JOURNAL OF PRAGMATIC CONSTRUCTIVISM

A framework for conceptualizing data properties as input to the construction of facts

Frederik Koeppe
Aalborg University, Denmark
frederikkoeppe@yahoo.de

Abstract

Data-related literature is permeated with fuzzy, multi-dimensional and ill-defined concepts like big data leading to low theoretical or practical usability to actors trying to succeed in an increasingly complex data environment. To successfully construct facts from data, a data language is needed which instead is grounded in well-established, precise and unambiguous concepts and terminologies. To contribute to this objective, this conceptual study works out the multiple data characteristics described in literature to clearly define those, to set unambiguous label names (data classes) for them, and to classify them into data dimensions representing specific meta-perspectives relevant to actors. By reviewing a wide array of data-related literature and the use of five classification principles, 63 data classes and 23 data dimensions have been identified. They have, sometimes in detail, been discussed, labelled and defined using the three dimensions of meaning from pragmatic constructivism, and they have been located inside a data value model which illustrates the key processes between a measured phenomenon and data practices. The resulting classification framework provides a shared understanding of universal data concepts and contributes to a more unambiguous data language in theory and practice.

1 Introduction

Since the introduction of Microsoft's ChatGPT in November 2022, Artificial Intelligence has been on everybody's lips. So-called AI stocks like NVIDIA providing the computing power to train AI models and to run their related prediction use-cases have skyrocketed, triggered by massively increasing demand for AI solutions and the phantasy that AI will change our way of living as dramatically as the internet has around 30 years ago. The range of AI applications imaginable is believed to be enormous, ranging from better data analytics, prompt-based content creation (text, picture, video, music), content summarization, automation of a high range of human activities, monitoring, action suggestions, digital assistance, scientific discovery and more. What all these applications have in common is that they rely on a massive amount of training data, which, together with the model design, determines the degree of usefulness of its applications or validity of their predictions, and that they lead to practices, whose underlying reasoning are less understandable or visible to actors if at all. Why machine learning algorithms are predicting a specific outcome often cannot be answered even by those tech engineers having designed and trained the underlying model. Specific outcomes can be wrong, biased or amplify unwanted behavior and, therefore, can lead to decisions that are socially harmful and reinforce inequalities.

AI on the other hand, can also be used to purposefully and effectually create digitally harmful content with the objective of creating mistrust, social divide and confusion. AI-generated deepfakes are constantly getting more believable and realistic and AI-generated fake news article distribution is increasing (cf. Verma, 2023) and then shared through social media. This leads to the conclusion of the recently published report 'Earth for All', which was published under the Club of Rome organization, that "the inability to tell fact from fiction" is claimed to be the "biggest challenge in the world today" (Dixson-Decléve et al., 2022, p. 101), substantially caused by social media with the result that global collaboration to solve existential threats like global warming is getting compromised since even agreeing on simple basic facts is not guaranteed anymore. The relevance and increasing awareness of false or misleading information and statements scattered digitally throughout the internet is also evident by the increasing amounts of terms describing this new phenomenon, like misinformation, alternative facts or fake news. As it is uncertain what the data

signifies, it can be argued that these developments lead to higher risks of the existence of what in pragmatic constructivism literature is described as illusive elements since actors might act based on wrong beliefs of what is true and false (cf. Nørreklit, 2017, p. 30).

More specifically, pragmatic constructivism argues that intentional actors construct their relationship with the world through the use of language. However, not all languages are equally good at creating intentional results. Thus, four dimensions of reality need to be integrated to create successful action leading to intentional results (cf. Nørreklit, 2017, p. 32 ff.). These four dimensions are facts, possibilities, values and communication. Along those dimensions, the validity of digital produced information can be examined. Firstly, for effective decision-making, planning, and control, possibilities presented by the actors should not be mere fabrications of imagination but have a factual basis. Thus, information must be grounded in sound evidence. The establishment of facts requires that the language is crafted in accordance with principles for sound conceptualization. As values motivate choices, the actors' values must be within the range of producing valid information. Finally, communication as the integrative dimension helps participating actors to succeed in complex practices through interacting and cooperating with each other.

