and challenge existing norms and standards (Boyd & Crawford 2012). Such newness and uniqueness, however, would mean that social norms and technological standards are not accurately fitted for such a novel concept as big data.

It is, therefore, essential to incorporate the variety of normative perceptions into an epistemological conceptualization and embrace the hermeneutical bias of researcher and machine. The way in which big data are perceived can be categorized,

and there are, in fact, already various categorizations in existence. Kitchin (2014a) divides big data into technical, ethical, political and economic, temporal and spatial, and philosophical viewpoints. De Mauro et al. (2014) classify big data by means of the following four themes: information, technologies, methods, and impact. Boyd and Crawford (2012) categorize big data into cultural, technological, and scholarly phenomena.

As previously mentioned, therefore, the term is opaque, and there are also a variety of definitions available. De Mauro et al. (2015) conducted a survey of existing definitions as shown in table 3. The authors systematically reviewed the literature until on the 3rd of May 2014, their corpus had reached a volume of 1,437 conference papers and articles with the term 'big data' as part of their title or on the list of keywords. The data coincide with the list of definitions postulated by Ward and Barker (2013).

Table 3: Existing Definitions of Big Data

Source	Definition	I	T	M	P
Beyer & Laney (2012)	High volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.	X		X	X
Dijcks (2012: 3–4)	The four characteristics defining big data are volume, velocity, variety and value.	X			X
Intel (2012: 3)	Complex, unstructured, or large amounts of data.	X			
Suthaharan (2014: 71)	Can be defined using three data characteristics, cardinality, continuity and complexity.	X			
Schroeck et al. (2012: 5)	Big data is a combination of volume, variety, velocity and veracity that creates an opportunity for organizations to gain competitive advantage in today's digitized marketplace.	X			X
NIST (2014: 5)	Extensive data sets, primarily in the characteristics of volume, velocity and/or variety that require a scalable architecture for efficient storage, manipulations, and analysis.	X	X		
Ward & Barker (2013: 2)	The storage and analysis of large and/or complex data sets using a series of techniques including, but not limited to, NoSQL, MapReduce and machine learning.	X	X		

Source	Definition	I	T	M	P
Microsoft (2013)	The process of applying serious computing power, the latest in machine learning and artificial intelligence, to seriously massive and often highly complex sets of information.	X	X	X	
Dumbill (2013: 1)	Data that exceeds the processing capacity of conventional database systems.	X	X		
Fisher et al. (2012: 53)	Data that cannot be handled and processed in a straightforward manner.	X		X	
Shneiderman (2008)	A data set that is too big to fit on a screen.	X			
Manyika et al. (2011: 1)	Data sets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.	X	X	X	
Chen et al. (2012: 1166)	The data sets and analytical techniques in applications that are so large and complex that they require advanced and unique data storage, management, analysis, and visualization technologies.	X	X	X	
Boyd & Crawford (2012: 663)	A cultural, technological, and scholarly phenomenon that rests on the interplay of technology, analysis and mythology.		X	X	X
Mayer-Schönberger & Cukier (2013: 12)	Phenomenon that brings three key shifts in the way we analyze information that transform how we understand and organize society: 1. More data, 2. Messier (incomplete) data, 3. Correlation overtakes causality.	X		X	X
Legend: I – Information, T – Technology, M – Methods, P – Impact.					

(adapted from De Mauro et al. 2015: 102)

From the above, De Mauro et al. derived the following definition: "Big data represents the information assets characterized by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value" (2015: 103). Beyond this extensive list, there are further definitions of the term 'big data'. Kitchin (2014b: 1–2) characterizes big data as massive in volume, quick in velocity, distinct in variety, exhaustive in its domain, granular in its resolution, relational in its structure, flexible, and scalable in its consistence. Kitchin (2014a) derived four other characteristics from additional literature (Boyd & Crawford 2012, Dodge & Kitchin 2005, Marz & Warren 2012, Mayer-Schönberger & Cukier, 2013):

- "*exhaustive* in scope, striving to capture entire populations or systems (n = all), or at least much larger sample size than would be employed in traditional, small data studies;
- fine-grained in *resolution*, aiming to be as detailed as possible, and uniquely *indexical* identification;
- *relational* in nature, containing common fields that enable the conjoining of different data sets;
- flexible, holding the traits of *extensionality* (can add new fields easily) and *scalable* (can expand in size rapidly)" (Kitchin 2014a: 68).

In addition to these definitions, Dutcher (2014) asked 43 "thought" leaders to give their own definition of big data. Ashlock (Chief Architect of Data.Gov), for example, defines big data the following way:

> "While the use of the term is quite nebulous and is often co-opted for other purposes, I've understood 'big data' to be about analysis for data that's really messy or where you don't know the right questions or queries to make – analysis that can help you find patterns, anomalies, or new structures amidst otherwise chaotic or complex data points."

Many definitions converge towards those suggested by Kitchin. Upadhyay (CEO of Lattice Engines), however, implies that big data may be an umbrella term bearing diverse meanings. O'Neil (Columbia University) attributes rhetorical potential to big data, thus identifying it as a tool of manipulation. To Murphy (Consulting Data Scientist), the word *big* is the key. His qualitative evaluation of the term 'big data' emphasizes its complexity of definition. *Big* data are more than meets the eye.

Due to the fact that big data are too wide-ranging and vague in definitions, some practitioners declare big data as already dead (e.g. de Goes 2013). Although big data may be vague and cannot be precisely pinpointed, the term itself describes the current challenge of datafication as faced by society, organizations, and individuals alike. A precise definition of big data is, therefore, probably not possible, as such a definition would turn out to be *big* as well. Jacobs, however, suggests the following meta-definition, picking up the umbrella concept of Upadhyay: Big data refer to "data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time" (2009: 44). Subsequently, "The challenge is not just a technological one: the selection, control, validation, ownership and use of data in society is closely related to social, ethical, philosophical and economic values of society" (Child et al. 2014: 818). Although Anderson proclaimed the end of theory in asserting that "with enough data, the numbers speak for themselves" (Anderson, 2008), big data seem to, thus far, have fostered theory-building (Boellstorff 2015, Tohki & Rauh 2015). In addition to the possibility of asking questions that could not be asked before (and obtaining different answers) (Weinberger 2013, Hand 2016), theoretically untangling the construct of big data is a Herculean task in itself. Big data have long been entangled with society, and we need to cope with that (Floridi 2012) – a discourse that is necessary (Rayport 2011, Barabási 2013), especially considering the omnipresence of big data.

Big data are already living a life of their own. Big data also connect to different flourishing concepts already in existence. This thesis aims at the delimitation of big data from data mining, its connection to algorithms, machine learning, and artificial intelligence. All above concepts are often used synonymously, and are closely intertwined with big data.

2.1.3 Delimitation from Related Terms

2.1.3.1 Data Mining

Big data are often associated with data mining and sometimes both terms are incorrectly used interchangeably. Although both terms deal with data, there are significant differences. Data mining is the computational process of discovering patterns in (large) data sets (Han et al. 2012). The process is often described as knowledge discovery in databases (Fayyad et al. 1996): the use of algorithms, statistical tools, and machine learning in order to extract patterns previously unknown. By identifying clusters, detecting anomalies, locating dependencies, and finding correlations, data mining supports the process of data analysis (Larose 2014).

Data mining is a method or a tool used in handling (big) data. Using data mining merely reveals patterns, and represents only *one* step in the process of knowledge discovery (Fayyad et al. 1996). This embeddedness of data mining into a *bigger* process can be integrated in the steps of knowledge discovery in databases, as proposed by Fayyad et al. (1996: 41):

- Selection
- Preprocessing
- Transformation
- Data mining
- Interpretation/Evaluation
- Knowledge

Data mining means scouring a haystack of data in order to find something other than hay, and maybe the metaphorical needle. It is the process of searching around *existing* data sets and finding information or knowledge that has thus far been unknown. The patterns and signals that may be hidden amidst the noise are what such a system is mining for. Data mining, therefore, does not serve the purpose of collecting data, it merely uses available data. Selecting data and interpreting data does not lie within the realm of data mining. Researchers, thus, point out that the term 'data mining' fails to sufficiently describe the actual process. Han et al. (2012: 6) coin a more suitable term: "knowledge mining from data". Other terms for data mining are knowledge extraction, data/pattern analysis, and data archeology.

Data mining is related to the concept of uncovering patterns without devising preliminary hypotheses (Scholz & Josephy 1984). This explains researchers' tendencies to talk about and deal with data dredging (Selvin & Stuart 1966), data snooping (Sullivan et al. 1999), or spurious correlations (Jackson & Somers 1991). It is

crucial, however, to emphasize that data mining can be part of a big data analysis, even though it marks only one step in the analysis. Data mining on its own leads to several statistical (Hand 1998), as well as ethical issues (Seltzer 2005). Many of those issues, such as overfitting (Elkan 2001), could be tackled by integrating data mining into a holistic big data value chain (Miller & Mork 2013).

2.1.3.2 Algorithms and Machine Learning

In order to describe algorithms, it is essential to understand their relevance. Cormen et al. (2009: xii) clarify that "before there were computers, there were algorithms. But now that there are computers, there are even more algorithms, and algorithms lie at the heart of computing". Due to the exponential growth of digitization and the abundance of data, those algorithms have become more relevant, increasing their influence on society. Beer attributes to algorithms "the capacity to shape social and cultural formations and impact directly on individual lives" (2009: 994). Algorithms influence everybody's everyday life (Pasquale, 2015). Some researchers foresee a future of algorithms "running the world" (Lisi 2015: 23). An algorithm can be defined as follows:

> "Informally, an *algorithm* is any well-defined computational procedure that takes some value, or set of values, as *input* and produces some value, or set of values, as *output*. An algorithm is thus a sequence of computational steps that transform the input into the output" (Cormen et al. 2009: 5).

The connection with big data is evident, as data constitute an algorithm's input and are generated as its output. This makes algorithms a tool to transform data. On the basis of explicit instructions, a computational device follows an algorithm step by step, with a finite amount of input (Boolos & Jeffrey 1974).

The line between algorithms and machine learning is relatively hazy and sparks the question: "Can machines think?" (Turing 1950: 1). Although Turing states that a machine probably cannot think, he suggests that we consider its potential to learn.

> "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E" (Mitchell 1997: 2).

As opposed to humans, a computer program may learn from errors or learn by executing tasks repetitively, but will in fact have "no idea what it's doing" (Schank 2015: 132). Algorithms are incapable of learning how to learn (Argyris & Schön 1996). A machine's learning process can be allocated to two distinct learning scenarios (Mohri et al. 2012). *Supervised learning*, during which a teacher (human) teaches a student (machine) new things, means that training data is available to reveal the instances in which input and output are correctly connected. *Unsupervised learning*, on the other hand, has no sample output data, forcing an algorithm to search for patterns, correlations, or clusters to discover similarities. Such a definition has similarities to that of data mining, despite data mining being more static and using

a finite set of data as well as a strict method of mining. Another type of learning is through the method of reinforcement learning (Sutton & Barto 1998). When learning through reinforcement, there is no knowledge about the correct output. Measuring the correctness of said output, however, becomes possible in interaction with, and through feedback from, the environment.

In his seminal work, Minsky (1961) categorized existing problems for algorithms as search, pattern-recognition, learning, planning, and induction. Minsky reviewed these methods and concluded that they still display many inefficiencies. Kosko (2015) claimed that not much has changed from an algorithmic perspective. He explains that people are still using algorithms that are decades old. Most of the progress achieved in recent years can be attributed to the increase in computational power, as well as to the increase in the amount of data being processed. Similar to the data mining delimitation, algorithms or machines are "not thinking, nor anything like thinking" (Schank 2015: 132). They are, however, now capable of analyzing huge piles of data, albeit still fairly inefficiently, as noted by Minsky (1961). It seems that in the early development of computational algorithms, elements such as speed, compactness, and elegance (Knuth 2011) were admired, while nowadays the use of brute-force (Fellows et al. 2012) and number-crunching (Vaux 2013, Schank 2015) may suffice due to the exponential increase in computational power. Kosko proposes that one of the most prominent algorithms in unsupervised learning is still k-means clustering (in MacQueen 1967). Even though it now carries diverse names, the general idea behind the algorithm remains unchanged. In the context of supervised learning, a popular algorithm is backpropagation (Rumelhart et al. 1986). Kosko concludes that the future means "old algorithms running on faster computers" (2015: 426).

This development may point towards the foreseeable future use of algorithms. Big data, however, have affected the potential of algorithms dramatically. Although most algorithms are based on a simple logic, they exhibit a tendency to become more complex; so complex and advanced, in fact, that the outcome is inscrutable and incomprehensible, even for the engineers behind the respective algorithm (LaFrance 2015). Although this may be bearable, there are several aspects that reveal an underlying problem. (1) People rely *overconfidently* on data (Miller, C. C. 2015). (2) Due to the opaqueness of data, everything that follows an organizing principle may be misinterpreted as transparency (Gillespie 2012). (3) Algorithms maximize myopia due to their focus on solving problems in short- or even real-time (Luca et al. 2016). (4) Algorithms are designed by people. They could implement anything they want in an algorithm and, as stated in (1), understanding these algorithms is becoming increasingly difficult. It is possible to include a certain ideology (Maher 2012) in an algorithm, or opportunities to commit fraud (Parameswaran 2013), and loopholes that allow for "gaming the system" (Rieley 2000).

2.1.3.3 Artificial Intelligence

There are counter-positions to the statement by Kosko and the halt of algorithmic development. Valiant (1984, 2013), for example, proposes the framework of, probably approximately, correct learning. This understanding of the learning process implies that computers may be capable of acquiring knowledge in a similar way to humans and, therefore, "in the absence of explicit programming" (Valiant 1984: 1134). This kind of ability is essential for improving the work between teacher (human) and student (machine) "where humans may be willing to go to great lengths to convey their skills to machines but are frustrated by their inability to articulate the algorithms they themselves use in the practice of the skills" (Valiant 1984: 1142). Another emerging field is deep learning (Arel et al. 2010), as a type of learning inspired by neural networks (Cheng & Titterington 1994). Deep learning deals with "multiple […] layers of nonlinear information processing and methods for supervised or unsupervised learning […] at successively higher, more abstract layers (Deng & Yu 2013: 201). Consequently, improvements in machine learning are strongly connected to the research and development of *artificial intelligence* (Deng & Yu: 2013). The development of artificial intelligence was predicted by Turing (1950: 8): "I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted." The statement may be bold, yet we are already surrounded by a variety of artificial intelligences (e.g. Siri or Cortana).

Generally speaking, artificial intelligence can be defined as the "design of intelligent agents" (Poole et al. 1998: 1). This is a general definition of a broad term, but marks a significant difference to fields concerned with the human mind, like psychology: researching artificial intelligence does not merely involve understanding, but also building artificial intelligences (Russel & Norvig 1995). For that reason, an abundance of more precise definitions can be sorted into the following categories:

- Systems that think like humans
- Systems that think rationally
- Systems that act like humans
- Systems that act rationally (Russel & Norvig 1995: 5)

The two categories linked to humans reveal a difficult bottleneck. The reason for this is that "We don't have sufficient ability to observe ourselves or others to understand directly how our intellects work" (McCarthy 2007: 1175). As a result, such definitions are often used when approaching the issue from a theoretical perspective.

Artificial intelligence can, therefore, currently be codified into three types, the first being *weak* (Searle 1980) or *narrow AI* (Hutter 2009). Intelligence of this type specializes in one specific area and is only capable of performing well in this field. The chess machine Deep Blue, while capable of beating humans at chess, is not able to do much else. In today's world, we are surrounded by such AI, for example spam filters, Google Translate, autopilot, self-driving cars, and so on. The second type of AI is the *strong artificial intelligence* (Searle 1980), sometimes called *full* (Bainbridge

2006), *human-level* (Nilsson 2005), or artificial *general* intelligence (Voss 2007). This type of AI refers to the idea that such an intelligence could perform any task. A machine of that kind is intelligent beyond a narrow spectrum and no longer specialized in only one specific field, thus capable of many tasks. Such an AI would pass the Turing Test (Turing 1950) and the Chinese Room Test (Searle 1980) and would be indistinguishable from a human. Finally, there is the third type of *superintelligence*. As yet, this type of AI is merely a theoretical mind game and a popular theme in science fiction. The very idea, in fact, connects artificial intelligence and its potential for human extinction (Barrat 2013). Bostrom defines it the following way:

> "By a 'superintelligence' we mean an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills. This definition leaves open how the superintelligence is implemented: it could be a digital computer, an ensemble of networked computers, cultured cortical tissue or what have you. It also leaves open whether the superintelligence is conscious and has subjective experiences" (Bostrom 2006: 11).

In addition to potential dystopian consequences, this type of artificial intelligence sparks an interesting discussion: would we even be able to understand an artificial superintelligence? In today's world, artificial intelligence predominantly mimics intelligence (Munakata 1998) without actually understanding (Hearst 2015). As McCarthy states, "Much of the public recognition of AI has been for programs with *a little bit of AI and a lot of computing*" (2007: 1175). Consequently, the development of artificial intelligence has not come as far as some people believe (Dreyfus 1965, Hopgood 2003, Epstein 2015). Others have identified AI as a threat to humanity (Hawking et al. 2014). Artificial intelligence is still predominantly human (de Biase 2015), due to the fact that AI is initially programmed by humans. For that reason, Dobelli (2015) refers to artificial intelligence as *humanoid thinking*. Any artificial intelligence will be restricted by a certain humanoid framework. There may also be the possibility of *alien thinking* of sorts. Such thinking will differ greatly from anything we know. This type of thinking, however, requires its own evolutionary path, "not just evolutionary algorithms" (Dobelli 2015: 99). Alien thinking, while imaginable, will therefore need some time to evolve. Artificial intelligence may be seen as something alien when humans no longer understand the underlying algorithms. For the sake of a clearer distinction, Kosslyn (2015) recognized the difference as close AI and far AI. Clark (2015) replies that, although artificial intelligence exposed to big data and deep learning will lead to knowledge that seems opaque, it will "end up thinking in ways recognizably human" (Clark 2015: 156). As a result, current artificial intelligence depends heavily on big data, without which many current systems would perform insufficiently. Any current self-driving car depends on data, be it generated by the car itself through sensors, or data from other sources. Google Translate is an example of big data rather than a weak AI system.

2.1.4 Big Data Pitfalls

In today's society, many people attempt to describe big data and outline their potential. Big data have been placed on a pedestal as being something unique and precious (Mayer-Schönberger & Cukier 2013). Today's discourse of big data makes it appear to be the solution to all problems of society (Steadman 2013), and capable of making the world a safer and better place (Olavsrud 2014). Inherent in this discourse is the belief that making something more data-based or data-driven will lead to more objective and considered decisions (McAfee & Brynjolfsson 2012), however, generating and using big data is a process that is not as clean or sterile as some researchers suggest. Many believe in "objective quantification" (van Dijck, 2014: 198), but big data are "messy" (Harford 2014: 14) and even just the collection of data causes a manipulation or "preconfiguration of data" (van Dijck & Poell 2013: 10). For that reason, ascribing an "aura of truth, objectivity, and accuracy" to big data would be a fallacy (Boyd & Crawford 2012: 664). Nevertheless, literature links big data to a variety of pitfalls. In the related literature, many researchers often refer to three papers (Boyd & Crawford 2012, Richards & King 2013, Dalton & Thatcher 2014). These papers present a systematization of obstacles in the usage of big data. In the following I will present them and explain why I chose Boyd and Crawford's (2012) systematization.