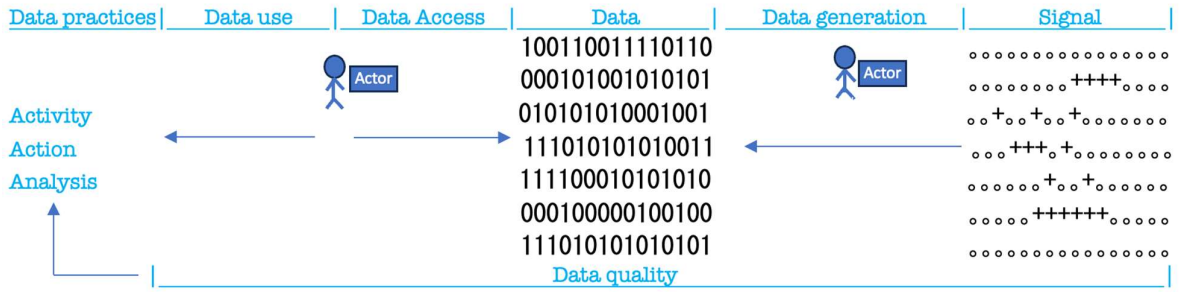
In general, pragmatic constructivism can help us to better understand functioning or failing practices and no recent technological development has had so much influence on human practices as data-based and digital tools, applications and devices. Being able to work with those is expected for an increasing number of employees, whereas data analysis skills are needed in almost all office departments today like marketing, procurement, accounting, sales or production. Those actors, however, often act at the end of the data value chain, meaning they are dependent on upstream processes designed by other actors. Those can be internal business intelligence departments managed by IT-affine employees, which are somehow accessible to the actor, but also hidden global actors operating disguised inside big tech corporations. Those corporations are very active in communicating the assumed positive application values of their tools, and therefore 'possibilities', but are mostly silent regarding their assumptions, principles, intentions, and beliefs worked into their models and, therefore, their 'values'. That means that values not only relate to those actors analyzing data for decision-making but also to choices made by tech engineers designing AI models, especially regarding what training data will be used and what moral or ethical key parameters will be manually added, since those significantly will impact the data content accessed by other actors. The importance of understanding values embedded in those applications became currently very visible with the criticism of Google's AI Model Gemini of being too 'woke' (cf. Chowdhury, 2024). As this example depicted a very easily visible value difference between Googles tech engineers and their user base, it shows the power of a few actors inside big global corporations able to dictate their worldview on a massive user base dependent on using their tools. Additionally, since those actors are outside the reach of most people, communication to those or influence on those are mostly impossible. In pragmatic constructivism, however, the importance of cooperation of actors through communication is emphasized to succeed in practice (cf. Nørreklit, 2017, p. 33). When, however, lagging communication as an intrinsic aspect of those new data practices cannot be overcome, the importance of actors identifying data characteristics as an indicator of possible benefits or harms on their practices increases in importance.

Figure 1 shows the data value chain with its three main data processes: data generation, data access and data use, which finally leads to practices. The starting point is always the motivation of an actor to generate data by setting up an infrastructure to capture a signal, signal variation, object or phenomena from the world. The term 'signal' in this context describes the transformation process of information about a changing state or behavior of real-world-objects and natural or technical phenomena like persons, animals, organisms, computer systems, machines or devices. Signals can for example be motions or sounds, which would then need visual- or audio-based sensors to be able to get observed and detected. By observing these signals, information about the object or phenomena can be gathered and actors can store the observed information as evidence in data. The interplay between the signal and the data is important because here there are always specific varying choices, limitations, applied methods or purposes in place which define in which form or characterization the observations or phenomena are stored in datasets. Datasets can have different dimensionalities, relationalities, granularities, origins, scopes, processing degrees, generation purposes and so on, which finally define how the phenomena are represented as data content inside the dataset.

When datasets are generated and stored, they need to be accessed before anybody can use them. The accessing actor does not need to be the same actor responsible for creating the data in the first place, and the requested datasets do not need to have the same content as those datasets of the data generation process since access limitations might be in place. Additionally, accessing actors might not access the same data compared to other actors with similar queries due to personalized systems favoring engagement instead of neutrality, leading to bubble effects. Therefore, challenges for the actor using the data especially exist when information about the preceding data-generating processes regarding infrastructure and involving actors are inaccessible, opaque or difficult to understand. Actors then might be misguided about what the data represents or if it fits their use cases. The risks of invalid reality constructions increase. That's why actors must educate themselves on data-related topics and be critical of what is presented to them. Data has become the universal transportation surface of information around the globe, and although consuming it passively for the average

uncritical actor has become easier, the risks of biased, harmful, dividing and incoherent reality constructions obviously have increased for society as a whole.

Figure 1: The data value chain



Therefore, this paper is concerned about the factual dimension of data especially the aspect of input data for AI or other data-based applications and the output data created by those tools. The content and characteristics of the input datasets define what those models are capable of, i.e. the factual basis of the data defines their possibilities, whereas the output data, in consequence, can have positive or negative real-life implications, i.e. value, on practices and society. It is therefore critical to understand the relationship between facts, evidence and data, especially the different forms in which evidence can be represented in data since data is not a neutral and objective medium representing the reality of the world but rather constructed and dependent on its social context and production infrastructure (cf. Kitchin, 2014a, p. 20 ff.). It is exactly the production infrastructure and context that can create misleading, untruthful, biased, manipulated, invalid, false or incoherent data representations, and identifying, estimating, describing and communicating those are pivotal for actors to reduce risks of unsuccessful practices emanating from AI or other data-based activities. Therefore, actors need data-specific concepts and frameworks which show up the different data forms in which evidence can be represented and adapt those to their specific use cases to construct facts which are trustworthy and reliable.

Generating good data-related practices, especially with AI, means understanding data first, and it is argued below that not enough work has been done to embrace the vast complexity and forms in which data can emerge. In literature a lot of data-related terms are used, however, often without defining them or delineating them to other data terms. On the other hand, probably the most famous data concept of 'big data' evolved into a term possessing dozens of definitions by now. Those definitions show that the current data landscape is complex, but they do not provide conceptual and understandable dimensions or perspectives with which we can better understand data and its relation to the phenomena the data is representing. Instead of using one term with dozens of confusing multidimensional definitions, which practically are not helping actors in their drive to understand how to succeed in their data practices, a framework describing the multidimensional characteristics of data through outlining distinct perspectives with which actors can practically analyze the validity of datasets in use, would increase understanding and cooperation between actors in their data practices. Therefore, this paper aims to enhance data-related language in theory and practice by delineating data by their individual characteristics they can possess regarding critical defining (meta-) perspectives or views independent of their specific content. To the knowledge of the author, such work has not been done before, and therefore, this paper aims to fill this gap by providing a conceptual framework in which dataset characteristics can be described, delineated, compared and evaluated.

The framework contributes by providing a comprehensive set of label names to differentiate aspects of data properties (views). Normally authors just use the notation of 'data type' to differentiate dataset properties. Depending on what is discussed, 'data type' can imply all kinds of different views, which makes a meta-view on different data properties based on this notation impossible. More precise communication between those actors providing the factual basis and the decision makers acting upon it requires a language that is able to describe the data landscape on which actors are trying to create successful actions. Instead of just accepting presented information as facts, decision-makers are able to understand the role and relevance specific datasets play in their reality construction. Only then decision-makers and data experts become co-authors in an interactive learning cycle and establish successful practices in a dynamic and complex environment in which they have to cooperate. By distinctively classifying datasets across the presented dimensions actors can identify weaknesses, shortcomings or risks in their current database or strategy by

helping to estimate data quality and by showing up alternative data configurations. Overall, it provides a basis for addressing the uncertainties in data production, with a reminder to be mindful of these when generating facts.

The paper is structured as follows: Section 2 provides core data concepts on which basis the research question is raised. Three main perspectives on how 'digital data' can be interpreted are developed, of which the computer-oriented perspective builds the main assumption and understanding of what data is for this study. Additionally, a review of the historical developments of data practices is presented, emphasizing the increasing risks of illusionary practices in today's environment, which cannot be addressed by fuzzy concepts like big data. Section 3 explains the methodology of the paper. It describes the principles which were used to generate the classifications and, therefore, the conceptualization process. Chapter 4 outlines all 23 data dimensions and all data classes in no specific order, while chapter 5 finalizes the framework by locating the data dimensions inside the data value chain model (figure 1). Chapter 6 concludes this study.

2 Development in concepts of data

2.1 The notion of digital data

To develop the data classification framework first it must be clarified on which general data definition these classifications are grounded on. This is done by outlining common theoretical definitions on the following terms or concepts: facts, data, datafication, digital, digitization and digitalization. Three theoretical data perspectives (representation-oriented, computer-oriented and signal-oriented) are worked out and it is explained why especially the computer-oriented perspective builds the foundation of the framework.