In their 2012 article, Boyd and Crawford develop six aspects of potential pitfalls of big data:

1. Big data change the definition of knowledge
2. Claims of objectivity and accuracy are misleading
3. Bigger data are not always better data
4. Taken out of context, big data lose their meaning
5. Accessibility does not make them ethical
6. Limited access to big data creates new digital divides

Dalton and Thatcher (2014) postulate a different systematization. They do, however, discuss big data at a more philosophical level and contribute to the meta-level discourse about big data. Big data are a highly social entity, which is why the authors focus on the lack of objectivity in big data:

1. Situating 'big data' in time and space
2. Technology is never as neutral as it appears
3. 'Big data' do not determine social forms: confronting hard technological determinism
4. Data are never raw
5. Big isn't everything
6. Counter-data exist

On the other hand, Richards and King (2013) deconstruct big data by explaining that big data always come with a tradeoff. Although big data contribute to transparency, identity and power equality in certain ways, big data increase opaqueness,

anonymity and power inequality in other ways. Richards and King, therefore, postulate the following paradoxes:

1. The transparency paradox
2. The identity paradox
3. The power paradox

There are several other papers (e.g. Mittelstadt & Floridi 2015, Saqib et al. 2015, Hilbert 2016) that tackle challenges related to big data, but most are relatively congruent and point out similar pitfalls.

In the following description of big data pitfalls, I will use the six aspects introduced by Boyd and Crawford (2012), because they are broader than the paradoxes by Richards and King (2013) and more precise than the systematization given by Dalton and Thatcher (2014). Boyd and Crawford (2012) seem to cover all relevant pitfalls of big data and, furthermore, they are giving a precise description of those pitfalls.

2.1.4.1 Big Data Change the Definition of Knowledge

Boyd and Crawford (2012) claim that big data will fundamentally change the way we view the working world and the production process. They compare the situation to Fordism (Amin 1994), which is closely linked to the theory of Taylorism (Taylor 1911), which dehumanized work in the early 20th century and only focused on the mechanic parts and automation involved in the work process. The changes that accompanied mass production extensively transformed the relationship between work and society. Computerization and digitization (Zuboff 1988, 2014) are also currently in the process of changing this relationship radically, and trends like automation may have ground-breaking consequences for the labor market (Frey & Osborne 2013).

Big data may contribute to those changes (Davenport 2014), but predominantly change the way people think. Boyd and Crawford (2010: 153) cite Latour as follows: "Change the instruments, and you will change the entire social theory that goes with them." Puschmann and Burgess figuratively describe this change: "before, we were starved for data; now we are drowning in it" (2014: 8). The abundance of data has changed the perception of knowledge drastically. Big data allow for researchers and practitioners to access information in real-time, and enable them "to collect and analyse data with an unprecedented breadth and depth and scale" (Lazer et al. 2009: 722). Big data have led to a supposedly epistemological paradigm shift (Puschmann & Burgess 2014). The computational turn in knowledge (Thatcher 2014) and the proposed fourth paradigm of data-intensive science discovery (Hey et al. 2009) have led to an environment in which "gather data first, produce hypotheses later" (Servick 2015: 493) represents an approved research approach. Kitchin describes these new data analytics as "seek(ing) to gain insights 'born from the data'" (2014b: 2). Following this argument, Anderson (2008) almost polemically proclaimed the end of theory:

> "This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves" (Anderson 2008).

Consequently, there is no need to construct hypotheses or conduct experiments (Prensky 2009). Formerly unknown patterns will be discovered (Dyche 2012), or, in other words, correlation trumps causation (Lycett 2013). Kitchin summarizes this argumentation the following way:

1. "Big Data can capture the whole of a domain and provide full resolution;
2. there is no need for a priori theory, models or hypotheses;
3. through the application of agnostic data analytics, the data can speak for themselves free of human bias or framing, and that any patterns and relationships within Big Data are inherently meaningful and truthful;
4. meaning transcends context or domain-specific knowledge, thus can be interpreted by anyone who can decode a statistic or data visualization" (Kitchin 2014b: 4).

Such a description prompts a scientific déjà vu of the statistical beliefs of empiricism and positivism. Comte's formula in particular, "Savoir pour prévoir, prévoir pour pouvoir" (Comte cited in Merton 1936: 898), translated into "to know in order to predict, to predict in order to control" (Clarke 1981: 90), can be linked to these beliefs concerning big data. Kitchin, however, characterizes this particular empiricist view as "fallacious thinking" (Kitchin 2014b: 4) and points out the limitations of data-driven decisions (Lohr 2012). Frické (2014) used the words of Popper (1963) to emphasize that although data can now be collected easily and in a more granular fashion than ever before, big data will not translate into useful observations by means of big data:

> "The belief that we can start with pure observations alone, without anything in the nature of a theory, is absurd; as may be illustrated by the story of the man who dedicated his life to natural science, wrote down everything he could observe, and bequeathed his priceless collection of observations to the Royal Society as inductive evidence. This story should show us that though beetles may profitably be collected, observations may not" (Popper 1963: 478).

Claims such as that there is "no need for a priori theory" are being strongly debated, for example by Frické (2014) who defines data as, per se, entangled with a priori assumptions of sorts. Knowledge is generated by means of big data in various ways and there may be an end to a certain type of theory, but big data calls for entirely new types of theories (Boellstorff 2015, Tokhi & Rauh 2015). Generating hypotheses and then collecting data may no longer be feasible, since the required data are already available (Frické 2015). In the conclusion of his paper, Frické counters Andersons's claim, demanding focus on thorough and precise scientific work rather than deleting theory and using only big data:

> "The ability to cheaply and easily gather large amounts of data does have advantages: Sample sizes can be larger, testing of theories can be better, there can be continuous assessment, and so on. But data-driven science, the 'fourth paradigm', is a chimera. Science needs problems, thoughts, theories, and designed experiments. If anything, science needs more theories and less data" (Frické 2015: 660).

Consequently, big data can be seen as requiring a meta-theory which consists of various theories and is, therefore, steeped in theories and methods (Mayer-Schönberger 2014). Big data are both "theory-laden" and "tainted by theory" (Frické 2014: 652). Big data lead to an abundance of data and, therefore, more information. This is what may make distilling knowledge from big data substantially more difficult (Jacobs 2009), and calls for a focus on scientific rationale of methodological rigor (Tokhi & Rauh 2015) and awareness (Ruths & Pfeffer 2014), two variables more strongly demanded than ever before (Lazer et al. 2014).

2.1.4.2 Claims of Objectivity and Accuracy Are Misleading

The basis of numerous approaches favoring big data is the underlying claim that, due to the mere abundance of data, big data are highly objective (Lukoianova & Rubin 2014) and, due to the variety of sources, more accurate (McAfee & Brynjolfsson 2012). People believe overconfidently that data-driven decisions are superior due to their objectivity (Miller, C. C. 2015).

> "A decision made by the numbers (or by explicit rules of some other sort) has at least the appearance of being fair and impersonal. Scientific objectivity thus provides an answer to a moral demand for impartiality and fairness" (Porter 1996: 8).

This claim makes sense at first sight, but big data are, in fact, highly subjective (Dalton & Thatcher 2014). Some, however, attribute objectivity to big data as a result of technical dispositions. After all, data are collected from sensors, saved into log-files, and processed by computers. Their objectivity, thus, results from their mechanical nature. For the sake of the argument, let us *assume* there is such a thing as *truly objective* big data. The critical problem here is that big data are not self-explanatory (Bollier 2010). There is an essential need for somebody to interpret this certain data set in order to extract any knowledge from it. Said interpretation could be by a human or by a machine. Human and machine often collaborate in some way, but one party always leads the interpretation. For that reason, any interpretation is influenced by either a human or a machine. A set of big data will be manipulated and transformed in one way or another, which renders it less than objective (Dalton & Thatcher 2014). Bollier questions the objective truth of big data, because any interaction with data will lead to subjective contamination: "Can the data represent an 'objective truth' or is any interpretation necessarily biased by some subjective filter or the way that data is 'cleaned'?" (2010: 13). Therefore, Metcalf and Crawford (2016) call for a theory on data subjectivity.

A closer look at the human interpreter reveals several interpretation biases. "Data are perceived and interpreted in terms of the individual perceiver's own needs, own connotations, own personality, own previously formed cognitive patterns" (Krech & Crutchfield 1948: 94). The following statements are similar arguments: "Disciplines operate according to shared norms" (Gitelman & Jackson 2013: 7). "Individuals construct their own subjective social reality based on their perception of the input" (Bless et al. 2004: 2). In a recent experiment, Silberzahn and Uhlmann (2015) gave 29 research teams the same data set and let them analyze it. Interestingly, the results varied greatly and the authors attributed this variance to the fact that "any single team's results are strongly influenced by subjective choices during the analysis phase" (Silberzahn & Uhlmann 2015: 191). Although it is unclear what influenced the researchers, whether it was their own preconceptions or the data set itself (Griffiths 2015), it seems as though the process of data interpretation itself is "inherently subjective" (Boyd & Crawford 2012: 667). Decisions are based on subjective perception and, therefore, deviate from rational decisions (Tversky & Kahneman 1974). Such biases (Kahneman & Tversky 1973) permeate any decision and, therefore, will influence data interpretation. Arnott (2006) identified more than 37 cognitive biases in the literature and Yudkowsky (2008) reports twelve cognitive biases (only five overlaps). Consequently, the lists are less than comprehensive and, especially in the context of data interpretation, far from exhaustive. For Table 4, I selected the most relevant biases from the work of Arnott (2006) and Yudkowsky (2008) and added further biases that fit the context of data interpretation. These authors used different sources for the definitions of the types of biases, due to the reason that these definitions are more precise and more concise.

Table 4: Selection of Cognitive Biases

Types of Bias	**Definition**
Hindsight Bias (Fischhoff & Beyth 1975)	"… refers to people's tendency to alter their perception of the inevitability of an event once they know the outcome of the event" (Christensen-Szalanski & Willham 1991: 147).
Correlation Bias (Tversky & Kahneman 1973)	"The subjects markedly overestimated the frequency of co-occurrence of natural associates, such as suspiciousness and peculiar eyes. […] In their erroneous judgments of the data to which they had been exposed" (Tversky & Kahneman 1974: 1128).
Confirmation Bias (Wason 1960)	"… means that information is searched for, interpreted, and remembered in such a way that it systematically impedes the possibility that the hypothesis is rejected – that is, it fosters the immunity of the hypothesis" (Oswald & Grosjean 2004: 79).

Types of Bias	Definition
Overconfidence Bias (Adams & Adams 1960)	There are "3 distinct ways in which the research literature has defined overconfidence: (a) overestimation of one's actual performance, (b) overplacement of one's performance relative to others, and (c) excessive precision in one's beliefs" (Moore & Healy 2008: 502).
Apophenia (Conrad 1958)	"… the tendency to find meaningful patterns in meaningless noise" (Shermer 2008).
Base Rate Fallacy (Meehl & Rosen 1955)	"The base-rate fallacy is people's tendency to ignore base rates in favor of, e.g., individuating information (when such is available), rather than integrate the two" (Bar-Hillel 1980).
Naïve realism (Ross & Ward, 1996)	"People think, or simply assume without giving the matter any thought at all that their own take on the world enjoys particular authenticity and will be shared by other open-minded perceivers and seekers of truth" (Pronin et al. 2002: 369).

(on the Basis of the Lists of Biases by Arnott 2006: 60–61 and Yudkowsky 2008: 91–119)

Although this is only a sample of potential cognitive biases, those biases reveal the subjectivity of the interpreter. There may be several ways to mitigate those cognitive biases in order to achieve more rational decisions (Burke 2007), but a machine might be far superior to a human in terms of subjectivity, especially since a machine, algorithm, or artificial intelligence, is supposedly more rational than a human. A machine may be a superior interpreter. This is, however, not the case, since "there is no automatic technique for turning correlation into causation" (Spiegelhalter 2014: 264).

Statistics are, generally speaking, a process that takes place post hoc (Frické 2015). There are several statistical biases that potentially distort any data set in one direction or another. One strong distortion factor can be attributed to the p-values in statistical findings. Significant results are essential for publishing research findings (Vidgen & Yasseri 2016).

> "P-values are used in Null Hypothesis Significance Testing (NHST) to decide whether to accept or reject a null hypothesis (that there is no underlying relationship […] between two variables)" (Vidgen & Yasseri 2016: 1).

Common understanding is that p-values lower than .05 are sufficient to classify findings as statistically significant (Ioannidis 2005). Taking that into account, there is no complete certainty as to whether or not hypotheses on the basis of statistical data will be rejected. There is a chance that a hypothesis could be falsely rejected (type I error) or falsely not rejected (type II error), as shown in Table 5.

Table 5: Type I Errors and Type II Errors

		Null Hypothesis (H0) is	
		True	False
Judgment of Null Hypothesis (H0)	Reject	Type I Error (False Positive)	Correct Inference (True Positive)
	Fail to Reject	Correct Inference (True Negative)	Type II Error (False Negative)

(Sheskin 2004: 54)

Although the threshold value of .05 seems rigorous, there is a chance that out of 100 independent tests, approximately five tests are either false positives or false negatives (Frické 2015). Such an example simplifies the problem too drastically. The type of error, however, occurs commonly in statistical findings (Ioannidis 2005). Although data "speak for themselves" (Anderson 2008), the results derived are erroneous. Even the most rigorous data analyses are susceptible to type I and type II errors, and these errors have consequences, as Ioannidis (2005: 696), well documented, describes: "It can be proven that most claimed research findings are false." According to this, we discover potential patterns where there are none and we overlook others (Shermer 2008). Those correlations can be attributed to logical and comprehensible explanations and lead to spurious correlations (Jackson & Somers 1991). Big data intensify those problems due to their nature, in the sense that there are many observations (n) and an even larger number of parameters (p) which aggravate these obstacles.

> "Big Data means that we can get more precise answers; this is what Bernoulli proved when he showed how the variability in an estimate goes down as the sample size increases. But this apparent precision will delude us if issues such as selection bias, regression to the mean, multiple testing, and overinterpretation of associations as causation are not properly taken into account. As data sets get larger, these problems get worse, because the complexity and number of potential false findings grow exponentially. Serious statistical skill is required to avoid being misled" (Spiegelhalter 2014: 265).

There are also several other error sources, such as sub-setting, overfitting, stepwise regression, univariate screening, and dichotomizing continuous variables that could lead to data-driven "hornswoggling" (Frické 2015: 657).

> "'Correlation is enough'. We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot… With enough data, the numbers speak for themselves" (Anderson 2008).

Consequently, Anderson's proposition is wrong.

Finally, data interpretation seems to be highly subjective, regardless of the type of interpreter. The interpreter, however, is not the only source of subjectivity. At this

point, the initial claim of truly objective data proves to be non-maintainable. Even with more data, granular observation, and further parameters, sampling biases remain inherent to any set of data (Crawford, 2013). "Indeed, all data provide oligoptic views of the world: views from certain vantage points, using particular tools, rather than an all-seeing, infallible God's eye view" (Kitchin 2014b: 4). In conclusion, any researcher "can be too influenced by preconceptions, or too influenced by the data" (Griffiths 2015: 141). Any type of data, however, is subjective and big data especially are often collected with a certain intent and often repurposed for different goals (Schneier 2015). Consequently, from the point of data creation onwards, any data exhibit a hereditary subjectivity, and removing this distortion resolves around a "careful application of statistical science" (Spiegelhalter 2014: 265).

2.1.4.3 Bigger Data Are Not Always Better Data

Another claim commonly propagated in the field of big data research is that big data are interchangeable with whole data (Boyd, & Crawford, 2012), or that big data are *n = all* (Amoore & Piotukh 2015). Mayer-Schönberger told Harford (2014: 17) in a personal discussion the following:

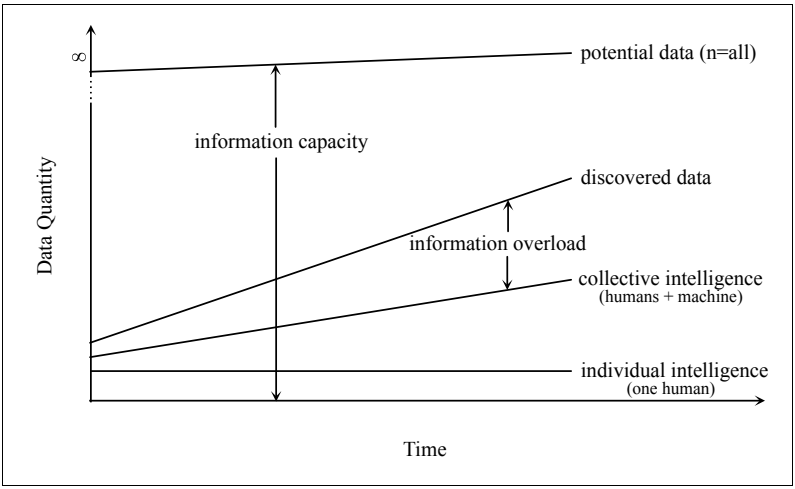
> "A big data set is one where "n = all" – where we no longer have to sample, but we have the entire background population. […] And when "N = All" there is indeed no issue of sampling bias because the sample includes everyone."

Claiming n = all seems like an audacious assumption and many researchers reject the thesis of big data as not being merely a sample but being a complete set (e.g. Boyd & Crawford 2012, Bowker 2014, Lagoze 2014, Amoore & Piotukh 2015, Tokhi & Rauh 2015, Hand 2016). Boyd and Crawford (2012) refute this assumption using the example of Twitter, which is often used as a source representative of "all people" (2012: 669). In accordance with this argument, Leonelli notes that "having a lot of data is not the same as having all of them; and cultivating such a vision of completeness is a very risky and potentially misleading strategy" (Leonelli 2014: 7).

Following the analogy that big data are n = all, humans are, metaphorically speaking, surrounded by data. "The world and its universe are, to anything or anyone with senses, incomprehensibly big data" (Andrejevic 2014: 1675). It is however not possible for anybody to access all the data. An intelligence of that kind would be comparable to Laplace's daemon:

> "We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes" (Laplace 1951: 4).

Figure 2: Conceptual Evolution of Data over Time (Scholz 2015a)



As shown in Figure 2, data are discovered at staggering speed (Miller 2010). Data are also becoming more granular, more singular, and more heterogeneous (Kucklick 2014). Any individual would be overburdened by this deluge of data (Anderson 2008) and could only grasp a small portion of all data. Data absorption can be increased by means of collaboration with other humans, as well as with machines. This is often referred to as collective intelligence (Bonabeau 2009). Deriving from the former statement that computational power is growing according to Moore's law and the acquisition of data, here data discovery follows Kryder's law: the discrepancy between discovered data and collective intelligence has been increasing. This information overload (Eppler & Mengis 2004) cannot be conquered even with the help of technology. Following the description of wholeness of potential data, it seems logical to assume that potential data cannot be exhausted at any time. This assumption becomes clear in the phenomenon that is known as the "complexity barrier" (Gros 2012: 183). The complexity barrier describes the situation wherein the effort of gaining scientific insight from a research field rises exponentially when approaching a certain threshold of complexity. If 'whole data' stands for the intellect of Laplace's daemon, then there may be data out there that are out of reach for any intellect other than such a daemon. The completeness of whole data is, therefore, not achievable, and consequently "big data and whole data are not the same" (Boyd & Crawford 2012: 669). Harford concludes with the following warning: "Found data contain systematic biases and it takes careful thought to spot and correct for those biases. 'n = all' is often a seductive illusion" (2014: 18). Big data will not lead to the theory of everything (Mainzer 2014).

2.1.4.4 Taken out of Context, Big Data Lose Their Meaning

Data are collected for a certain purpose, from a distinct vantage point, with distinct methods, and various tools (Kitchin 2014b). "However, in both its production and interpretation, all data – 'big' included – is always the result of contingent and contested social practices that afford and obfuscate specific understandings of the world" (Dalton, & Thatcher 2014). Several researchers conclude that raw data do not exist. In her edited book, Lisa Gitelman unmasks "'raw data' as an oxymoron" (Gitelman 2013). Bowker then explains that claim even further, judging that "'raw data' is a bad idea" (Bowker, 2005: 184). Data, and especially big data, are contextualized by the process of data generation (Kitchin 2014c) and, although the data set seems objective, distortion has already taken place in the process of collecting specific data. Information about the organization of data collection is essential (Lynch 2008), and, therefore, Maturana proposes that "anything said is said by an observer" (1970: 4). Applied to big data, the sentence could be transformed into: any data collected are collected by an observer. Von Foerster also introduced a variation to this assumption: "Anything said is said to an observer" (1979: 5) which could be adapted for big data as well: any data collected are collected for an observer. Data appear to always be firmly entangled with the entity that collects them. George et al. (2014: 322) categorize the sources of big data into the following five types:

- public data (collected by governments or governmental organizations)
- private data (collected by private-firms or individuals)
- data exhaust (passively collected and collected for a different purpose)
- community data (social media data)
- self-quantification data (collected by individuals to monitor or track themselves)

It seems obvious that there are reasons for collecting certain data, depending on the data collector and observer, but a certain tendency can be observed wherein big data "is […] not used for the purpose for which it was collected" (Puschmann & Burgess 2014: 1699) and repurposing data has become common practice (Schneier 2015). The context of data collection is also often lost in transfer (Pasquale 2015), due to various types of invisible "access constraints and platform-side filtering" (Ruths & Pfeffer 2014: 1063). This fosters "the need for increased awareness of what is actually being analyzed" (Ruths & Pfeffer 2014: 1064). There is evidence of a self-selection bias, especially in the context of social media (Schoen et al. 2013). Taking Twitter social media data out of context, therefore, means assuming that Twitter generates a representative sample. This claim is flawed (Tumasjan et al. 2010) however, generating distorted results (Ruths & Pfeffer 2014).

In his reply, Seaver (2015) tackles the claim originally brought forward by Boyd and Crawford (2012) that big data can be completely taken out of context, and states that context is king, context is key, context is questioned, context is constructed, and context is contested. He explains that context derives current business, and emphasizes that context-aware systems do exist and actually work quite well. In his discussion, however, he predominantly focuses on the inadequacies of the term

'context' which is subject to constant debate (e.g. Dilley 1999, Johns 2006). In the context of big data, the term may lack precision. Boyd and Crawford (2012), and in particular Dalton and Thatcher (2014), Kitchin (2014b), and Pasquale (2015), highlight the problem that the transfer of data inevitably causes a loss of information. How were the data derived? Who collected the data? For what purpose were the data generated? These questions may sound like contextual information, but they also resemble the definition of metadata (or more precisely extended metadata).

> "To any data element, or to any of the component cells of a composite, can be associated, in a binary relationship, certain data elements which represent data 'about' the related element. We refer to such data as 'metadata' and call the relationship one of 'secondary association'" (Bagley 1968: 91).

For (extended) metadata, it will necessarily contain all the information about the observer and the methodological constraints that lead to a distortion of big data. These contextual aspects will be included in any data set in order to incorporate the influences brought by the data collection into other contexts. Losing, or getting rid of, this contextual metadata will change the data and will lead to a substantial increase in the number of statistical traps and obstacles.

2.1.4.5 Accessibility Does Not Make Them Ethical

Big data have led to an abundance of data available to anybody and obtainable without great effort (Fanning & Centers 2013). Many companies also act as data brokers (Otto et al. 2007) and sell a variety of data at low cost. People reveal more and more information about themselves (Enserink & Chin 2015), especially due to the recent trend of quantifying the self as a process of self-tracking aspects, such as running habits (Ruckenstein & Pantzar 2015). Aggravating the effect of the masses of data makes people slaves to their habits, and further facilitates their identification (Eagle & Pentland 2006). Some researchers (e.g. Tene & Polonetsky 2012) view big data as a contributor to the anonymization of the individual due to the sheer mass of data. The majority (e.g. Ohm 2010, Richards & King 2013, de Montjoye et al. 2015, Schneier 2015), however, would much rather discuss the "myth of anonymization" (Clemons 2013). Several companies have evolved, over time, to become proper "Datenkraken" (Bager 2006: 168), a German compound that translates to 'data kraken'. These companies acquire data from all available sources and use them in order to paint a granular picture of any person (Kucklick 2014). To make matters worse, every person leaves behind a trail of data. This is why privacy alone is no longer enough (Matzner 2014), because "privacy as we have known it is ending, and we're only beginning to fathom the consequences" (Enserink & Chin 2015: 491), especially as "any information that distinguishes one person from another can be used for re-identifying data" (Narayanan & Shmatikov 2010: 24). Schneier, an established security expert, takes a fairly drastic approach to demystifying this belief:

> "It's counterintuitive, but it takes less data to uniquely identify us than we think. […] We can be uniquely identified by our relationships. It's quite obvious that you can be uniquely identified by your location data. With 24/7 location data from your cell phone, your name can be uncovered without too much trouble. You don't even need all that data; 95% of Americans can be identified *by name* from your four time/date/location points" (2015: 44).

Technological advancement will foster this development even further. One example is the location data generated by smartphones: due to the intimate relationship of humans with their smartphone (González, et al. 2008), people are now constantly accompanied by their phones. In consequence, data about the whereabouts of every smartphone user already exist and anonymizing it by removing personally identifiable information is absolutely insufficient. The amount of information that can be derived from meta data is even more shattering (Schneier 2015). Michael Hayden, former NSA and CIA director, was recently incensed enough to say that "we kill people based on metadata" (2014).

Big data have decreased and eradicated anonymity (Tucker 2013), and will be involved in the creation of a "goldfish bowl society" (Froomkin 2015: 130). Ethical use will be the most critical topic concerning big data, especially with regard to the numerous examples of the de-anonymization of large data sets. Two cases have been discussed in the literature in greater detail (e.g. Ohm 2010, Schneier 2015). AOL published search queries in 2006, and researchers were able to identify AOL users on the basis of this set of data (Barbaro & Zeller 2006). Netflix published their users' movie rankings, and researchers were able to de-anonymize people by comparing the data set with the Internet Movie Database (IMDb) (Narayanan & Shmatikov 2008).

The question is, therefore, no longer whether de-anonymizing big data sets is possible, but whether or not it is done. There are enough data available to conduct any analysis imaginable, but the ethical perspective has become prevalent. How do researchers (or big data analysts) act ethically (Boyd & Crawford 2012)? If big data are far from being anonymous, nobody can pledge privacy of data, but merely promise to use them in an ethical way. Mittelstadt and Floridi summarize this challenge in the context of biomedicine, but their statement holds true for any other discipline: "Data have been identified as particularly ethically challenging due to the sensitivity of health data and fiduciary nature of healthcare" (2015: 28) and, additionally, lead to "ethical responsibility development, deployment and maintenance of novel data sets and practices in biomedicine and beyond in the era of Big Data".

2.1.4.6 Limited Access to Big Data Creates New Digital Divides

If big data are really that ubiquitous and anonymization is not entirely possible, having access to data becomes a source of power. Andrejevic calls this phenomenon "the big data divide" (2014: 1673). Although everybody generates data, there are many with limited or no access to data at all. Manovich (2011) classifies these people into

three stages of big data involvement: data creator, data collector and data analyzer. The difference in access, however, leads to "asymmetric sorting processes and different ways of thinking about how data relate to knowledge and its application" (Andrejevic 2014: 1676). Especially in the context of predictive analytics (Shmueli & Koppius 2011), people or algorithms with access to data can potentially influence those without access to such data (Palmas 2011).

> "When you are doing this kind of analytics, which is called 'big data', you are looking at hundreds of thousands to millions of people, and you are converging against the mean. I can't tell you what one shopper is going to do, but I can tell you with 90 percent accuracy what one shopper is going to do if he or she looks exactly like one million other shoppers" (Nolan 2012: 15).

Such processes lead to a form of social sorting (Lyon 2003) and, even worse, a sorting of people without access to data by people with access. Such presumed knowledge of the future can create self-fulfilling prophecies (Merton 1948), especially when people are nudged in a certain direction (Thaler & Sunstein 2008). Being able to execute this form of steering is a new and weighty source of power. Unfortunately, assessments of this kind, deliberate or accidental, are quickly written in stone:

> "For example, one data broker (ChoicePoint) incorrectly reported a criminal charge of 'intent to sell and manufacture methamphetamines' in Arkansas resident Catherine Taylor's file. The free-floating lie ensured rapid rejection of her job applications. She couldn't obtain credit to buy a dishwasher. Once notified of the error, ChoicePoint corrected it, but the other companies to whom ChoicePoint had sold Taylor's file did not necessarily follow suit. Some corrected their reports in a timely manner, but Taylor had to repeatedly nag many others, and ended up suing one" (Pasquale 2015: 33).

This shift in power means many people are growing more and more powerless when it comes to their data. They do not know what is collected, who is collecting, and for what purpose. There will be winners and losers (Richards & King 2013), there will be people who are empowered or disempowered (Mansell 2016). Current developments in big data increase the complexity as well as the opacity of the system (Burrell 2016) and will intensify this divide even further.

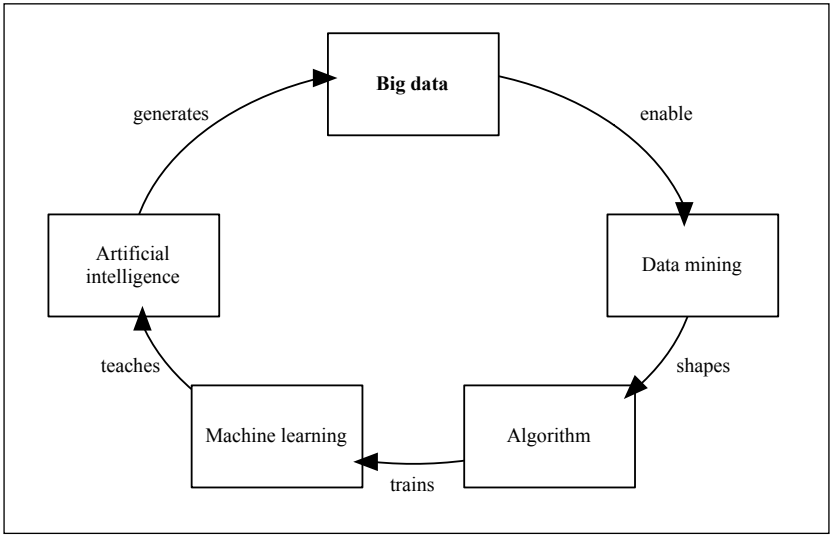
2.1.5 May Big Data Be with You

In summary, big data cannot be categorized into existing technological dimensions like data mining, algorithms and machine learning, or artificial intelligence. Big data are interconnected with those technologies and take a new form during this process. As artificial intelligence becomes smarter, more autonomous and opaque, big data are transformed in novel ways. Without big data and the abundance of data available, none of the current improvements in technology would be possible. IBM's artificial intelligence Watson, for example, is constantly learning from the internet (Madrigal 2013) and Google's artificial neural networks are capable of dreaming about their experiences within the internet (Mordvintsev et al. 2015). Those artificial

intelligences, therefore, generate data on their own and could potentially learn from themselves. For that reason, disentangling big data from this technological cycle is as impossible as it is unnecessary. Big data have led to a quantum leap in those fields and, although those technologies merely mimic intelligence, big data make them smarter.

Big data are entangled in a complex way with data mining, algorithms and machine learning, and artificial intelligence. Big data enable those technologies to be better (Lohr 2012). On the other hand, big data are enabled by these technologies (O'Leary 2013). Big data contribute to a cycle of technology and can be depicted as in Figure 3.

Figure 3: Big Data's Technology Cycle



Consequently, big data are difficult to grasp, which renders defining big data a complicated task. The term 'big data' may not be sufficient for describing the phenomenon (Manovich 2011). "Big data is less about data that is big than it is about a capacity to search, aggregate, and cross-reference large data sets" (Boyd & Crawford 2012: 663). Big data, therefore, better represent a new grasp of data. Big data as a *new understanding of data* already have and will have in the future an unarguably significant impact on society. Digitization and technological progress are predominantly driven by data. Artificial intelligences work because of big data and mimic real intelligence quite well. Big data act as a lubricant for any advance in modern society. Big data are the basis of and a resource for modern communication (be it human to human, machine to machine, or human to machine) and transform the entire communicative process. Big data are "already occupying a huge place in the

landscape of what technology is, what it might offer, and what it could be" (Bell 2015: 9). One could deduce that big data are everywhere (Cukier 2013). Big data have outgrown their lion's cage (Dietsch 2010). They float freely within the realm of both digital and analog worlds. The barriers between both worlds are also blurry and seem likely to merge in the future (Mayer-Schönberger & Cukier 2013, Kucklick 2014, Pasquale 2015). Big data will potentially contribute significantly to the theory of everything (Mainzer 2014). Metaphorically speaking, big data is a theoretical construct that surrounds everything, everywhere, and all the time. Big data are more than the *invisible hand* (Smith 1776) and sound similar to the description of 'the Force' in the Star Wars universe.

> "Well, the Force is what gives a Jedi his power. It's an energy field created by all living things. It surrounds us and penetrates us; it binds the galaxy together" (Obi-Wan Kenobi in Star Wars – A New Hope).
>
> "For my ally is the Force, and a powerful ally it is. Life creates it, makes it grow. Its energy surrounds us and binds us" (Yoda in Star Wars – The Empire Strikes Back).

In the context of big data, this sentence could be transformed into:

> Well, big data are what gives a human power. It's a data field created by all living things. It surrounds us and penetrates us; it binds the world together.
>
> For my allies are big data, and powerful allies they are. Life creates them, makes them grow. Their data surround us and bind us.

Although this metaphor may be a bit farfetched, it is crucial to state that big data are omnipresent, and many worldly things are no longer possible without *tapping into the realm of big data*. Big data are "here to stay" (Newell & Marabelli 2015: 10), "too big to ignore" (Simon 2013: xxi), and most importantly "here to grow" (Floridi 2012: 437). Contrary to the Star Wars analogy, however, where not all people can become Force-sensitive and use the Force, people can become big data-sensitive and use big data to their purpose. Big data in general are initially neutral, but as Gitelman (2013) asserts: "Raw data is an oxymoron". Data are shaped by a data-sensitive entity (human or machine), thus following the logic of Kranzberg (1986: 545): "Technology is neither good, nor bad; nor is it neutral" because "technology is a very human activity" (Kranzberg 1986: 557), however, big data "act as intermediaries in almost every aspect of our existence" (Seife 2015: 480). Due to their influence, big data cannot be seen as something purely technological, even though that perspective is already a difficult one. Beyond that, they are heavily entangled with today's society and, therefore, are understood as a socio-technological phenomenon. It is for that reason that "We must look beyond findings from previous studies of emerging technologies because new technologies often help create new socio-technological contexts" (Michael & Michael 2013: 26).

2.2 Big Data at the Socio-Technological Level

2.2.1 Technology and Society

Big data are highly entangled with society (Mayer-Schönberger & Cukier 2013). Society deals with big data on a daily basis and the influence of this datafication is growing exponentially. Technologies driven by big data are growing in significance. One example is how people find their way from A to B. GPS-powered smart-phone applications or dedicated devices have largely replaced printed maps.

> "It is easy to dismiss technology as a mere object, without giving much consideration to how it is woven into our everyday life. Technology provides a means for us, its users, to get things done. Without giving it much thought, we leave our homes every morning with our cellphones (often more than one), laptops, MP3 players (such as iPods), headphones, watches, and other gadgets. Only when our technology fails us we suddenly realize the depth of our dependence on that technology" (Quan-Haase 2016: 1).

For that reason, technology is more than a mere tool designed by society, instead it shapes and changes it. "We are no longer looking at just a 'technology' and its 'users' but the event of their relationships, of their reciprocal configuration" (Giddings 2006: 160). Technology challenges beliefs and structures (Heidegger 2011). Technology and society influence each other reciprocally (MacKenzie & Wajcman 1985). Both humans and technology are relational to each other as "equal" objects (Bryant, L. R. 2011). Technology are more and more interconnected and intertwined with human actors (Quan-Haase 2016). Although claiming the equality of humans and non-human objects as conscious actors may seem confusing, the hypothesis is legitimate; any object becomes part of the network due to its interaction within. Giving technology the freedom to work on its own (e.g. machine learning), makes it independent from the supervision of a human actor.

Technology on its own is no neutral thing. It is always connected to its network and, thus, connected to human actors (Kranzberg 1986). A plain piece of paper is nothing more than a potential tool. Writing on it, however, transforms it into something much more than a mere piece of paper. This new object evolves into data or maybe even information, and is integrated into a distinct network. Writer and paper are put into a relationship which, on top of that, is highly contextualized. Where was the paper written on? In what way? The mere use of an object (medium) already creates a message (McLuhan 1967), even without knowing what is written on the paper. The purpose of technology is constituted by its connection to reality (Whitehead 1929). On the other hand, it is a convention (Latour 2005) that connects various realities (Berger & Luckmann 1966), thus shaping the social habitus (Bourdieu 1977).

Technology in a broader sense can be categorized into "(1) material substance, (2) knowledge, (3) practice, (4) technique, and (5) society" (Quan-Haase 2016: 4). The first understanding disregards the connection between technology and society completely and views technology only as a passive tool under human control (Feist et al. 2010). The second definition compares technology to knowledge. Technological

knowledge can be seen as a focus on the ability to create artifacts, objects that are created according to human intentions (Hershbach 1995), which re-presents the translation from technological ideas into designs, objects, and forms (Layton 1974). The third definition goes beyond the idea of creating artifacts and compares technology to a system, thus understanding it as a practice enclosed in human activities (Franklin 1999). Franklin uses this argument to reveal the negative potential of technology, especially in the context of becoming a normal part of society's routines and thereby changing people's behavior altogether. The fourth definition understands technology as a technique. 'Technique' is derived from the Greek word *technikos* and denotes human activity that possesses a certain goal and mechanism (Heidegger 1977). Technology is, therefore, not only a tool but rather a mechanism related to human activities (Ellul 1964). The final definition approaches technology with regards to society. Simpson (1995) sees technology as a change agent for society and some go even further: "It doesn't push things forward or transform the world, it becomes the world" (Baudrillard & Gane 1993: 44).