2.1.1 What is data?

In literature, definitions of data are manifold and often relate to other terms and concepts like facts, information, epistemological units, binary elements, organization assets and more. Historically, the first meaning of data derives from Latin and is 'given' relating to a fact of the physical world (cf. Mayer-Schönberger & Cukier, 2013, p. 78). When this fact is depicted in some physical form, we can interpret it as data representing the fact. This definition conforms with the ontological position of realism since it assumes that, independent of humans, there is an existing world out there. In this world, phenomena are predictable by identifying their cause-and-effect relationships, which then can be mirrored as factual statements. Positivism, which builds upon realism, then assumes that by applying scientific methods, facts from the world can robustly and neutrally be measured. This objectivity is the key prerequisite of a measurement that creates "*a bridge between reality, to which the object under measurement belongs, and the linguistic/symbolic realm to which the measurement results belong*" (Mari, 2007, p. 42) and is achieved by "*meaningful measurement concepts and scales outlined by qualities of content, reference, criteria and consistency*" (Mauro et al., 2024, p. 61).

Data as a 'representation of a fact or a collection of facts' conforms therefore with realism and positivism in that way, that facts are out there and just need to be empirically measured, collected and then stored in datasets. However, this realism view is often challenged by scholars with the argument that there is no objective reality out there and that knowledge and measures therefore are socially constructed. Pragmatic constructivism combines both extremes by differentiating between the world as a realist concept (everything that exists) and a reality, which is constructed by actors interacting with the world. Facts are not about mirroring ontological objectively existing things but involve an epistemological process that establishes a bridge between the language use and phenomenon of the world, which can be not only objectively existing but also subjectively existing, i.e. human constructs, feelings and emotions. They are constructed by actors, but not random since they require sound evidence to work in practice. Facts are one of four dimensions of reality (facts, possibilities, values, communication), which are needed for reality constructions that lead to successful practices. Therefore, instead of considering data as a 'representation of a fact', a definition in agreement with pragmatic constructivism would rather be that data can be understood as a 'representation of evidence'. From that evidence then actors can construct facts by accessing and analyzing the data. However, to be able to establish evidence from the world as data (i.e., to datafy), some form of structured language or coded system is needed to accomplish the transfer. At the same time, the actor accessing the stored data also needs the knowledge of how to decode the data to interpret its meaning since, otherwise, the recorded data creates no value. Semiotics is the scientific field addressing these sign processes between sender and receiver, which not only is used to store evidence but also to establish communication. To prevent misinterpretation and miscommunication between sender and receiver (i.e., the recorder and decoder), both actors need to agree on common codes for how evidence or meaning should be transferred. Therefore, based on this interpretation, data can be defined as coded evidence, which creates value for accessing actors with decoding knowledge. If the decoding knowledge is not present, meaning and information cannot be transferred as two persons cannot linguistically communicate with each other when each of them speaks a different language. We see that with this broad definition, data can be understood without considering computers or modern digital processing at all.

2.1.2 What is datafication?

Datafication is the first process step explained above meaning the recording process of signals and evidence from the world as data (i.e., to datafy). It is, therefore, the transformation process from real information into some structured form, leading to data representing the information with the goal to be recallable at a later time or place. Imagine a bird scientist going into the forest, observing birds in their natural habitat, and collecting the size, color, location and behavior of those birds in a tabular form on a piece of paper. For each instance of an observation, he uses one row and writes down specific codes for observed size, color, location and time, etc. in the respective columns. He datafies observed information on paper to collect evidence about the world. Other scientists with the knowledge about how to decode the tabulated data can understand the meaning of the data and gain conclusions about the real observations made.

Datafication often in literature is defined as to transforming a phenomenon into a quantified data format (cf. Mayer-Schönberger & Cukier, 2013, p. 15 and 78; Mazzocchi, 2015, p. 1255). This is achieved by the structured coding format in the datafication process. E.g., by adapting a tabular recording format, the bird scientists can afterwards count how many birds he observed, how many of those where black or calculate their average size. He will have problems to do this completely out of his memory. Therefore, the datafied information enables the quantification of observations, the interpretation of evidence and the knowledge transfer.

2.1.3 What is digital, digitization and digitalization and is data always digital?

The term *digital* is best explained in its comparison to its opposite, namely *analog*. *Analog* describes the capability of a phenomenon to exist in infinite states or variations, while *digital* describes a capability to only exist in finite states. Nature (uninfluenced by humans) per se is in an analog state. It can exist in endless variations of colors, sounds or smells, and it can variate endlessly between two time points. The reason for this is that color, audio or time are continuous dimensions, meaning that between two states, those change on a continuous, i.e., stepless scale. On the other hand, a digital representation of color, audio, or time has discrete scales meaning a limited number of possible states and is therefore quantifiable and uniformly measurable. The technical process of transferring analog information into a digital data format is called *digitization* while *digitalization* is rather a term used in politics, media or business to state the importance of "*potential changes in the processes beyond mere digitizing of existing processes and forms*" (Mergel et al., 2019, p. 12) since representations are increasingly getting purely digital as digital cameras, digital audio or biometric passports are becoming the standards of today. Computers can only read and process digital data since its foundation of processing data is based on binary code, meaning processing code sequences on the basis of only two different states (like 1/0, true/false, on/off, open/closed). This is the minimum discrete scale to store information since only one state could not differentiate or measure anything. A variable using these two states is termed a *binary digit* or *bit*, and since computers use these primary units of information, the processing power and storage capabilities of digital devices are measured in bits and bytes.

Table 1: Examples of general data perspectives

Examples/Perspectives	Representation	Computer	Signal
Bird observation recorded on paper	Digital data	Not data at all	Analog data
Bird observation recorded in Excel	Digital data	Digital data	Analog data
Analog camera picture	Analog data	Not data at all	Analog data
Digital camera picture	Digital data	Digital data	Analog data
Computer generated data	Digital data	Digital data	Digital data

This information processing computer-oriented perspective is what is defining the *digital* nowadays. However, the representation of the bird scientist observing information from the real world and coding it on paper in a structured manner would not be different from him typing the same data into Microsoft Excel directly. The reduction of complexity of the analog world into a specific limited number of states for each observation attribute (the columns) on a paper or on a computer only differs in the capabilities of a computer to read and interpret the data. Therefore, strictly speaking, we see *digital data* in both cases. I would call this the **representation-oriented-perspective**. In this perspective, the difference between *digital* and *analog data* is a question of if the transformation process has changed the analog continuous signal into a discrete format or not. The **computer-oriented perspective**, however, sees data as *digital* as soon as it is formatted into the binary code and is, therefore, computer-readable. Everything else is analog and not even data at all. Therefore, in this perspective, the observations on the paper are *analog* information not digitized into data yet. The third perspective is the **signal-oriented perspective**. This perspective describes data by the character

of the real phenomenon it represents. Here, for example, audio data processed by a computer can be labelled as *analog data* because original sound waves are analog. *Digital data* here can only be signals produced by computers themselves since all real-world signals by nature are analog.

We see that a simple differentiation between analog and digital can already have multiple perspectives. The framework will follow the computer-oriented perspective and, therefore, the notion of data always being digital since the goal of this framework is to differentiate and characterize datasets processed by digital devices and computers with which actors constantly must deal with in our current time.

2.2 The development in data practices and the 'big' problem of big data

The risks of illusionary data practices have increased historically. First, with the beginning of the computer age, possibilities emerge to save collected observations as datasets locally. Companies and scientists started to use this opportunity to collect internally generated datasets while control and understanding of the production infrastructure of those can be described as high. With further technological advancement, collaboration between scientists through data-sharing was possible, and the first openly accessible scientific databases emerged. Companies increasingly started to use externally generated datasets to complement the internally generated data to better understand customers, suppliers, competition, brand values and more. These externally generated datasets are not in the control of those actors directly, and therefore, questions of understanding and measuring the data quality, provenance, validity and biases of those datasets have gained much more importance. Now, with the start of AI producing output data, the way how output data are produced is even less understandable to those actors using the data, especially when the specific AI model's design, training data and embedded values are unknown.