The evolution of definitions reveals that technology and society are indeed highly entangled. One could even propose a fusion of society and technology (Quan-Haase 2016). In terms of big data that claim may already be genuinely true, but even though they appear as two distinct actors within a network (Latour 2005), society influences big data and big data influence society. The relationship between big data and society is a deterministic one. Following this logic, sharpening the distinction between technology and society is essential to understanding the potential interferences between society and big data. Quan-Haase derived the following definition from a similar argument:

> "Technology is an assemblage of material objects, embodying and reflecting societal elements, such as knowledge, norms, and attitudes that have been shaped and structured to serve social, political, cultural, and existential purposes" (2016: 9).

On this basis, the interconnection and ubiquitous force of big data surrounding every human can be deconstructed and separated. This separation is essential to understanding both the technological deterministic and social deterministic viewpoints. According to the respective literature, this unilateral deterministic view falls into the category of hard determinism (Marx & Smith 1994). Both aspects are essential to deriving a better understanding of the socio-technologically intertwined relationship between big data and human actors.

2.2.2 Technological Determinism

According to the theory of technological determinism (e.g. Blauner 1964, Smith & Marx 1994), technology is the driving force behind social change. Any change in technology will cause society to adapt. Technological determinism identifies technology as the driving force behind social change in history (Kunz 2006) and is described as "the most important cause of change" (MacKenzie & Wajcman 1985: 4). This concept sees technology as independent and autonomous, guided by an internal

logic (Quan-Haase 2016). The term 'technological determinism' was coined by Thomas Veblen (1921/2001) and built up on observations by Karl Marx: "The hand-mill gives you society with the feudal lord; the steam-mill, society with the industrial capitalist" (Marx 1971: 109). Although Marx criticized technological determinism and broadened it in order to include social productive forces (MacKenzie 1984), the theoretical debate substantiates the influence of technology: "The uses made of technology are largely determined by the structure of the technology itself that is that its functions follow from its form" (Postman 1992: 7). This determinism is strongly linked to the idea that "given the past, and the laws of nature, there is only one possible future" (van Inwagen 1983: 65). Moore's law (Adee 2008), for example, can be seen as self-directed and following its own form. The number of transistors is growing at an exponential rate, apparently without any influence from society. Technological progress in this field is determined by technology itself. To strengthen this potential of self-direction, Heilbroner (1967) points out the phenomenon of simultaneous discoveries (Merton 1961, Bikard 2012) and, consequently, supposes a general predictability of technology (Bellow 1959, Martin 2010). Heilbroner reports the absence of sudden technological leaps. This argument, however, may lose its validity in modern times (Sood & Tellis 2005). Technological life, consequently, dominates social, political, and economic life (Ellul 1964), as well as technological norms of practices, under the aegis of rationalization (Habermas 1970). This perspective, though frequently described as bold (e.g. Bimber 1994), describes the idea that society adapts to technological change (Miller 1984). Winner defines technological determinism as a form of technological somnambulism (2004). He asserts that the behavior of society compares to sleepwalking when it comes to technology. Technology creates new worlds and restructures everyday lives. Interestingly, Winner (2004) highlights that people tend to be blinded by the usefulness of technology and, therefore, do not realize the societal transformations brought upon them by technology. Technology acts as a change agent for changes that often go unrecognized by society.

The general idea behind technological determinism seems insufficient when trying to grasp the interaction between technology and society (e.g. MacKenzie & Wajcman 1985, Degele 2002), but is still popular as it conveys an understanding of the force of technology in today's world, especially due to the assumption that technology will eventually solve today's problems, and deliver a technological fix (Drengson 1984) capable of clearing the way to a technological utopia (Segal 1984). Interestingly, the idea of a technological fix re-emerged in the discussion about big data. Morozov criticizes the belief that every complex problem is solvable "if only the right algorithms are in place" (2013: 5), for which he coins the term 'technological solutionism'. His argument underlines the potential of big data to be technologically deterministic. The effect seems increasingly prevalent because big data can be opaque, obscure, and overwhelmingly complicated. Big data as a type of technology appears about "to slip from human control" (Heidegger 1977: 289).

Big data have an inherent ability to enforce social change, which is clear in the current applications of big data. Based on algorithms, big data generate a certain

and distinct form of reality. One aspect of changing reality is the filter bubble (Pariser 2011) that shields society from certain data. Chomsky (2013) clarifies: "We can find lots of data; the problem is understanding it. And a lot of data around us go through a filter so it doesn't reach us." Due to the abundance of data, people tend to over-estimate the effect of data drastically, and are more inclined to rank results as objective and, consequently, truthful. Big data have a tendency to mirror reality (Bolin & Schwarz 2015). This is why big data construct social reality, hiding behind a veil of objectivity and granularity, but are in fact adhesively subjective and imprecise. This creates a kludge that is comparable to an assemblage (Kitchin & Lauriault 2015) or mosaic (Sprague 2015) of reality. Society, however, recognizes this mosaic as reality which makes big data representative of a type of social constructivism or, to be more precise, a form of "data constructivism" (Scholz 2015a: 10).

On the basis of data constructivism, social cognition is guided by big data and the underlying algorithms. Algorithms, however, always include an ideology of sorts (Mager 2012). Such an ideology is not bound exclusively to subjective measures, but similarly connected to statistical measures. Pentland (2014) argues that society follows certain rules and abides by statistical regularities. Such regularities can be seen by means of big data and, therefore, used to optimize society (Mayer-Schönberger & Cukier 2013). Carr (2014), however, criticizes this approach as follows: "Pentland's idea of a 'data-driven society' is problematic. It would encourage us to optimize the status quo rather than challenge it" (2014), as it could potentially result in a deadlock (Rätsch 2015). Big data will transform into civic "thermostats" (McLuhan 1967: 68) while declaring the ideal temperature (Carr 2014). People may become "slaves to big data" (Hildebrandt 2013: 1).

Social engineering of this type strongly influences people's behavior, not because the data are extremely precise, but because it shows a socially accepted path. People are nudged (Thaler & Sunstein 2008) into a distinct direction which renders them more willing to follow that path. Anticipatory obedience (Lepping 2011) turns that prediction into a self-fulfilling prophecy (Merton 1948) or even self-preventing prophecy (Brin 2012). Such a prediction merely needs to appear probable for it to actually happen. Data constructivism has substantial power over people and will lead to a form of social control (Scholz 2015a). Any data-driven society is determined purely by the use of available data, which may be insufficient. While statistical analyses may be rigorous, they may lead to completely different results following either the Gaussian or the Paretian statistical principle (Bolin & Schwarz 2015). Even though the results are objective and statistically correct, they are only a chimera of reality (Brenner 2013).

Assuming that big data shape society leads to a data constructivism of reality. Even though this may sound promising, a deficiency of big data will not fix all societal problems but will lead to a specific type of social engineering. People more and more commonly act according to rules that are disguised as socially acceptable. Frischmann (2014) even proposed a Reverse Turing Test, which does not aim at measuring the ability of a machine to act like a human being, but rather a human's ability to resemble a machine. The test is intended to reveal whether humans are,

indeed, controlled by nudges from technology such as computers, smartphones or wearables (Yeung 2016).

All things considered, big data provides society with as much information as possible, consequently overwhelming it (similar to Huxley 1932). It may be so overwhelmed, in fact, that society as a whole is nudged towards an entity, such as an algorithm that selects "relevant" information (similar to Orwell 1949) for society. Such algorithms conceal the objective truth from society on the basis of big data and, perhaps not even purposefully, result in the data constructivism of reality.

2.2.3 Social Determinism

The opposite relationship is social determinism (e.g. Pannabecker 1991, Green 2001). Supporters of this concept believe that society creates technology in order to fulfill a certain need in society (Quan-Haase 2016). This theory identifies people as central drivers of change. As a result of societal change and the corresponding societal needs, people develop new technologies which then lead to technological progress. Contrary to technological determinism, technology is not autonomous and self-directed (Winner 1993). Society or social groups attribute meaning to technology and its use or impact. For that reason, technology is influenced by a variety of social factors such as history, economics, and ideology (Giddings 2006). Technology is a social construct that receives its meaning and relevance from society (Winner 1993). Following that idea, technology is being developed to saturate society's needs, thus overcoming human limitations.

The social deterministic view has spiked several advances within the discourse, and has contributed to the emergence of the academic field "science and technology studies" (influenced by the seminal work of Kuhn (1962)). The academic field rapidly rejected both technological determinism and social determinism (Bijker et al. 1999), and realigned its focus on the mutual shaping (Boczkowski 2004) of society and technology, however, in the beginning of the field, some researchers called this social deterministic influence the social construction of technology (e.g. Bijker et al. 1987). They also proposed a superiority of society over technology in that it is society that shapes technology. The use of technology is also influenced by social context (Klein & Kleinmann 2002). Pinch and Bijker (1984) define this concept on the basis of four key terms: (1) relevant social group, (2) interpretative flexibility, (3) closure and stabilization, (4) wider context. The authors argue that technology can only gain meaning and consequently survive when receiving societal support (Pinch 2009). A lack of support or interest from society prevents the development of a certain technology, or as the authors state: "a problem is only defined as such, when there is a social group for which it constitutes a 'problem'" (Pinch & Bijker 1984: 414). On this basis, interpretative flexibility suggests that technology is not neutral, as its meaning can vary according to social context. This can be exemplified by the discovery of microwave ovens. They were accidently discovered in a radar station when, by coincidence, a melted candy bar was found in close proximity to the source of radar waves (Andriani & Cohen 2013). Microwaves as a specific

technology were moved from one socio-cultural context into a completely different one by a certain social group. While the technology (microwaves) remains the same, its meaning for society obviously differs drastically depending on whether the waves are used for tracking planes or cooking food. This change in usage can be described as cultural transduction (Uribe-Jongbloed & Espinosa-Medina 2014). Andriani and Cohen (2013) also established the concept of closure and stabilization. Closure is reached once a social group comes to an agreement about the purpose of a certain technology, and stabilization is achieved when the technology seems ready for the market. Finally, there is the wider context. Even if a certain social group supports and demands a new technology, it may still be rejected. It seems obvious for example that the future belongs to the electric car, however, many people still currently reject it (Pierre et al. 2011). Various social groups support the electric car, and it tackles a specific societal need, but in the wider context a majority choose a different technology (Winner 2003).

Following the idea of social construction, Winner (1993) links research into the social construction of technology with social constructivism. Lawson (2004) goes further by arguing that any social-deterministic view is rooted in social constructivism. Following this argument, social determinism can be linked to big data as a technology, and big data understood as being shaped by a form of social constructivism. Social constructivism, therefore, serves as the counter-argument to data constructivism. One essential aspect of the following discourse is that there are social groups that shape technology, and as Winner (1993) reports, those groups will take over elitist roles and obtain great power. Floridi (2012) reasons that the game of using big data will be won by those with the ability to use big data and, as he quotes from Plato, "those who 'know how to ask and answer questions' (Plato, Cratylos, 390c)" (Floridi 2012: 437). Floridi raises a crucial concern that big data will lead to a big data divide (Andrejevic 2014) and that the people with the power, as well as the ability to use big data, will shape the social reality of those without access to, and knowledge of, big data. Boyd and Crawford (2012) separate society into big data rich and big data poor. The big data rich will have the power to use social constructivism.

The most defining work regarding social constructivism was written by Berger and Luckmann (1966) who claimed that society can be seen as both an objective and subjective reality. In their sense, however, objective did not carry the same connotation as the commonly used term 'objectivity' as the authors clarify: "It is important to keep in mind that the objectivity of the institutional world, however massive it may appear to the individual, is a humanly produced, constructed objectivity" (Berger & Luckmann 1966: 60) and, to be more concise that "society is a human product" (1966: 61). As far as subjective reality is concerned, this mainly focuses on the reality constructed by individuals on their own in developing their subjective reality through socialization and interaction with nature. Berger and Luckmann elaborate on the idea as follows: "Man is biologically predestined to construct and to inhabit a world with others. […] man produces reality and thereby produces himself" (Berger & Luckmann 1966: 183).

The authors claim that objective and subjective reality co-exist and mutually shape social reality, however, the process of social constructivism follows a certain logic. Social reality is created through a certain form of externalization. A particular social group comes to a preliminary consensus, thus creating a set of norms which in repetition becomes habitualization. If this constructed reality is reproduced by others, it becomes institutionalized, will be aligned with new institutions, and form a shared language. This stage is followed by a phase of objectification during which the construct of reality receives legitimation. This institutionalized and legitimated reality will now be passed down from generation to generation and consequently internalized by society. The result of this process is human-constructed objectivity.

When considering only a small social group, the assumption of such processes of social interaction constructing reality appears plausible. New technologies like big data, however, can be used alongside the constitution of social reality. Big data can help repeat certain social realities, dominate the language in use, reveal the need for certain institutions, support the process of legitimation and transfer such a reality to other individuals and generations. Big data are, therefore, a powerful tool with which to internalize a certain social reality. Such effects are intensified by changes in society's communicative behavior. Berger and Luckmann have already envisioned a future that seems similar to today's digitized communication:

> "The social reality of everyday life is thus apprehended in a continuum of typifications, which are progressively anonymous as they are removed from the 'here and now' of the face-to-face situation. At one pole of the continuum are those others with whom I frequently and intensively interact in face-to-face situations – my 'inner circle', as it were. At the other pole are highly anonymous abs actions, which by their very nature can never be available in face-to-face interaction. Social structure is the sum total of these typifications and of the recurrent patterns of interaction established by means of them. As such, structure is an essential element of the reality of everyday life" (Berger & Luckmann 1966: 33).

In the future, structure will be an essential parameter for social constructivism and a process of structuring is already ongoing. As previously explained, big data divide society into two groups, with the elitist group shaping the use of big data. One group will always be using new means for their own purposes. Such behavior is commonly referred to as Campbell's law:

> "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (Campbell 1979: 85).

While the question of whether big data would knowingly be used for harmful purposes may be debatable, there is a temptation to use big data to nudge people's behavior in a certain direction (Schroeder 2014). One popular example is the potential of the media to steer, manipulate, and control social reality (e.g. Chomsky 2002). Despite great controversy, differences in coverage inevitably change the narrative

of that coverage as observed, for example, in the context of climate debate (Feldman et al. 2012). Selecting certain information and omitting other information can be used to change social realities. This effect becomes even more pronounced when not all available information can be accessed. Society is "drowning" (Puschmann & Burgess 2014: 1699) in data and seeking any available reduction in complexity. Such a reduction in complexity is controlled by different social groups and, therefore, entangled with certain agendas and ideologies (Mager 2012).

Information is selected by certain social entities in order to nudge people into certain behavior (similar to Orwell 1949). People unconsciously commit themselves to a social entity, as they are overwhelmed, and scared away from any other social entity (similar to Huxley 1932). Such social groups shape their own realities, and have the ability to spread and impose their reality on other people, resulting in a social constructivism of reality.

2.2.4 Socio-Technological Concurrence

Speculating about the effects of technology on society and those of society on technology eventually reaches its limits. This is why the interdisciplinary field of science and technology studies has developed a stance of general rejection towards technological determinism and social determinism and instead focuses on the mutual shaping of technology and society. Technology cannot be seen separately from society, much as society does not evolve independently from technology. Both are integral parts of a holistic socio-technical system (Bijker et al. 1999). A first approach to describing the relationship between technology and society is technological momentum (Hughes 1969). Hughes uses technological and social determinism but connects both models from the perspective of time. He claims that "a technological system can be both a cause and an effect; it can shape or be shaped by society" (Hughes 1994: 112). Over time, a technological system will move between the extremes that are technological determinism and social determinism. Hughes reasons that a technology is shaped by society at first, but over time evolves into a technology shaping society. It gathers *momentum.*

Elaborating on previous research, Callon, Latour, and Law developed Actor Network Theory (ANT) in the 1970s (Murdoch 1997). Building upon the concept of mutual shaping, ANT understands technology as dynamic process (Latour 1987). It therefore influences the social network, which renders it a part of the network. The theory attributes to technology the ability to act in a certain way, however, and influence the network. In this way, it varies from the perspective of social constructivism (Giddings 2006). As Latour highlights by himself: "[ANT] entirely bypasses the question of 'social construction' and the realist/relativist debate" (Latour 1999: 22). ANT is, thus, incompatible with theories such as the structuration theory by Giddens (1984) and, therefore, does not follow the duality of technology within organizations (Orlikowski 1992). The relationship is described as follows:

> "If human beings form a social network it is not because they interact with other human beings. It is because they interact with human beings and endless other materials too […] Machines, architectures, clothes, texts – all contribute to the patterning of the social" (Law 1992: 382).

ANT links human beings and non-human beings and goes even further in the sense that it ascribes to those non-human beings (any kind of technology) the role of an actor (Latour 2005). Both share a form of social assemblage and "enter a stable definition of society" (Latour 1991: 129). MacKenzie and Wajcman (1999), interpreting the relationship between society and technology as described by ANT, conclude that they are made out of the same material, thus linking human and non-human actors in the same way. Latour underlines this claim in claiming to "see only actors – some human, some non-human, some skilled, some unskilled – that exchange their properties" (Latour 1992: 236). Other researchers agree, such as Bryant (2011a), for example, who understands both humans and technology as part of a relational network and, within the network, equal objects. Giddings (2006) dissolves the difference between technology and users, which represents a shift from traditional viewpoints. Latour (2005), however, uses the example of an airplane. The airplane is not only controlled by a pilot, but by an uncountable number of different actors that, within their relational network, keep the airplane flying. The focus shifts from the question of why relationships between actors exist (human and non-human beings), to the question of how these relationships work (Quan-Haase 2016).

When ANT is applied to big data and society, it fulfills the need for mutual shaping. The idea that big data and humans within the societal network are viewed from a relational perspective (Giddings 2006) aligns with the analogy of big data being a force surrounding everybody. Connecting big data to society in such a way mimics the influence of big data on social reality. This thought does not entail that one dominates the other or vice versa, so much as both actors enriching each other, which leads to a form of socio-technological concurrence of big data and society, mutually contributing to the construction of reality.

Proposing a certain concurrence, however, means that technology and society are changing, growing, and evolving together, but not at the same speed. Big data in particular are growing at a pace so mind-boggling that people compare big data to an avalanche (Miller 2010). The videogame "Mass Effect 2" (issued by Bioware in 2012) includes a quote by Mordin Solus (he is an alien which accounts for his unusual English) that targets the problems of such an imbalance:

> "Disrupts socio-technological balance. All scientific advancement due to intelligence overcoming, compensating, for limitations. Can't carry a load, so invent wheel. Can't catch food, so invent spear. Limitations. No limitations, no advancement. No advancement, culture stagnates. Works other way too. Advancement before culture is ready. Disastrous."