Table 2: Historical phases of data-related practices

	Phases	**Visibility**	**Control and understanding**
1.	Use of internally generated datasets	Data production fully visible	High
2.	Addition of externally generated datasets	Data production visibility limited	Middle to Low
3.	AI generated datasets	Only output data visible	Low to none-existent

Although all three phases might nowadays take place simultaneously, it shows that good data-related practices become more challenging as we move into an AI-dominating environment. Along with the more challenging and complex periphery in which actors must act and interact, it becomes more important to be aware of the language and meaningful concepts to avoid illusions that lead to bad practices. This is, however, not what is seen in the current literature, where many fuzzy concepts are imbued with especially the one which grounded and preceded the AI hype, namely big data. Big data is discussed in countless papers, especially on how it will change professional activities (cf. Warren et al., 2015, p.397 ff.), which opportunities, challenges and risks it possesses or what it actually is or should be characterized, defined or conceptualized (cf. Akoka et al., 2017, p. 105 ff.; Wamba et al., 2015, p. 234 ff.). Rarely, however, do those papers go into detail about the specific characteristics datasets nowadays can possess, since big data is more used as a term to describe a trend in society or technology and not to distinguish concretely 'big' datasets from 'small' datasets. Over time, big data has developed into a buzzword, where authors couldn't put enough effort in to outdo each other in extending the original V-characterization of big data, which initially just was data with high volume, high velocity and high variety, with additional terms starting with the letter V like veracity (cf. Marr, 2017, p. 87; Janvrin & Weidenmier Watson, 2017, p. 3; Japkowicz & Stefanowski, 2016, p. 3), validity (cf. Shah, 2018, p. 40), value (cf. Marr, 2017, p. 87; Janvrin & Weidenmier Watson, 2017, p. 3; Japkowicz & Stefanowski, 2016, p. 3; Ylijoki & Porras, 2016, p. 77; Gandomi & Haider, 2015, p. 139; Chen et al., 2014, p. 173), variability (cf. Japkowcz & Stefanowski, 2016, p. 3; Gandomi & Haider, 2015, p. 139), verification (cf. Shah, 2018, p. 40) or volatility (cf. Shah, 2018, p. 40).

With those, it seems impossible to concretely label datasets as big or to distinguish them from smaller datasets since the term is open to highly differing interpretations and views. Table 3 shows big data definitions extracted by the researcher of this paper. When those are analyzed and taken apart, we can find groups of defining elements in those, which belong to one specific perspective of data qualities. Every group in the table can answer a distinct question (second column) about a dataset, like why the dataset was created, if the dataset is related to other datasets or how fast the data was generated. These perspectives, however, never get worked out in literature into a data classification

framework. The table also illustrates that the data can be produced and manipulated by different technologies that have consequences for what they are evidence of and what they can be deduced for.

Table 3: Big data definition groups

Definition groups	Questions the definitions can answer
"…collected through devices and technologies such as … increasingly, WiFi sensors, electronic tags." (Chua, 2013, p. 10)	What is the origin of the data?
"…text…" (George et al., 2016, p. 1493); "…videodata…" (Calvard, 2016, p. 67); "…videos…" (George et al., 2016, p. 1493); audio…" (Chen et al., 2014, p. 173); digital trace…" (George et al., 2016, p. 1493)	What is nature of measurement used to create the data?
"…size beyond the ability of typical database software tools to capture, store, manage and analyze." (Moffitt & Vasarhelyi, 2013, p. 4); "…multi-structure data…" (Moffitt & Vasarhelyi, 2013, p. 5); "…unstructured data…" (Chen et al., 2014, p. 171; Gandomi & Haider, 2015, p. 138), "…NoSQL…" (Ward & Barker, 2013, p. 2; Akoka et al., 2017, p. 106); …unstructured in nature…" (Kitchin, 2014a, p. 68); "…structured…in nature…" (Kitchin, 2014a, p. 68); "…structured…" (Ylijoki & Porras, 2016, p. 74)	How difficult is it for a computer to analyze the data?
"…high velocity capture…" (Akoka et al., 2017, p. 106); "…created in or near real-time…" (Kitchin, 2014a, p. 68); "…speed at which new data is generated…" (Marr, 2017, p. 87); "…speed and immediacy of data creation…" (Calvard, 2016, p. 66); "…speed at which data is generated." (Phillips-Wren & Hoskisson, 2015, p. 90)	How fast is the data created?
"…social media content…" (Calvard, 2016, p. 67); "…social media…" (Chua, 2013, p. 10); "…social media data…" (Chen et al., 2012, p. 1165); "…sensor … data…" (Chen et al., 2012, p. 1165); "…from sensors…" (George et al., 2016, p. 1493); administrative data…" (Arnaboldi et al., 2017, p. 764)	What is purpose of generating the data in the first place?
"…exhaustive in scope, striving to capture entire populations or systems (n=all)…" (Kitchin, 2014a, p. 68)	What is the relation of the data to the population of interest it represents?
"…challenge of managing data quality…" (Buhl et al., 2013, p. 68); "…inconsistencies…" (Japkowicz & Stefanowski, 2016, p. 3); "…messiness or trustworthiness of data." (Marr, 2017, p. 87); "…data accuracy and reliability of data…" (Janvrin & Weidenmier Watson, 2017, p. 3); "…quality of data…" (Japkowicz & Stefanowski, 2016, p. 3); "…uncertainty surrounding data integrity and trustworthiness…" (Phillips-Wren & Hoskisson, 2015, p. 90); "…unreliability…" (Gandomi & Haider, 2015, p. 139)	What is the quality of the data?
"…integrated from different sources and joint together." (Sparks et al., 2016, p. 33); "…embeddedness…" (Chen & Yu, 2018, p. 17); "…uniquely indexical in identification; relational in nature…" (Kitchin, 2014a, p. 68)	How relational is the data?
"…automatically machine obtained/generated…" (Moffitt & Vasarhelyi, 2013, p. 4 f.)	How automated is the data created?
"…big data is about predictions." (Mayer-Schönberger & Cukier, 2013, p. 11); "…draw inferences from correlations not possible with smaller datasets." (Moffitt & Vasarhelyi, 2013, p. 5)	Can the data be used for predictions?
"…cannot be managed by standard software…" (Chua, 2013, p. 11)	How restricted is the data format?
"…embeddedness, i.e., the places, the spaces within … social interaction are operated." (Chen & Yu, 2018, p. 17); "…temporally and spatially referenced…" (Kitchin, 2014a, p. 68)	Are there descriptive information saved in the data indicating background information about the dataset itself?
"…fine-grained in resolution…" (Kitchin, 2014a, p. 68)	How granular does the data represents its content?

We can see that the big data literature does not provide a conceptual framework or trustworthy language to describe, evaluate, classify or compare specific datasets in use, how to show up alternative sources of better-equipped datasets or how to evaluate AI-generated output data, and, therefore, they do not provide a basis to assess the validity of data in use. Good data-related practices need the development of meaningful measurement concepts to deal with the increasing challenges practitioners in that area face nowadays. The multidimensionality and messiness of current big data definitions and concepts do not provide this.

On this background, we raise our research question: What are useful perspectives describing distinct ranges of data characteristics with which actors can better understand and question the usefulness of data in their data practices? Identifying those perspectives and the fundamental manifestations of data characteristics relating to those perspectives would then contribute to a more universal data language in theory and practice.

3 Methodology

This study is designed to enhance our understanding of data in relationship to the construction of valid information for their effective application in human practices. Instead of just accepting presented information as facts, decision-makers should be able to understand the role and relevance specific datasets play in their reality construction. The study aims to unfold as a conceptual investigation aiming to classify datasets regarding their properties to facilitate a data language to guide the construction of facts. Understanding and assessing the quality of data requires a conceptual framework to link the evaluation of data in relation to the establishment of facts. In view of this, we need to outline quality criteria of concepts to produce facts along with the principles that can be used to generate the classifications of the elements (views) shaping the dataset, along with principles for what is needed to make reliable conclusions from a dataset.

Overall, the investigation draws on the criteria of good concepts outlined in the literature of pragmatic constructivism. Based on Nørreklit, 2017, concepts are cognitive structures that are used in the construction of reality (cf. Nørreklit, 2017, p. 24). Good concepts "*…can grab phenomena…hold on to them…describe and understand realities … and … consequently do things.* [They] *should be precise descriptions of realities, preferable with an unequivocal meaning, freed from ambiguities and the possibility of misinterpretations….*" (Henriksen, 2016, p. 31). For information to be effective in practice, its concepts must be outlined according to purpose along with being clearly defined and grounded. Although the degree of precision should be applied pragmatically in relation to context, the meaning of these concepts must be well-established and shared to function properly.

It follows from pragmatic constructivism that intentional actors produce the digital data, but the multicentric production of digital data poses important challenges in terms of what is the actor's intentions behind the digitally produced data. Accordingly, terms like 'big data' or 'data type', due to their subjective interpretational character, are not sufficient as descriptors to understand the realities in which actors must establish good data-related practices. We need more sophisticated concepts to grasp what might be behind them to evaluate their data quality for decision-making.