Solus confirms the concurrence of technological and societal progress and underlines the importance of a certain balance between both aspects. In a nutshell, big

data can be seen as an actor within the social network, shaping society as much as society shapes big data. There is, however, a need to master this new technology (Heidegger 1977). Merely making big data an actor within the societal network would not suffice, as they would be too raw and too vague. In a next step, therefore, big data need to be explained from within an organization and by means of existing organizational theory. This is especially reasonable considering big data's current reputation as merely being a form of "unclear technology" (Cohen et al. 1972) within organizations.

2.3 Big Data at the Organizational Level

2.3.1 Epistemological Framing

The previous chapters reveal the difficulty of grasping big data as a theoretical construct. Big data are not clearly outlined as a type of technology separated from society, but need to be treated as technology that interacts intensively with society. In fact, the meaning and influence of big data evolve through this delicate interaction. There is no deterministic direction but a concurrence between both actors. There is also a certain friction between big data and society, which is why their relationship mimics duality. In this case, duality denotes the instance of both functions seeming contradictory, while in fact being complementary (Evans & Doz 1992). Janssens and Steyaert define this as follows:

> "Duality has the most general meaning of the three concepts: paradoxes and dilemmas can be seen as dualities, but not all dualities can be seen as paradoxical or simultaneously contradictory, or involving an either-or situation or an impossible choice" (Janssens & Steyaert 1999: 122–123).

The duality between big data and the members of an organization are even more apparent. Despite this, the task of grasping big data as part of an organization is more complicated. Big data are not something that can be 'gripped' and, therefore, are not always restricted to one organization alone. But there are what is referred to as 'organizational big data', big data that are uniquely contextualized towards one particular organization. When trying to explain big data within an organization, on the basis of organizational theory, one criterion to pay attention to is the closeness and openness of a system. The following chapters outline a series of theories located somewhere between predominately closed systems and completely open systems, and searching for theoretical explanations of the behavior of big data within an organization.

Several preceding comments on big data and organizational theory are necessary, especially concerning three assumptions that recur in the discourse of organizational theory, and that will be distorted by the implementation of big data within organizations: the assumption of bounded rationality, the modernist and postmodernist view of organization, and the discourse about the iron cage.

The term *bounded rationality* was coined by Simon (1959) and describes the fact that people make decisions on the basis of their limited information. Based on knowledge gained from incomplete information, people use heuristics to decide things and, therefore, sometimes act irrationally. People decide within certain constraints, which can be without apparent contextual, procedural reason or result from various other influences (March 1978). People lack the ability to make perfect rational choices (Pescosolido 1992) and are only capable of choosing the option that seems best to them. Allison (1969) vividly illustrates this in his analysis of the decision process during the Cuban Missile Crisis. People are bound by their incomplete rationality and especially by the limitations of the information available to them. Big data increases the amount of available information and, therefore, apparently contributes to more rational decision making.

Such a conclusion, though on a smaller scale, was discussed by March (1978) who highlights the risk of self-evident empirical truths. People tend to fall for the "illusion" of rationality that accompanies big data, but in order to form a basis for rational choice, big data need to be a precise image of reality. Its range, however, although far beyond that of an individual, is limited to only a small portion of reality. In the words of Wittgenstein: "The limits of my data [originally: language] mean the limits of my world" (1922: 74), but since big data are incomplete, any decision on the basis of big data will always rest upon bounded rationality. No organization on its own will have access to all big data, which is why any organization is regulated by bounded rationality. Decision making within an organization is, therefore, not only subjugated by the individual bounded rationality of people, but also by the bounded rationality that derives from accessible and available big data. Such *bilateral bounded rationality* establishes the constraints of organizational bounded rationality.

The next assumption regarding organizational theory is that theories can be categorized into a *modernist* and *postmodernist approach* and, above all, describe organizations differently (Cooper & Burrell 1988). Boisot and McKelvey (2010) proposed the provocative idea that modernist and postmodernist perspectives can be bridged. In times of big data, it may be a compelling argument that there is a need to bridge these perspectives. The authors connect modernism to positivism and, thus, link empirical observation with objectivity. This is only possible through repetition and the replicability of events. This is often interconnected with the Cartesian view (e.g. Miller 2008), according to which any cause brings about an explainable effect as well as some form of stable environment. Although such a reductionist view has been vigorously challenged (e.g. Alvesson & Kärreman 2011), Gaussian statistics focusing on normal distribution are still widely applied in the (social) sciences (e.g. Greene 2003). The goal of modernism is to produce robust and objective knowledge, but the postmodernist critique is that the stories derived from the acquisition of knowledge (Calás & Smircich 1999) are socially constructed stories. The social world consists of a form of radical subjectivity (Foucault 1977) and is influenced by power (Townley 1993), the scope of interpretation (Latour 1988), or regional and cultural contexts (Soja 1999). On the basis of this idea Boisot and McKelvey (2010) state that postmodernist distrust laws that are derived from normal distributions.

Behind every result there is a story or narrative that leads to "infinite conversations" (Wyss-Flamm & Zandee 2001: 297). As opposed to a reductionist view, postmodernists try to engage in social complexity (Cilliers 1998). Consequently, there is no way for any researcher to find one objective truth, at least not in the social sciences. The authors see a way "to integrate the ordered world of modernists and the more 'chaotic' world of postmodernists" (Boisot & McKelvey 2010: 416), especially as both worlds describe an atomistic and a connectionist ontology that may be scalable.

In the context of big data, scalability is the clue. Big data can be granular, detailed, and precise, but also be general and universal. Big data can record the interactions of organizations on a global scale, as well as the interactions of individual employees on a local scale. In the context of an organization, particularly, a story or narrative can be monitored over time by means of big data (Kim et al. 2013), which may result in a unique organizational signature (Stein et al. 2016). Such use of big data, however, depends on the combination of both the modernist and postmodernist perspectives. On the one hand, the modernist view delivers an over-generalization in the sense that all organizations seem similar. On the other hand, the postmodernist view would lead to a conception of organizations, it was so contextualized that a comparison would be altogether impossible. Big data are not merely a tool to bridge both views, but also provide organizations with a form of *scalability* between both views.

The final assumption to be discussed before analyzing the theories is the existence of the *iron cage*. In 1952, the first English translation of Weber's book "The Protestant ethic and the spirit of capitalism" used the term iron cage. However, in the German original Weber was talking about *stahlhartes Gehäuse* which more precisely translates into "shell hard as steel" (Baehr 2001: 153). Baehr traces back this fundamental change in meaning to the free interpretation by the translator Parson. The author elaborates, furthermore, that cage means being trapped in something, but shell describes a "living space both for the individual who must carry it around" (Baehr 2001: 163). In a certain way, the actual meaning of Weber is comparable to an augmentation of the organization and as something the organization can carry around. Such a shell could be beneficial or harmful, but such a metaphor would "appear anticlimactic" (Baehr 2001: 164). Nevertheless, the term iron cage became popular in social science (Baehr 2001) and, therefore, is used commonly for describing the situation in which organizations and their members are caged within a bureaucratic rationalization. Organizations are shackled by a precisely organized and mechanically tuned bureaucracy and, due to its apparent superiority, other organizations tend to adapt to such a rationale. DiMaggio and Powell (1983) agree that organizations have a tendency to homogenize their structures, however, they do not ascribe the homogenization to bureaucracy but to structuration (Giddens 1979). The basis of this is the question of why there are so many similar organizations. DiMaggio and Powell (1983) claim that this homogenization can be described as isomorphism. They follow the description of Hawley (1968): "Isomorphism is a constraining process that forces one unit in a population to resemble other units that face the same set of environmental conditions" (DiMaggio & Powell 1983: 149). Bureaucracy is only one reason for isomorphic change. Other strong influences are

the distribution of power and the social legitimacy of an organization in comparison to other organizations. DiMaggio and Powell (1983) list the following mechanisms as driving forces for institutional isomorphic change. (1) *Coercive isomorphism* refers to institutions that pressure organizations in a certain direction. Although the introduction of new laws is the most prominent example, informal pressures are likewise possible. (2) *Mimetic pressure* is the belief that an organization becomes safer or more stable by mimicking another organization. (3) *Normative isomorphism* denominates structural change due to the professional background of an organization's members. Similar educational backgrounds encourage isomorphism. All three factors have an influence on the isomorphic tendency of one organization.

Big data can contribute to the isomorphic tendency of an organization through any of these three factors. External pressure is reinforced by supportive data, which leads to more coercive isomorphism. Big data reveal the fittest companies and provide enough information to mimic another organization completely, which, therefore, increases mimetic pressure. Big data also allow everybody to gain knowledge. Said knowledge, however, underlies a certain homogenization, as well as a form of Westernization (Wilson et al. 2006) which boosts normative isomorphism. Big data open up new sources of information to an organization. Assuming that big data contribute to anti-isomorphic tendencies, however, would be a false conclusion. Overall, big data increase and reinforce isomorphic tendencies, metaphorically securing the iron cage. Big data, being incomplete, lead to data constructivism and the creation of a certain reality. This increases the isomorphic tendency in a certain direction in accordance with this one data-constructed reality. The use of big data also gives any institution power and legitimation due to an apparent trust in numbers (Porter 1996). In addition to being reinforced, the iron cage becomes transparent. In their original article, DiMaggio and Powell (1983) give directions for those institutions that influence the isomorphism within an organization in a similar way to Bentham's panopticon (1843), in which the inmates of a prison are watched by a number of watchmen on a tower at all times. Organizations know that there is isomorphic pressure and, thus, adapt to normative expectations. In times of big data, however, organizations find themselves facing a post-panoptical scenario (Baumann 2000). In this case, organizations no longer know who is pressuring them and where the isomorphic tendency is directed, but the need to adapt to other organizations in the sense of homogenization remains obvious. When big data single out a certain type of organization as beneficial, organizations will change to match this structure, unaware of who decided it, and how this institution came to that conclusion.

In summary, all three assumptions reveal that big data are part of any organization and will play an integral role in understanding organizations. However, big data construct a new layer of reality within organizations. Big data cannot be seen as a mere source of information and, therefore, an external and objective factor, but much rather as an internal and subjective factor. Big data will contribute to certain solutions and intensify other problems. As the complexity of big data and their

implementation increases, big data do not stop at the border of an organization as they are becoming increasingly heterogeneous.

2.3.2 Organizations as Open Systems

Generally speaking, organizations are never completely closed systems which marks a major difference to other fields of research. Organizations always interact with their environment to a certain extent. Von Bertalanffy (1968) was, in the 1940s, one of the first to describe the difference between closed and open systems. While it is necessary to define closed systems within the realm of physics, all other systems that are organized in any form will differ because they interact with other organizations from the outside.

> "However, we find systems which by their very nature and definition are not closed systems. Every living organism is essentially an open system. It maintains itself in a continuous inflow and outflow, a building up and breaking down of components, never being, so long as it is alive in a state of chemical and thermodynamic equilibrium but maintained in a so-called steady state which is distinct from the latter" (von Bertalanffy 1968: 39).

In the social sciences particularly, organizations are never closed systems, but systems that are living or social (Luhmann 2011). Luhmann (2011) also describes a different form of closed systems: a system can be operationally closed, which means that there is an outer side to an organization that faces the environment, as well as an inner side of an organization that does not interact with the environment and conducts tasks and operations completely independently of it. Contrary to a completely closed system, only the operational tasks are separated from the environment. Normally, such operational tasks are precisely described and there is no need for external interaction. For example, the production of a sheet of paper takes place within an organization and without interaction with the environment and, therefore, is operationally closed, though everything else is done in interaction with the environment. Separating the operational perspective from the general system is reducing complexity and enables the researcher to focus on the observation of organizations.

Table 6: Overview over the Theories on Open Systems

Operationally Closed System		**Open System**
Cybernetics	Systems Theory	Population Ecology Theory
Complex Systems Theory		

In order to integrate big data into the network of organizational theories, I will use the structure shown in Table 6. All theories mentioned there have increasingly opened towards the environment. All three selected theories contribute to a dynamic perspective on organizations and are linked with the complex systems

theory. One final preliminary remark about the differentiation between cybernetics and systems theory: both theories are relatively similar and, therefore, the terms are often used synonymously (von Bertalanffy 1972). As a matter of fact, it is sometimes difficult to attribute a certain concept to one distinct field of theory and the correctness of the following selection of concepts is subject to debate. In the course of this thesis, however, both theories will be differentiated according to one specific aspect: cybernetics considers a system more from a predominantly technical or mechanical perspective (Ashby 1956), and systems theory rather from a social or organic one (Luhmann 2011). I assume that cybernetics will contribute more towards understanding the effect of big data on a social system, while systems theory will likely contribute to understanding the effect of a social system on big data.

2.3.2.1 Big Data in Cybernetics

The term cybernetics goes back to Wiener who, in the title of his corresponding book, defines them as "control and communication in animal and machine" (Wiener 1948). The term is derived from the Greek word *kybernētēs* and means steersman or pilot. Ashby adds to this by explicitly characterizing cybernetics as "the art of steermanship" (1956: 1). He also specifies it as "theory of machines" (1956: 1), but moves away from merely describing the machine in favor of trying to understand its behavior. Rooted in such mechanical thinking, cybernetics has influenced the computer sciences (Umpleby & Dent 1999), robotics (Arkin 1990), simulations (Forrester 1994), and the internet (Licklider 1960). The theory was also expanded to social systems and had a strong impact on the understanding of organizations (Morgan 1982). A prominent example of the use of cybernetics in a social system was the steering of Chile in the 1970s by a cybernetical system called CyberSyn, as envisioned by Stafford Beer (Medina 2006).

Cybernetics can be categorized into first order cybernetics and second order cybernetics (von Foerster 1979). The main difference is the role of the observer in the respective systems; in the first order, the focus is on the observed system. The second order, however, focuses on the actions of the observer. Umpleby (1990) summarized several definitions as depicted in Table 7, and expands the general definition by stating that first order cybernetics involves focusing on the model of a controlled system, and second order cybernetics makes the modeler central and treats the system as something autonomous. He contributes his own definitions which highlight the differences in interaction and the differences in the use of theory. Cybernetics shifted from a realistic or positivist view towards a more constructivist perspective (von Glasersfeld 1979).

Table 7: Definitions of First and Second Order Cybernetics

Author	First Order Cybernetics	Second Order Cybernetics
von Foerster	The cybernetics of observed systems	The cybernetics of observing systems
Pask	The purpose of the model	The purpose of the modeler
Varela	Controlled systems	Autonomous systems
Umpleby	Interaction among the variables in a system	Interaction between observer and observed
Umpleby	Theories of a social system	Theories of the interaction between ideas and society

(Umpleby 1990: 113)

One key concept within first order cybernetics is the law of requisite variety (Ashby 1956), which is often reduced to the quote that "only variety can destroy variety" (1956: 207). In more detail, Ashby explains that a fixed amount of variety is imposed by an external player D (disturbance) and that there is a variety of responses to come from a player R (response). He explains that "only variety in R's moves can force down the variety in outcomes" (1956: 206). Any R is capable of regulating the variety of outcomes due to the external input of variety by D. However, the "capacity as a regulator cannot exceed R's capacity as a channel of communication" (1956: 211). Ashby also states that there is the "hard external world, or those internal matters that the would-be regulator has to take for granted" (1956: 209), which he calls T (table). Thus, T is influenced by the variety of D and regulated by R. R is of utmost importance in regulating D and T, in order to influence the outcome.

Transferring this concept to organizations, there is a variety of external input as well as a variety of responses from organizations, that will lead to a variety of outcomes. Boisot and McKelvey (2010) call the spectrum of variety 'the Ashby space'. They propose the idea that an organization deals with a variety of stimuli and has a variety of responses. Both varieties can be low and high in this model. As defined by Ashby, however, a high variety of stimuli will lead to a high variety of responses. The regulation of variety is imposed by some form of ordering principle that tackles T – the authors use algorithmic compression as an example (Boisot & McKelvey 2010) – which makes it possible to categorize the Ashby space into an ordered regime (low variety of stimuli and low variety of responses), complex regime (medium variety of stimuli and medium variety of responses), and a chaotic regime (high variety of stimuli and high variety of responses).

Applying this concept to big data in an organization helps in understanding the general effect of big data within an organization. Big data can be seen as an external force that is taken for granted by the organization as well as an external force that disturbs the organization. Big data contribute massively to the variety of

stimuli and, if unfiltered, will lead to a massive increase in the variety of response and will exceed the communication capacity of any organization. The organization will require the capability to "block" (Ashby 1956: 212) harmful big data and let the beneficial big data in. The organization will respond towards big data and incorporate useful and relevant information. Consequently, there are big data that may be beneficial but also that may be harmful for the organization. Ashby, as well as Boisot and McKelvey (2010), highlight that such regulation is the task of an organization, especially where the response of regulation influences the chances of an organization's survival (Ashby 1956). Big data will be regulated and ordered in some form in order to destroy variety and lower the variety of outcomes. The law of requisite variety claims that big data will be regulated by the organization itself and not by any external source, thus enforcing the idea that any organization *will deal with their own big data on their own.*

Another popular concept in first order cybernetics is homeostasis (Wiener 1948, Ashby 1952, Boulding 1956), based on the homeostasis concept as introduced by Cannon (1926). He defines homeostasis as follows:

> "The highly developed living being is an open system having many relations to its surroundings […]. Changes in the surroundings excite reactions in this system, or affect it directly, so that internal disturbances of the system are produced. Such disturbances are normally kept within narrow limits, because automatic adjustments within the system are brought into action, and thereby wide oscillations are prevented and the internal conditions are held fairly constant. The term "equilibrium" might be used to designate these constant conditions. […] The coordinated physiological reactions which maintain most of the steady states in the body are so complex, and are so peculiar to the living organism, that it has been suggested […] that a specific designation for these states be employed – homeostasis" (Cannon 1929: 400).

Cannon already addresses the potential misinterpretation of stasis as being inflexible or even stagnating. Stasis also implies a certain condition, however, and in combination with the term 'homeo', meaning similarity, homeostasis is the concept of a system that is "to maintain uniformity" (Cannon 1929: 401). In this context, homeostasis is linked to the steady state concept (Lloyd et al. 2001), according to which such systems will remain constant despite influences from the external environment. Ashby (1952) calls this state a form of ultrastability, in which a system is able to change its internal structure in order to respond to the environment, causing the system to deal with external disturbances without compromising steadiness. Wiener (1948) formulized a form of feedback control that renders negative feedback as a critical source of reaction. Negative feedback is the response of a system to changes from a normal condition, from the steady state or the equilibrium, in order to move the system back to this normal condition. This is contrary to positive feedback, which would increase the departure from the normal condition. Therefore, in order for a system to be homeostatic, it needs negative feedback in order to react accordingly. A system is normally not able to achieve a stable homeostasis, but fluctuates around the equilibrium. Wiener (1948) expects oscillation and an eventual

oversteering of the system. Such behavior can be traced back to the idea that negative feedback does not work in real-time and that any feedback comes with a certain time lag. A homeostat, as Ashby (1952) denotes a system in homeostasis, thus oscillates around the equilibrium, but keeps the system in an ultrastable condition.

Introducing big data to a homeostat implies an external disturbance that will probably lead to a massive deviation from the normal condition. A functioning homeostat changes its interior appropriately due to its ultrastability. Although the system returns to its equilibrium over time, the system is internally transformed and adapts to the new input. Interestingly, there is already some discussion of the idea that algorithms may be acting like homeostats has already entered discourse (Schwefel 1994). The idea implies that any algorithm-based system such as a modern organization tends to stabilize itself but will be transformed by big data. The logic of the homeostat emphasizes that big data may at first be a disturbance, but in the end will be used by the algorithmic system to return to the normal condition, especially as big data themselves, seen as an environmental force, do not have ultrastable features. Big data are constantly changing and transforming, and Ashby would, therefore, probably see big data as a source of variation, noise, and disturbance (Ashby 1952, 1956).

That may be a reason why Wiener declined the idea of homeostasis in society. He stated that "in connection with the effective amount of communal information, one of the most surprising facts about the body politic is its extreme lack of efficient homeostatic processes" (Wiener 1948: 185). In recent times of big data, the amount of information has drastically increased since Wiener's times, but there is still a lack of homeostatic processes. He anticipatively traced it back to the factor of numbers and size, which leads to "anti-homeostatic factors in society" (Wiener 1948: 187). He added that the "control of the means of communication is the most effective and most important" (Wiener 1948: 187). Thinking this further, due to the large size of big data, they can contribute to, but definitely *will influence*, smaller homeostats (any organization). Big data are not a homeostat on their own, however.

To follow those two concepts of first order cybernetics is the general idea behind second order cybernetics or cybernetics of cybernetics. The most important aspect of this new type of cybernetics is the renunciation of an objective reality, and consequently the impossibility of deriving an objective truth. Maturana (1970) and von Foerster (1979) connect the reasoning behind this argument to the observer of such a system. They stipulate that the claim of objectivity is in no way achievable due to the properties of an observer. Any observer will influence the observation to a certain degree. Von Foerster (2003) uses the following example to underline his argumentation.

> "… a brain is required to write a theory of a brain. From this follows that a theory of the brain, that has any aspirations for completeness, has to account for the writing of this theory. And even more fascinating, the writer of this theory has to account for her or himself. Translated into the domain of cybernetics; the cybernetician, by entering his own domain, has to account for his or her own activity. Cybernetics then becomes cybernetics of cybernetics, or *second-order cybernetics*" (von Foerster 2003: 289).

This argumentation reveals the importance of understanding the effect of an observer in any cybernetic system. This observer is not independent and is part of the observed system. This type of observer effect (Robins et al. 1996) is also known in quantum physics, where observing a quantum will change its properties, or where the observers influence their own observation by the mere act of observing (Heisenberg 1927). There is always an interaction between the observer and the observed. Observers influence the observed system with their eigenbehavior and the observers invent their environment (von Foerster 2003). Von Foerster paraphrases the effect as follows: "cognition → computing a reality" (2003: 215).

In the context of big data, the observers are not capable of separating themselves from big data at all, but now have enough information to compute a granular version of their reality. Within second order cybernetics, von Foerster (2003) covers the problem of memory. Memory, following his argumentation, is influenced by hindsight as well as foresight and, to make it even more difficult, the concept of big data is self-referential (Puschmann & Burgess 2014). Big data use data to generate new data in order to analyze data to generate even more data. In this recursive feedback loop, hindsight influences big data through experiences, and foresight is influenced by potentially desirable outcomes. Any observer will push any big data analysis into a new direction (willingly or unwillingly) and these new results will influence the existing observer or a new observer, in a different or the same way. Big data are part of a vicious cycle. Von Foerster summarizes it at the end of his chapter on constructing reality with the following claim: "reality = community" (2003: 227).

Although I suggest that this reality is a subjective one and not the objective reality, big data enable any form of community to construct their own. As noted earlier, big data lead to a data constructivism that exhibits a similar effect. The observer will influence the reality generated through big data and big data will carry this influence even further, creating a distinct eigenbehavior of the observer that spreads through big data. This eigenbehavior competes with other observers' eigenbehaviors, and will eventually lead to an eigenbehavior of the community. Such an understanding strongly supports the argument that big data are subjective (Boyd & Crawford 2012), and that raw data are an oxymoron (Gitelman 2013). Big data generate observations within the system observed by big data. On this basis, big data compute a subjective reality and *will not achieve* an objective reality or the objective truth, because big data are *both observer and observed object at the same time*.

Analyzing this selection of ideas in the theory of cybernetics from the perspective of big data reveals, above all, big data's inability of reaching the objective truth. Claims that big data will lead to an end of theory (Anderson 2008) can be disproven, and cybernetics implies that big data are bound to make understanding reality even more difficult. Cybernetics shows that an organization is obliged to deal with big data on its own in terms of variety, especially if big data influence a homeostatic organization and are influenced by the observer of big data, that is the organization itself. Big data and organizations interact in so many ways and so often that organizations influence big data and big data influence organizations. Any organization could use big data to *achieve ultrastability* within the modern turbulent

environment. What ultrastability means, however, depends on the eigenbehavior of the organization and the resources it invests in dealing with the variety of responses.

2.3.2.2 Big Data in Systems Theory

The term 'systems theory' was coined by von Bertalanffy, although he more commonly addresses General Systems Theory (1968). According to the prefix 'general', the aim of the original theory is truly towards the full page as it is a type of metatheory. Von Bertalanffy described it as a type of *weltanschauung* – world view (Pouvreau & Drack 2007). He did, however, identify a need for a systems approach for organizations in the 1920s (von Bertalanffy 1972) and explained this necessity, and the emergent overflowing of theories concerning organizations as follows:

> "[…] we are looking for another basic outlook on the world – *the world as organization.* Such a conception – if it can be substantiated – would indeed change the basic categories upon which scientific thought rests, and profoundly influence practical attitudes.
>
> This trend is marked by the emergence of a bundle of new disciplines such as cybernetics, information theory, general system theory, theories of games, of decisions, of queuing and others; in practical application, systems analysis, systems engineering, operations research, etc. They are different in basic assumptions, mathematical techniques and aims, and they are often unsatisfactory and sometimes contradictory. They agree, however, in being concerned, in one way or the other, with 'systems,' 'wholes' or 'organization'; and in their totality, they herald a new approach" (von Bertalanffy 1968: 187–188).

In his later research he narrowed down his general and somewhat holistic approach to shape the modern view of systems theory. He states that it is impossible for any person to grasp the objective reality or even the objective truth. It is only possible to mirror some aspects of this reality due to certain models (von Bertalanffy 1965). A system or organization can only be translated into models through a certain perspective, and such logic follows the concepts of second order cybernetics (von Foerster 1979) and radical constructivism (von Glasersfeld 1995).

On the premise of understanding organizations, system theorists have developed a variety of concepts that explain the behavior of organizations and the interaction of organizational members. There is, however, one concept that focuses purely on the description of an organization according to its input from the external environment and its output to, or the reactions of, the external environment. The inside of the system remains unknown and is described as a *black box* (Luhmann 1991): "The constitution and structure of the box are altogether irrelevant to the approach under consideration, which is purely external or phenomenological. In other words, only the behavior of the system will be accounted for" (Bunge 1963: 346). The observer is not able to see inside of the black box. The term 'black box' is also used in the context of programming. During what is called the 'black-box-test', a tester is to determine whether or not a piece of software is performing according to specifications, without

knowledge of the software's inner workings. The opposite of this procedure is the 'white-box-test' and is conducted by the programmers involved in the development of the software. They possess knowledge about the code and are able to see the inside of the (white) box. The concept of the black box has become an integral part of today's programming culture.

These concepts from programming have an impact on the perception of big data. There is currently a tendency to put big data into a black box (Pasquale 2015), mostly because they seem too complicated to understand. Luhmann (1991) noted such complexity as another reason for black boxes. Nevertheless, the need to open the big data black box is urgent, especially as big data are currently capable of influencing every organization. Data are put into the black box that is big data, and completely new data potentially emerge from it. Especially in the context of organizations that are also potential black boxes (Sirmon et al. 2007), the interactions between both black boxes will seem difficult to follow. If big data construct a new type of reality and are on their way to becoming truly ubiquitous, seeing big data as a black box we will likewise place the entire system, and everything, inside of an enormous black box. With everything inside said box, everybody will also be inside which makes focusing on the input and output stimuli impossible. Although big data are currently observable to a certain degree as a black box, the interaction and especially the diffusion in, or fusion with, society will make these observations more and more complicated – at least to a certain degree. Ultimately, being within the big black box will create the necessity of *dealing with big data* as a white box.

The ability of an organization to self-organize (Nicolis & Prigogine 1977) and its capability of autopoesis (Maturana & Varela 1972) is connected to systems theory. Self-organization is the ability of an organization to achieve order out of itself. In a team, for example, self-organization is conducted by its team members. Self-organization will eventually lead to some form of spontaneous order (Kauffman 1993) and is presumably faster (Weick 1979). Autopoiesis is the potential of an organization to renew itself. Such an organization is autonomous from other organizations and capable of surviving due to its structure (Froese 2010). Both concepts assume that organizations or team members are independent and act freely, without influences from the outside. There are, however, restrictions to those concepts. Self-organization can be externally induced (Pongratz & Voß 1997, Stein 2000) and means that an organization can create structures that support self-organization and decrease centralized steering (Gomez & Probst 1980). Bounded autopoiesis (Scholz 2000) involves restricting the absolute autopoietic potential of an organization by putting in place a certain regulation rule that keeps the organization coherent.

Both concepts describe the behavior of big data quite well. Big data are in some form self-organized and autopoietic in one way or another. Data within big data interact with each other freely and will potentially find some form of order. In addition to that, big data constantly generate new data to achieve self-renewal. Although big data are capable of self-organization, they depend on a technological structure. People and machines generate data and interact with them, data will not create themselves without any external influence which renders self-organization

impossible. Big data are also constantly renewing themselves. New data are added to the stockpile of data, but the blessing or the curse is that big data never forget (Solove 2011). This constitutes the need for some regulatory law that supports interaction with big data. Time, for example, may be a reasonable regulator, especially in the context of corporations. Corporations benefit greatly from big data, but the use of outdated data may cause serious harm to corporations. Although big data can self-organize and conduct autopoiesis, when interacting with organizations, the organizations have an interest in influencing the self-organization and autopoiesis in order to benefit from big data. Organizations want to *impact* and *interfere* with big data in order to *control* and *manipulate* the relevant portion of big data in their own interests.

Although there are many intersections with cybernetics, it is systems theory that underlines the claim for an inability to achieve objective truth. Nonetheless, big data represent a challenge to systems theory. Due to the ubiquitous amount of data, big data mimic a type of 'whole', but being so large will not fit into any kind of black box. Systems theory strengthens the idea that big data, as something that is everywhere, need to be researched and understood. Putting big data into a black box will not be sufficient, as the input and output of this big black box is also difficult to observe. Systems theory says that the internal structures of big data need to be observed in order to understand the effect of big data on any system. Systems theory also contributes to the idea that big data in their vastness cannot be understood, influenced or even changed by any system at an organizational level. Such a system can influence its perspective on big data. Big data that are relevant for organizations can be affected and interfered with. Such a task is also in the inherent interest of organizations: interfering in such a regulating way will help to harness the relevant portions for this system of big data. Such a system *acts proactively* and does not react to the output with which big data present it.

2.3.2.3 Big Data in Population Ecology Theory

The following theory focuses on the evolutionary approach to organizations. Rooted in the theory of evolution (Darwin 1859), population ecology theory considers the population of organizations and their battle for survival and, consequently, the survival of the fittest as a guiding rule. The survivability of an organization is only one aspect; more relevant are the evolutionary processes and the question of why some types of organization are more fit than others. Population ecology borrows the ideas of variation, selection, and retention or diffusion from biological evolution (Aldrich et al. 1984). Aldrich et al. (1984) describe these three stages as follows: *Variation* always takes place if a new organization is created, as this newly formed organization is influenced by existing ones and will blindly or purposefully vary from those organizations. *Selection* takes place because some organizations are more fit for the environment than others. Those organizations are able to acquire sufficient resources from the environment and will survive; other less fit organizations will have access to fewer resources and are bound to fail over time. This is a selection process

that thins out the population and favors certain types of organization. *Retention* or *diffusion* concern the preservation of certain types of knowledge. This process not only focuses on the organizations as part of the population, but mainly targets the members of such organizations. Knowledge will be passed down in some way from existing members to new members, if it appears to contribute to the survivability of the organization. Aldrich et al. (1984) added the principle of struggle for existence, as organizations face fierce competition, although the time span of this competition as well the effects on an organization are more than decades long (Aldrich 1979, Hannan & Friedman 1977). The authors reason that organizations act as though they are in an evolutionary competition; observing this over such a long period of time, however, may be difficult. The concept introduces time as a factor for organizations, comparable to lifecycles (Hurst 1995). Organizations change over time in one way or another; organizations are created and will potentially die.

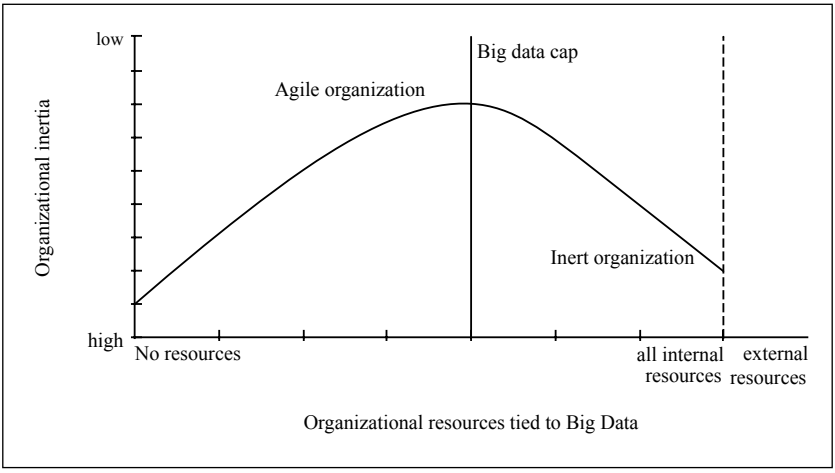
Big data can also be seen as a temporal construct. Big data and organizations are always in a relationship: one generates the other, which influences the first, and so on, until the starting point is no longer known. Big data, thus, underlie an evolutionary process. Data are generated and selected. If data are not objective, however, there will be errors in the analysis (Lazer et al. 2014). Big data mutate over time. On the one hand, mutations could be driven by data and lead to an autonomous evolution, which I propose as data-driven mutation. On the other hand, mutations could be described as organization-driven mutations. Both mutations are possible and probable; data-driven mutations, however, are currently popular and seem to be more adaptable to the environment (Provost & Fawcett 2013). Over time, there is the chance of a convergence of organizations as well as their fossilization. Data-driven mutations depend on big data as the source of selection, variation, and retention. As a result of big data, organizations become more specialized for a certain environment and make such organizations susceptible to environmental changes. Big data reinforce certain structures as big data also favor a standardization or normalization (Scholz 2015a). All of those factors add up to a hardening of the structures in organizations, because big data will act as Occam's razor (according to Wittgenstein (1922)). Here, data-driven mutations focus on plausible explanations while eliminating improbable ones. This is comparable to evolutionary degeneration or a reduction in variation in the population. Changing structures is difficult, and if reality is constructed around those structures, they turn into shackles.

As with biological evolution, organizations are sometimes not fast enough at adapting to new changes in the environment and are generally speaking not the fittest contestants in the population. In 1984, Hannan and Freeman developed the concept of *structural inertia* which refers to organizations that are not capable of understanding and predicting changes in the surrounding environment (Hannan & Freeman 1977), or unable to change internally due to a certain path dependence (Sydow et al. 2009). Such organizations are inert and will have lower potential survivability than more agile organizations. Although Hannan and Freeman (1984) propose that organizations rarely change, this claim is criticized by March (1981) as well as by the emergence of the research field of organizational change (Todnem By

2005). On top of that, the idea of differences in the capacity to change clarifies the reasons for the potential death of organizations (Freeman et al. 1983).

Introducing big data into the concept of organizational inertia will, at first, seem like a contribution. Big data support organizations with a variety of information, and may allow them to simulate or predict changes in their nature. Methods such as predictive analytics (Shmueli & Koppius 2011) allow organizations to prepare for changes in the environment and will lower structural inertia substantially but, having access to a huge stockpile of data concerning their environmental nature, will lead to more elaborate models and various different simulations. There is a natural complexity barrier (Gros 2012) within any of these big data models. One example is big data contributing to weather predictions (Hampton et al. 2013). Beyond a certain threshold, however, we see diminishing returns from using resources to deal with big data. It is possible to invest more resources into big data in order to get better weather predictions, but the amount of resources tied to improving weather predictions outweighs the outcome. Tying resources to predicting changes in nature will make organizations less agile, as organizations need more resources to foster their agility with big data. Big data will, therefore, make any organization more agile until a certain big data cap, beyond which any organizational resources bound to dealing with big data will decrease in agility and lead to an increase in structural inertia. Big data decreases inertia at first, but will eventually lead to an increase in inertia as shown in Figure 4.

Figure 4: Organizational Inertia and Big Data Cap



Population ecology theory supports several implications of big data and focuses predominately on the temporal perspective of big data. It seems that big data, at least in the short term, can have very beneficial effects on most organizations. Over

time, however, these effects are subjugated to some form of diminishing return. As a matter of fact, big data may even make an organization less survivable when flowing unregulated into organizations. More and more resources are required to deal with big data which will generate a *data desirable structure*. Data-driven organizations are more specialized, but at the cost of their differentiation in structure and their potential agility. Population ecology theory identifies positive effects of big data to a certain point. At this tipping point, the potential benefits of big data are at their maximum. Above that threshold, however, big data become harmful to the survivability of an organization. Big data can be compared to oxygen: any living organism needs oxygen to survive, but an excess becomes toxic. Big data are essential for the survivability of any modern organization, but from a certain point onwards, big data are lethal for any modern organization.

Subsequently, the question arises: How can an organization discover this tipping point? It is essential to highlight that an organization cannot use big data for the discovery of such a tipping point, because any more resources allocated to big data will shift the organization further towards the big data cap or even worse beyond the big data cap. Therefore, the big data cap will be monitored by other means. This could be a monitoring by the respective experts within the organization and the observation of relevant indictors. Furthermore, the measurement does not require infinitesimal accurateness but ranges in certain intervals. The organization will already change in agility on the way to the big data cap, so the expert panel will perceive a diminishing return. The goal is to prevent the organization from reaching the big data cap and it will be sufficient to avoid a certain interval before this big data cap, in which agility is slowing down and/or decreasing.

2.3.2.4 Big Data in Complex Systems Theory

At the moment, many organizations are trying to solve problems using the classical playbook, and are focusing on simplification, predictability, equilibrium and linearity (Marion 1999). Barabási indicates the inadequacy of such an approach as follows: "As companies face an information explosion and an unprecedented need for flexibility in a rapidly changing marketplace, the corporate model is in the midst of a complete makeover" (2003: 201). Organizations need to move beyond reductionism (Barabási 2012) to a world where change is the new stability (Farjoun 2010). Complex systems theory focuses on unpredictability, non-equilibrium and non-linearity (Maguire et al. 2011).

The field of complex systems theory (or complexity theory) has a long history and is heavily influenced by cybernetics, systems theory, and evolutionary theory (Merali & Allen 2011). Although this theory seems like a loosely connected conglomeration of various concepts picked from different theories, the common notion or understanding of complex system theory is explained by Lissack, in that "within dynamic patterns there may be an underlying simplicity" (1999: 112). Scholz (2015b) points out that, as an organizational theory, complex systems theory has evolved

from a "remarkable new vista" (Anderson 1999: 229) to "its time to change" (Andriani & McKelvey 2009: 1068) within recent years. Eoyang (2011: 320) even goes so far as to question everything we know: "Everything that supported stability and continuity of organization [is] compromised".

The field of complex systems theory is researched by numerous researchers, and there is a European school and a North American school, and many disciplines influence the field (Maguire 2011). For that reason, there is currently no concise definition available. There is, however, unanimity regarding the features of complex systems. Many researchers (e.g. McKelvey 2004, Sullivan & Daniels 2008, Maguire 2011) cite the description of Cilliers, concerning complex systems. He lists the following ten features:

1. "Complex systems consist of a large number of elements
2. A large number of elements are necessary, but not sufficient
3. The interaction is fairly rich, i.e. any element in the system influences, and is influenced by, quite a few other ones
4. The interactions are *non-linear*
5. The interactions usually have a fairly short range
6. There are loops in the interactions
7. Complex systems are usually open systems
8. Complex systems operate under conditions far from equilibrium
9. Complex systems have a history
10. Each element in the system is ignorant of the behaviour of the system as a whole, it responds only to information that is available to it locally" (Cilliers 1998: 3–4).

Those features are moving away from the general idea of reductionism and linearity. Their more complex direction is beneficial when it comes to big data. Big data consist of many elements and, even though the variety of elements may not be huge, their impact is ample. Big data, organizations, and especially the members of organizations, interact constantly, and this interaction is truly intensive. For Cilliers, the aspect of non-linearity is of utmost importance which is why he himself put it in italics. He explains that any large *and* linear organization will eventually split into similar but smaller organizations. Large organizations exist due to non-linearity. He reports that non-linearity "guarantees that small causes can have large results, and vice versa" (1998: 4) and implies phenomena like the butterfly effect (Lorenz 1963). Cilliers' fifth principle denotes the instance that interactions are of short range. To put it into context, big data may not directly influence a member in an organization, but the effects of big data are often the result of someone handling data within organizations (Rubinstein 2013). In addition to that, the system displays loops of interaction. Big data influence organizations as much as organizations influence big data, so there is a constant feedback loop in a complex system that is infused by big data.

Generally speaking, big data depend on the idea that an organization is an open system. Big data are big due to the idea that all data from every source are available.

Cilliers (1998) then compares equilibrium to the death of an organization. Although this may be a bit far-fetched, big data represent a constant source of disruption. Complex systems remember their history which is even more true for big data. Big data will remember everything that is collected about an organization. The information about such organizations is available for eternity. History may be ignored, but that decision would be made by organizations. Finally, Cilliers specifies that the elements in a system are nescient to the behavior of the whole system. This is reasonable because any element that understands the whole would inherit all the complexity of such a system. In terms of big data, no member of an organization will completely understand the wholeness of big data, or the impact of big data on an organization. Cilliers emphasizes a previous claim in a different context: only local (or organizationally relevant) big data are of interest to an organization, and organizations are only capable of dealing with those portions of big data. That means that within complex systems theory, big data cannot be completely grasped by any system.

On the premise that complex systems theory is rooted within the theories presented earlier, those concepts can be recognized within concepts of complex systems theory. They are expanded in certain ways and are part of advanced concepts tackling the same phenomena. As shown in table 8 and explained below, many of the concepts have a certain counterpart in complex systems theory. They may sometimes not precisely tackle the same phenomenon in an organization, but they are capable of describing the interaction between big data and an organization in more detail.

Table 8: Inclusion of Organizational Theory Streams in Complex Systems Theory

Theory	**Understanding Big Data**	**Expansion within Complex Systems Theory**
Cybernetics	Law of Requisite Variety	Complex Entropy
	Homeostatic	Homeodynamic
	Second Order Cybernetics	Third Order Cybernetics
Systems Theory	Black Box	Emergence
	Self-Organization	Self-Organized Criticality
	Autopoiesis	Fractals
Population Ecology Theory	Selection, Variation, Retention	Adaptation and Co-Evolution
	Organizational Inertia	System Fitness

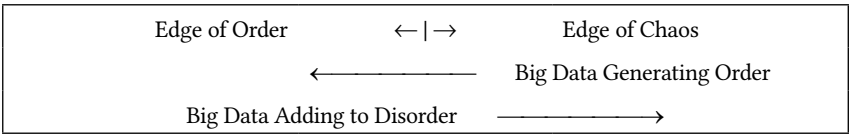
Complex entropy. The first approach only vaguely fits the law of requisite variety, but it tackles the situation of how the system deals with variety in the environment. There is a metaphorical link between the law of requisite variety and complex

entropy. The theory, however, moves away from the interaction between disturbance and response towards order and chaos (Whitfield 2005), or order and disorder (Morin 2008). Morin compares entropy to disorganization and uses negentropy (Brillouin 1953) with reorganization. He explains a certain paradox concerning entropy, in that the universe has a tendency to entropy (maximal disorder) on the basis of the second principle of thermodynamics, but that the universe also seeks to organize itself (maximal order). In the reductionist view, there is the assumption that we either observe order or observe disorder, but complex entropy allows for a system to go beyond such limitations. Morin claims that "for 'either/or' we substitute both 'neither/nor' and 'both/and'" (2008: 33). The paradigm of complexity for the author is not, therefore, the assumption that order and disorder are logically contradictory, but that order is linked to disorder. Order emerges from disorder and disorder is born in order. This view may not, however, be confused for a deterministic one, the question of who determines whom is irrelevant; what happens in the conjunction of order and disorder (Morin & Coppay 1983) is what matters. Complex systems theory is, therefore, concerned with the intersection at the edge of order and the edge of chaos (Waldrop 1993), thus tackling the first critical value (Bradford & Burke 2005) and the second critical value (Beinhocker 1997). Morin highlights the importance of linking apparently contradictory concepts when observing organizations:

> "If we think already that there are problems of irreducibility, of indeductibility, of complex relations between parts and whole, and if we think moreover that a system is a unit composed of different parts, one is obliged to unite the notion of unity and that of plurality or at least diversity. Then we realize that it is necessary to arrive at a logical complexity, because we should link concepts which normally repel each other logically, like unity and diversity. And even chance and necessity, disorder and order, need to be combined to conceive the genesis of physical organizations […]" (Morin 2006: 9).

The conceptual grasp of this notion of big data is that they are simultaneously in a state of order and disorder. Any organization that tries to gravitate around the tipping point between the edge of order and chaos can use big data to achieve some sort of orbital stability. However big data are not inherently orderly or disorderly, but resemble a dynamic system that is influenced by people using big data, with big data changing themselves through algorithmic evolution. This means that either somebody external can enforce a predefined order, or big data can discover a spontaneous order. This order follows a certain rule and, as a result, will not be objective but subjective. Assuming that this order is subjective implies that disorder is also subjective. Consequently, both order and disorder will constantly compete against each other to achieve a certain type of order and disorder within big data. Big data within organizations will also follow such a process, trying to order and disorder the organization. If balancing between the edge of order and the edge of chaos already imposes a challenge in itself (Waldrop 1993, Marion 1999), however, having such strong and sometimes rampant force will bring about a destabilizing power that moves organizations far away from the 'sweet spot' shown in Figure 5.

Figure 5: Big Data as a Destabilizing Power for Order and *Disorder*



In a recent study on information dynamics in social media, researchers discovered that within the diffusion of information, there is a phenomenon they called the "order-disorder-transition" (Borge-Holthoefer et al. 2016: 6). They propose that, if information is spread far enough, any information network will transform from centralized to decentralized, and consequently, shift from order to disorder. The researchers assume that such social networks are not controlled or steered by an external force. Big data, therefore, possess the ability to add disorder in big data and within an organization. Big data are able to move an organization away from the edge of order. Rätsch (2015) discusses the potential of big data to prevent innovation and lead to an organizational stalemate. Picking up his argument, organizations using big data for the sake of order become more average. In a sense, being average is not a bad thing (Scholz 2013a), but it will lower disorder and eliminate variety in organizations. Furthermore, such developments reinforce themselves, if big data suggest that a certain structure is beneficial and big data obtain beneficial results, the organization will be forced to follow its path, structures will become shackles (Scholz 2015a), thus generating a data-driven structure. Big data will generate order and move organizations away from the edge of chaos. Using big data deliberately is a premise to keep an organization in orbit around the joint between the edge of order and the edge of chaos.

Homeodynamic. If an organization can only gravitate around this point of order and chaos, it will cope with ordering and disordering forces, but achieving a form of homeostatic steady state appears impossible (Lloyd et al. 2001). To conquer these obstacles, Yates (1994) developed the concept of homeodynamics, a concept that has some similarities to dissipative structure (Prigogine & Stengers 1984) and homeokinetics (Soodak & Iberall 1978). Trzebski (1994) describes the main difference between the concepts as follows:

> "Homeostasis (is) state oriented homeostatic steady state, stability close to equilibrium, Program (set point)-driven system. Homeodynamics (are) rate-oriented homeodynamic stability, not very far from equilibrium, fluctuating and oscillating or close to 1/f noise informationally, not fixed program-driven systems with easy generation of new activity patterns" (Trzebski 1994: 111).

Organizations try to achieve a certain type of homeodynamic stability (Scholz & Reichstein 2015): stability here does not refer to a steady state, but rather to the actual idea of stability. If organizations deal with external changes and are able to self-organize internally, stability becomes a dynamic concept. Farjoun (2010: 203)

describes stability, and reliability, as follows: "In their more dynamic sense they can also be viewed as long term efficiency and robustness against failure and persistent perturbations […] and they therefore require variation to sustain". He, therefore, contradicts the idea of a steady state and implicitly proposes a more homeodynamic view of organizations.

Talking about big data and a steady state or equilibrium also sounds paradoxical. Big data are, above all, dynamic and largely imbalanced. The only steady thing about big data is their exponential growth. But even the exponential will change one day, due to the complexity barrier and diminishing returns (Gros 2012). For big data, the steady state is growth. For an organization this means that it copes with the influence of a growing amount of data. In order to achieve some form of homeodynamic stability the organization changes so as to oscillate near the equilibrium, transforming or filtering the increasing big data noise into something useful. Homeodynamics are, therefore, achieved by dampening or even amplifying the effects of big data in a way that causes organizations to achieve a (temporary) homeodynamic balance.

Third order cybernetics. In cybernetics there is already a differentiation between first order and second order cybernetics. Recent years have seen discussions about a third order cybernetics (e.g. Boxer & Kenny 1990, Johannessen & Hauan 1994, von Foerster 2003). The discourse about third order cybernetics is closely entangled with the emergence of virtual reality concepts such as the internet and other media (Kenny 2009). Although the singular term virtual reality is defined more loosely than its plural virtual realities such as World of Warcraft or other video games, the general idea is that people face life or reality in a virtual environment. Big data contribute massively to the data-constructed reality that is happening in this virtual reality. Kenny (2009) explains that nothing is real and that it no longer seems possible to identify the observer. He also says that it is questionable whether the observer has seen anybody or just observes phenomena derived from data. Big data allow for more granular data to be gathered about individuals, but always correlate to many other people (Tene & Polonetsky 2013). In order to generate Amazon recommendations, for example, the system takes a user's recent purchase and compares it to those of millions of other customers and the products they ended up buying. Users simply mimic the behavior of others and others mimic theirs. So, is the system really observing individuals or just a mass of people? Big data make this differentiation more difficult. Kenny (2009) asks if anybody even knows their observer. We are now living in a world of mass surveillance and 24/7 connectivity. We are observed at all times, but do not know by whom. Similar to the post-panopticon (Baumann 2000), we are well aware that observation takes place; but that is all.

Kenny (2009) proposes third order cybernetics in the sense of big data, an expansion already described by the example of big data. It is essential, however, to emphasize that big data engulf the idea of an observer. Everybody observes everybody and everybody is observed by everybody. Big data are also never real, neither in actual nor in virtual reality. Big data construct a subjective reality in both real and virtual worlds. These two constructed realities are not necessarily identical or

comparable and may resemble the idea of the presentation of self, as proposed by Goffmann (1959).

Emergence. Black boxes are evidently incapable of grasping big data, as big data are something completely new and stimulate a massive increase of thinking. Big data cannot be put into any existing black boxes. This newness is the source of new knowledge, new ideas, and change in a system exposed to big data. Consequently, any system connected to and influenced by big data will change over time and generate new patterns. Some form of emergence is inherent. Emergence is not a new concept and in terms of social systems can be attributed to Durkheim (Bellah 1959). Sawyer (2002) analyzes Durkheim's contribution to emergence research in social systems. From the interaction of individuals, Sawyer explains, some form of patterns emerges at the social level. It is what McKelvey (2016) describes as the bottom-up emergence of strategies (Mintzberg & McHugh 1985), ideas (Hamel 1998), networks (Feldman & Francis 2004), groups (Barry & Stewart 1997), hierarchies (Illinitch et al. 1996), or emergent innovations (Oster 2009). Emergence is a phenomenon that, especially in complex systems (Holland 1995), has a strong impact at the higher level, as the following quote explains:

> "There can be no sociology unless societies exist, and … societies cannot exist if there are only individuals" (Durkheim 1897/1951: 38).

The quote applies to big data as well: there can be no big data unless data exist, and data cannot exist if there are only *datum points*. Out of big data emerge big data and big data generate big data. It may be unclear, however, due to its complexity, what exactly emerges from big data. Big data have a generative capacity. Similar to the concept of generative grammar (Chomsky 1965), according to which grammar as a set of rules generates language, it is known that the internet is also capable of being generative (Zittrain 2006). New patterns will be able to emerge from big data as form of informational grammar.

Self-organized criticality. Such emergence also needs a form of self-organization which is a concept popular in complex systems theory (e.g. Kauffman 1993, Krugman 1996). Bak et al. (1988) extended the concept of self-organization with the idea of self-organized criticality. In a later book, Bak (1996) describes the concept as follows: "I will argue that complex behavior in nature reflects the tendency of large systems with many components to evolve into a poised, 'critical' state, way out of balance, where minor disturbances may lead to events called avalanches, of all sizes. […] The state is established solely because of the dynamical interactions among individual elements of the system: the critical state is self-organized" (Bak 1996: 1–2). He clarifies the idea employing the example of a sand pile. Adding sand to a sand pile is understandable and observable at first, but adding sand will make the system grow, causing it to eventually establish a dynamic of its own. Avalanches may occur, when more sand is added to the sand pile. Bak concludes from this idea that, although individual actions are understandable, they become complex when embedded in a bigger environment (e.g. nature). Individual and local interactions are still possible in such a dynamic system. Those small interactions, however, can

grow into big avalanches. As the system grows, these effects can become stronger, a phenomenon known as the butterfly effect (Lorenz 1963).

Big data as a system are huge and full of interactions. However, all interactions take place independently and autonomously. Local interactions between data take place everywhere and at any time. Big data are submitted to self-organized criticality. One example is high-frequency trading (Buchanan 2015) and the flash crash (Kirilenko et al. 2014) which was caused by a small error in one algorithm, and led to a massive stock market crash in 2010. Although the error was quickly discovered and corrected, the market did not rebound completely. Ultimately, one algorithm interacted with another algorithm and these local interactions disrupted the entire system. One of the advantages of big data is that they are fast and are found almost in real-time which means that there are many interactions between data within big data. Big data in an organization are also highly self-organized, they interact with members of organizations on a regular basis, and the sand pile of big data in the organization will grow over time within an organization. Self-organized criticality will also grow, rendering big data within organizations highly complex.

Fractals. Autopoiesis is often linked to the idea of self-similarity (Andersen 1994) and seeks from this similarity the potential for something to renew itself. Maturana (1987) criticizes the application of autopoietic behavior within social systems, but is intrigued by the idea that there is some form of governing rule. Some researchers (e.g. Eoyang 2011, Scholz 2015b) explain self-similarity within organizations using the concept of fractals (Mandelbrot 1977) and describe a geometric shape that, if split into parts, seems like a smaller copy of the whole. The most popular example of a fractal is the snowflake. Falconer (1997) even reports the possibility of generating a fractal iteratively through a non-linear equation. In organizations, fractals are used as a metaphor (Eoyang 2011) and refer to knowledge (Nonaka & Takeuchi 1995), ideas and innovation (Zimmermann, & Hurst 1993), and corporate identity (Bouchikhi & Kimberly 2003). Knowledge, ideas, and identity are quasi-fractals at the individual level and the organizational level. They have a certain self-similarity and are essential to the individual-organizational fit.

Fractals within big data are also more of a metaphorical concept. Big data at the societal level can be similar to those at the individual level. From a statistical perspective, there is a certain inherent self-similarity. Individual data are aggregated at a societal level and big data are consequently fractals, big data can become fractals due to this idea. Scholz (2015b) proposes the following argument, that if an organization uses a normal (Gaussian) distribution and, therefore, focuses on the average, fractals enforce a more centralistic view. The majority decides what those fractals look like. The minority accepts and adjusts their fractals. Big data have the capability of reinforcing such behavior. Big data are, therefore, not fractals, big data *create* fractals on the basis of a certain governing rule and which makes them *bounded* fractals.

Adaptation and co-evolution. Complex systems theory, in general, is heavily influenced by evolution theory. The main premise of selection, variation, and retention is dominant in complex systems theory (e.g. Holland 1995). The main change is

the understanding of evolution as a dynamic system. Organizations constantly adapt to internal and external changes (Siggelkow 2002) and establish an environment of co-evolution (Rindova & Kotha 2001). Individuals in organizations also influence adaptation and co-evolution (Stacey 2001).

Big data and organizations dynamically adapt to each other and co-evolve. Such behavior can potentially cause a rat-race and mimic the behavior noted in the red-queen hypothesis (van Valen 1973). Both systems constantly try to improve their survivability in competition with other opposing systems. Van Valen derived the term from Alice in Wonderland: "Now, *here*, you see, it takes all the running you can do, to keep in the same place" (Carroll 1991). Big data are ubiquitous and any organization has access to a vastness of data. There is an impending need to use big data and many organizations use them simply because other organizations do. Such an evolutionary race is already happening, and organizations are running in a certain direction unaware of whether it is the direction with the *highest* survivability. This concept is linked to self-organized criticality (Adami 1995). The current evolutionary path could lead to an evolutionary dead end (Takebayashi & Morrell 2001).

System fitness. Finally, there is the concept of organizational inertia and, in the context of complex systems theory, the ability or inability to react quickly to internal and external challenges is often referred to as the fitness of a system (Anderson et al. 1999). Evolutionary adaptation could eventually lead to an evolutionary dead end, but some populations are capable of changing direction completely. Such populations are fitter and can change dynamically according to changes in the landscape. Kauffman (1995) borrowed the term "fitness landscape" from Wright (1932) and theorized that some populations are more adaptable than others. This form of fitness is visualized as height in this landscape (Provine 1986). The higher a population, the fitter it is. Over time, such landscapes can change dynamically and something that was defined as being fit may become less fit.

After exemplifying the expansions of big data within complex systems theory as depicted in Table 8, this dynamic approach is especially relevant for big data in organizations. It may seem less cost-efficient to focus on one type of fitness with respect to big data. That could mean using only one type of big data analysis for all big data problems. Within a static environment, however, such an approach is evolutionally correct and will result in the fittest solution. Specialization trumps generalization in this fitness landscape, but within a dynamic environment the fitness landscape is dynamic and changes constantly. One solution may sometimes help, but may otherwise be pointless. An organization is, therefore, able to conduct a variety of big data analysis. However, organizations are also able to identify a fitting analysis for current respective evolutionary obstacles. Generalization trumps specialization in this fitness landscape. An organization that stays *homeodynamically agile* will be fitter than a highly specialized one.

Big data can be grasped by complex systems and, above all, reveal the need to deal with big data within an organization. Being complex, however, does not equate to the idea of making something complicated and does not denote a decision between reductionism or holism. As stated by Morin (2008: 56): "Complex thought does not

all reject clarity, order, or determinism." Complex systems can be governed by simple rules (Eoyang 2007, Sull & Eisenhardt 2012) even if the system is dynamic and flexible (Falconer 2002). I, therefore, follow the proposition of Farjoun, of moving beyond dualism towards a type of duality: "Duality resembles dualism in that it retains the idea of two essential elements, but it views them as interdependent, rather than separate and opposed" (2010: 203). This aligns with the demand by Morin (2008: 33) to "substitute either/or for both/and". There is a need to balance "both stability and flexibility, both continuity and disruption, both ties to the old and stretches to the new" (Eoyang 2011: 326). In summary, big data within an organization will tackle the order and disorder with drastic measures and will continuously influence organizations. Organizations, therefore, will find ways of dealing with big data. In order to gain a competitive advantage and an evolutionary lead from the use of big data, it is necessary for an organization to achieve a *homeodynamic stability* and stay *homeodynamically agile* in the context of big data within organizations.

2.4 Big Data at the Human (Resource) Level

2.4.1 Current Status of Big Data in Human Resource Management

Big data will have an extensive impact at the social level, the organizational level, and the individual level. Especially within an economic organization the effect of big data will transform the way people are working. Initially big data will change the way the HR department is working, and only after this change will the effects of big data influence every employee within the organization. Consequently, the impact at the human resource level precedes the impact at the human level, although the human resource level already comprises an influence at the human level – within the HR department.

Nevertheless, big data in HRM is currently underresearched (e.g. Angrave et al. 2016, George et al. 2014, Huselid 2015) and, although, big data will influence human relations (Harvard Business Review 2013) the current discussion is driven by practitioners rather than researchers. The relation between HRM and big data is quite interesting as HRM holds the competence to support human actors as well as the strategic potential to implement big data into organizations, although its technological competencies are currently underdeveloped (Stone et al. 2015).

In the context of big data and HRM probably the most cited case is that of Moneyball (Lewis 2004). The author discusses Billy Beane and his experience as the general manager of the baseball team "Oakland Athletics". The book represents a fitting example of big data in HR, because the players are the most valuable asset a sports team holds. The team's narrow budget forced the manager to search for different ways of acquiring talent. Using and analyzing big data, he managed to form a team that was unusual, but competitive and highly successful. He discovered new indicators to evaluate the performance of players providing a competitive advantage towards other teams. Today's baseball teams employ so-called sabermetricians in

order to level the playing field by analyzing empirical data (Baumer & Zimbalist 2014). Beane's competitive advantage is now available to every team in the league. Those approaches to the use of statistics have become popular, especially in team sports – for instance, in ice hockey (Mason & Foster 2007), basketball (Oliver 2004), and soccer (Anderson & Sally 2013) to name a few. This example reveals a strong focus on strategic HRM, consequently, there is a link between big data and strategic HRM (Angrave et al. 2016).

The field of strategic HRM emerged as a research stream in HRM in 1984 and can be traced back to the research by Beer et al. (1984) and Fombrun et al. (1984). Over the time the definition of strategic HRM changed and this progress is described in an article by Kaufman (2015). In his review article about the evolution of the term strategic HRM, Kaufman summarizes the central elements of strategic HRM as follows:

> "HRM as the people management component of organizations, a holistic system's view of individual HRM structures and practices, a strategic perspective on how the HRM system can best promote organizational objectives, HRM system alignment with organizational strategy and integration of practices within the system, and emphasis on the long-run benefits of a human capital/high-commitment HRM system" (2015: 396).

This synopsis highlights the integral role HRM plays in organizations and the general strategy. There is a fit between the work of the HR department and the strategy implementation within the organization; consequently, strategic HRM contributes towards the competitive advantage of an organization (Becker & Huselid 2006). Therefore, if the HR department is a source for strategic decisions and, by that, contributes to the competitive advantage, this HR department needs to have a high differentiation in its architecture (Lepak & Snell 1999). Furthermore, Becker & Huselid (2006) mention that in order to contribute towards the strategic direction and the potential competitive advantage, the HRM focuses on its system rather than operational tasks. The strategic goal of HRM is to contribute to a sustainable competitive advantage. However, the focus will not purely lie at the organizational level but also at the individual level (Gerhart 2005). Strategic HRM is, therefore, a link between the strategic direction of the organization and the impact of such strategy at the individual level.

Due to the reason that big data influence the organization extensively, strategic HRM will deal with those changes in a strategic way to generate a competitive advantage out of big data. It is important to highlight that in this case, the competitive advantage is generated by combining people with big data. Big data aligned with the current digitization resemble a paper by Lepak and Snell (1998) talking about the virtual HR department. Big data enable the HR department to have access to all the relevant information as well as communicate with every employee everywhere. Interestingly, they highlight that "perhaps the most dramatic impact of IT on structural integration within HR is its transformational role" (Lepak & Snell 1998: 220). Derived from that, big data will transform HRM – HRM, however, will exploit the technological potential of big data, in order to do its work in a more flexible, more dynamic, and more responsive way. Big data enable the HRM to strategically realign

itself, in order to transform the working environment for its employees. The focus shifts from operational tasks towards a more strategically oriented management of the organization and the relationship between people and big data. Technology is seen as a catalyst for the change of the HRM function (e.g. Parry 2014) and, therefore, big data enable HRM to focus on the strategic perspective and to create an environment for the employees that may lead to a competitive advantage.

Big data could lead to freeing up resources in the HR department that are currently used to do operational tasks. Tasks which can, potentially, be automated and would enable the HR department to focus more on strategic work. However, the current situation of big data in HRM is quite different. Although it seems obvious that big data will, predominately, require a strategic HRM and the Moneyball example highlights this necessity, current applications derived from Moneyball are on an operational level in areas like recruitment (gild 2013), talent management (Bovis et al. 2012), job performance (Armstrong 2012), and data-driven decision-making (Guszcza et al. 2013). Table 9 depicts further opportunities for the use of big data especially in the field of recruitment. Big data may aid in the search for candidates and provide insights into the recruiting process. The use of big data supplies additional benefits for workforce planning and the talent management of employees. However, big data are not seen from a strategic perspective in HRM. It can be stated that the strategic HRM perspective suffers neglect when it comes to the application of big data at the human (resource) level.

Table 9: Examples of Big Data in Human Resource Management Practice

Categorization by Armstrong (2014)	Operational HRM	Strategic HRM
People resourcing	**Workforce planning:** • Employee development planning (dm) • Labor management (Blue Cross Blue Shield, LBHF)	**Workforce planning:** • Demographic risk management (Deutsche Bahn) • Strategic personnel planning (EnBW) • Strategic workforce planning (ÖBB, Techniker Krankenkasse, Commerzbank) • Strategic scenario planning (REWE, Lufthansa, Bayer MaterialScience) • Scalable workforce planning (AccentCare)
	Recruiting: • Recruiting with focus on niche roles (Tripadvisor) • Recruiting with focus on experience-passion-fitness (fitbit) • Candidate engagement (Rapid7) • Candidate management (Recurly) • Intelligence & insight into candidates (red hat) • Candidate search efforts (StrongView) • Recruiting hidden talents (Taboola) • Recruiting and time-to-hire (Fitness First) • Candidate communication and employer-branding (Fitch Ratings) • Web-based recruitment management (Carillion) • Recruiting non-exempt employees (CARQUEST) • Hiring process system (Apollo Group)	**Recruiting:** • Establishing and maintaining talent pool (Gainsight)

Categorization by Armstrong (2014)	Operational HRM	Strategic HRM
	Talent management: • HR life cycle with focus on talents (Avaya) • Talent management system (Motorola, Nationwide, WakeMed)	**Talent management:** • Talent acquisition strategy (Advance Auto Parts)
Learning and development	**Human resource development:** • Learning management system (UAP) • Learning content management system (Potash) • Employee development program (NYC: Department of Education)	**Human resource development:** • Improving training attrition (JetBlue)
Performance and reward	**Performance measurement:** • Standardization of specific measures (Evonik)	
	Compensation and incentives: • Incentive management (Financial Service Company) • Compensation management (Exelon, Scotiabank)	
Employee relations	**Employee engagement:** • Employee engagement program (FRHI Hotels)	**Employee engagement:** • Change in morality and attrition (Nationwide Brokerage Solutions)
	Onboarding: • Onboarding process (H&R Block)	
Employee well-being	**Healthcare:** • Evaluation of healthcare costs (Wegmans)	

(Cases from Blue Yonder, Dynaplan, glid, Google, IBM Kenexa, PeopleFluent)

Nevertheless, neglecting the strategic HRM of big data will not be productive and could even be harmful (e.g. Peck 2013). Consequently, there is a certain research gap in the field of big data in HRM as shown in Table 10. The research in HRM currently struggles to transfer the basic research into some form of applied research and by that widening the research to practice gap (Huselid 2011). However, at the same time Anderson (2008) states that big data will work without any theory and, by that, proposing the sufficiency of data-driven applied research. At this point in time, data-driven applied research is dominating the field and there is no known theory-driven applied research. This becomes explicable in the fact that data-driven applied research is quicker than theory-driven applied research, especially as big data are still not sufficiently understood theoretically. Suggesting that theories are no longer necessary, and that with enough data valid results are possible regardless of theory, sounds compelling to many, especially to corporations. Therefore, data-driven applied research has a significant head start compared to theory-driven research. And it also explains the focus on operational applications.

Table 10: Hermeneutical Observation of Big Data in HRM

		Way of Gaining Insights	
		Theory-Driven	Data-Driven
Research Focus	Applied Research	?	See Table 9
	Basic Research	Scholz 2015a	Anderson 2008

Both approaches appear contradictory. Furthermore, the current dominance of data-driven applied research neglects strategic HRM and purely focuses on operational HRM. Therefore, every application of big data in HRM lacks a strategic fit (Scholz, C. 1987) towards the organization in any aspect. Consequently, there is currently in most organizations no link between HRM and big data strategies; however, such a link is essential to utilize all resources within an organization (e.g. Scholz, C. 2014a). Big data are decoupled from the strategic HRM, however, influence the operational HRM due to data-driven applications. Such applications lead to data-driven decisions and, therefore, are indirectly influencing the strategic HRM. Big data and strategic HRM cannot be separated and are highly linked, and whilst at the moment big data determine the work of strategic HRM, the task of strategic HRM is to strategically manage big data in HRM. The usefulness of such a function is highly debated (e.g. Cappelli 2015, Charan et al. 2015).

The current imbalance creates ground for the ongoing turf war within HRM. HRM is already facing an existential crisis (Ulrich et al. 2013). Data-driven applications take over several core fields of HRM and HRM is at the moment neglecting the chance to focus on the task of strategic HRM. Consequently, the role of HRM is shifting and its path is unclear.

2.4.2 Classification of Views

For the sake of terminological division of the two approaches, those supporting the data-driven approach shall be called *anti-guessworkers* and those following the theory-driven approach to be called *neo-luddites*. The two terms are not intended to be judgmental, but they describe a certain behavior or attitude of the groups.

Supporters of the data-driven approach do so in order to eliminate the "guesswork" (Evolv 2013) involved in HRM. Those people characterize the HR department as "being touchy-feely, but in the age of big data, it's becoming a bit more cold and analytical" (Walker 2012a). Block (as cited in Walker 2012b) even goes so far as to state that "software will supplement, if not supplant, many of the personnel decisions long made by instinct and intuition." Big data have finally led to HRM analyzing at least some data. By using distinct measures and metrics, it is possible to lower the employees' sick time, increase retention, lower attrition, and optimize payment (Walker 2012b). Another example is Google's Project Oxygen (Bryant, A. 2011): by analyzing data such as performance reviews and surveys, a team derived rules of leadership, such as being a good coach or having a clear vision and strategy for the team. Others have discovered that there is a correlation between the browser on an employee's computer and their performance (Economist 2013). These examples show that big data can tackle some important questions; it is essential, however, to select the right sense-making metrics (Bladt & Filbin 2013). A popular example is the aforementioned Moneyball example (Lewis 2004), a seemingly purely data-driven approach that led to the major success of the Oakland Athletics.

Contrary to the view on big data in HRM, the neo-luddites claim that HRM can work professionally without having to take such an intensively data-driven approach. The term luddite is derived from the anti-technological-progress movement in the beginnings of the industrial revolution (Baggaley 2010) and is picked up in recent years in a populist fashion in terms of automation, claiming for the "race against the machine" (Brynjolfsson & McAfee 2011) to be common. These neo-luddites, in the context of HRM, are especially offended by the fact that current HRM only uses guesswork instead of any distinct analytics (Lay 2012). They even accuse big data of "dehumanizing human resources" (Cukier 2013) in claiming that along the road of big data, humans will turn into nothing more than resources (Graham 2013). Decisions and processes will be outsourced to big data and the analysts. They also question whether or not the behavior and actions of employees can be sufficiently collected as data (Williams 2013). Even if the existing data are good, they will not necessarily lead to good decisions (Shah et al. 2012). The neo-luddites especially address the privacy aspects of big data in the light of the global surveillance disclosures in 2013, building up resistance towards the use of big data and reinforcing the HR-IT barrier. Claims that all data will be collected (Richtel 2013) and focusing on the data exhaust or the digital footprint (data that we leave behind) increase skepticism against big data. While vast amounts of data may be of interest for HRM, they make the employee transparent, and damage the trust between HRM and employees (Scholz, C. 2014b, Scholz, C. 2016), thus, therefore, destroying the key

to high employee morale (Graham 2013). The neo-luddites see parallels to Taylorism and call this new form of workplace surveillance "new digital Taylorism" (Parenti 2001: 26) or Taylorism 2.0 (deWinter et al. 2014). Some authors even go so far as to argue that data will be the resource of any knowledge and cognition (Anderson 2008). Big data create a *holistic* picture of employees and are already being used in determining an individual's use to an organization without that person having a chance to justify themselves – a procedure with striking resemblance to "Der Process" (Kafka 1925) in which the protagonist is prosecuted for a crime: although the crime he is charged with is unknown, the jury receives details of his life and consequently finds something incriminatory.

Even though both sides obviously exaggerate their claims and, apparently, strongly oppose each other, both contribute towards the *erosion* of the HRM function in an organization (Ulrich et al. 2013, Cappelli 2015, Charan et al. 2015, Stone et al. 2015). On the one hand, the anti-guessworkers are implementing data-driven structures that will eventually lead to the *obsolescence* of the HR department. Why does an organization need such a department, if everything it does can be done by a data-driven application, especially, if those applications are faster and apparently more precise (e.g. Brynjolfsson et al. 2011, Feffer 2015)? On the other hand, the neo-luddites contribute to a strengthening of the current HR-IT-barrier. The HR-IT-barrier describes the complications between the HR department and the IT department to communicate properly with each other. The HR department loses its connection to the organization and can no longer contribute to it. In these times of digitization, in particular, the HR department is also transforming into a more digitized function. Not using those new tools as well as limiting all dispositions of them leads to a decrease in usefulness of the organization. For this reason, both approaches seem to be insufficient in trying to grasp the relationship between big data and HRM.

2.4.3 Augmentation as an Alternative Path

The anti-guessworkers accuse the neo-luddites of being too human while on the other side of the spectrum, the neo-luddites blame the anti-guessworkers for being too mechanical. Both groups seem like specialists in their respective fields. Big data at the human level are simultaneously mechanical and human. The individuals in consideration are not a string of zeroes and ones but people and people following their instincts may potentially be beneficial or harmful for an organization. The emphasis here, lies on the word *may* as it reveals a certain uncertainty. Whether big data are good or evil is debatable, yet the question is far from constructive. Big data are here to stay and big data are shaping the reality of people. Big data are entangled with their lives, especially with their working lives.

Rather than denying the advantages of either side, I propose a different approach. Following the logic of duality, this approach shall be called *augmentation* approach. Augmentation derives from the Latin word *augmentare* and means to gain, add, foster, or increase. One thing is augmented by another thing and

becomes more, bigger, or better. In today's world (and fitting for the argument of this thesis), the idea of "augmented reality" (Azuma 1997) is a very popular one. This technology mediates the view of reality by adding an additional layer of vision to it. The perception or view of a person is enhanced by this additional layer (Graham et al. 2012). Users of augmented reality receive more information which enables them to improve their work. Using this as an analogy, augmentation seems to fit the case of HRM.

Augmentation also describes a certain direction of use. The human actor is augmented by technology to become *better* which does not exclude a data-driven approach rather than narrowing it down. A human is responsible for the decisions made and is only augmented by big data in order to make the best decision possible in a distinct case. There is room to be humane in certain cases, but also the potential of supporting decisions with big data. Big data and HRM can work together and their collaboration can be superior to either one working alone. The superiority of such collaboration has been proven in chess, in which the most successful combination is human and machine together (Kelly 2014, Ford 2015).

This augmentation also allows for big data to connect at the human level. People are capable of using big data for their purposes and are able to utilize and harness their potential. This is relevant to the use of big data at an individual level, but also on an organizational one. In the previous chapters it has become evident that big data are omnipresent and surround people and society, but both worlds seem to be separate from each other. Big data and society, however, are closely entangled and interact extensively. Implementing such an augmentation makes big data visible and usable for the HR department and, ultimately, for individual employees.

Interestingly, although Moneyball is often seen as an example of the superiority of big data in HRM, the current development shows a different path. Fears have developed that the big staff of coaches, scouts and others may become obsolete (Kim 2014). However, the use of the Moneyball principle had different outcomes. The Danish football team FC Midtjylland uses big data for their work (Biermann 2015) and has recently won the Danish championship for the first time. The result of this approach is that people view football differently. Big data have opened up new possibilities. Biased decisions by coaches or staff are debunked through big data which enables people to bring the best team together. Big data help to buy the best players for any budget, but money "can't buy team spirit" (Thomas & Wasmund 2011: 286). Big data augment and support people within organizations in making better decisions, but big data do not make people redundant in the process of decision-making. Biermann concludes that the competitive advantage of FC Midtjylland is not a result of big data but of the "synthesis of cold analysis and heart" (2015: 96). It seems that big data, at least in this case, are not the source of competitive advantage (contrary to the Moneyball case), but that the people are the competitive advantage. This may be even more true in a world where everybody has easy access to big data. Such a task sounds similar to the prime goal of HRM: *making people better*!