While pragmatic constructivism provides an overview of conceptual qualities for the measures, it does not delve into the specifics of data input production or numerical accuracy, and, therefore, we draw on relevant literature in relation to this aspect (Nørreklit, 2017). This is in line with pragmatic constructivism that advocates methodological plurality, where the choice of method should suit the phenomenon under investigation. Below, the methodology of the conceptualization process in relation to the following four processes and elements of analysis is explained:

- Process of collection data terms from a wide array of literature.
- Selection and grouping based on five classification principles.
- Setting label names and defining data dimensions and classes based on the three structural dimensions of Kure et al., 2017.
- Analysis of data dimension's locations inside the data value chain.

The different processes have been conducted in an iterative manner, therefore not in strict consecutive order. However, especially the five classification principles (step 2) have been followed to establish the classification whereas the three structural dimensions (step 3) have been followed in the final process of labelling, defining and describing the different dimensions and classes setup from step 2.

3.1 Process of collecting data terms

The collection process started with collecting and analyzing big data definitions (Table 3) in current publications, especially from Mayer-Schönberger & Cukier, 2013; Kitchin, 2014a; and Marr, 2017. By researching more data-related papers new data terms and definitions were additionally collected and stored. As the list grew over the years it was visible that groups of defining elements could be identified with which we could understand specific qualities or characteristics datasets can embody. But, grouping them into specific dimensions would require principles and test criteria to decide on which terms can be grouped together, or which terms should be excluded. Five principles were therefore developed to guide decisions made on what to in- or exclude from the framework.

3.2 Selection and grouping terms based on five classification principles

Meta-perspective: This means that all data classes should be able to answer general questions about aspects of the data production infrastructure, which could be applied to all datasets. Only then can the dimensions represent perspectives,

which can be applied to all datasets and give a comparable view of how to describe data characteristics from a meta-perspective independent of the specific data content. Therefore, data descriptions, which just describe its data content, i.e. weather data, should be differentiated from data descriptions which relate to a general data perspective. The labelling of data as weather data answers the question about its content (What is the data about? – It stores information about the weather), whereas labelling it as monitoring data can describe the generation purpose of why the data generated (Why was the data generated? – To monitor the weather). Data descriptions, which just describe its data content should be excluded from the framework.

Completeness: This meant that as soon as new data perspectives were identified, those were added as data dimensions. With this continuing process, a more and more complete overview of data properties could be developed, although, of course, total completeness can never be claimed. This principle, however, enables future expandability of the framework as soon as new perspectives come up or merging of views when two views are actually describing the same data perspective. Completeness also means that at least two classes need to be identified to span a dimension. When a dataset is labelled as possessing some kind of characteristic by an existing terminology, then there is also some other characteristic that this dataset does not possess. This other characteristic needs to be found and defined for a data dimension being considered as covering a specific view of multiple and at least two data properties as data classes. Sometimes those opposite data terminologies are not defined in literature. Of course, it is easy to say the opposing data to quantitative data is qualitative data, but what is the opposing data to metadata?

Relevance: Relevance means that inside a dimension only properties are worked out, which scientists and practitioners generally are going to face in their normal work. Too many classes in a dimension would inflate the framework unnecessarily. *Measurement nature* with eight first-level classes is the dimension with the highest number of classes. It is, however, possible that specific classes can be further broken down into subclasses.

Distinctiveness: Dimensions should distinctively represent a different view on datasets, therefore no redundancy with other dimensions should be present. Only by this, a possibility to identify a specific 'configuration' of a dataset, database or data strategy is given with which actors can discuss and evaluate the strengths and weaknesses of those building a factual basis upon it. This means that a dataset, database or data strategy could be described in its characteristic by a combination of classes from all data dimensions and, therefore, is 'configured' in a specific way. It can show up alternatives for managers and scientists in their search for valid and high-quality datasets or show up the need for a new data strategy or approach. It enables comparability between datasets.

Concreteness: This means that if specific terms of dimensions or classes are not precisely able to describe dataset properties and are somehow vague, they must be excluded from the framework. Big data is one of those terms.

Two other aspects to point out are that the differentiations inside a dimension are, for one thing, not always singled matched towards a dataset, and secondly, dimensions can have gradual transitions between classes. 'Not-singled matched' would mean that a dataset can possess aspects of multiple classes, e.g., a dataset could have raw data and processed data at the same time, audio data could have temporal metadata included or the dataset is used for descriptive and predictive uses at the same time. The classes, therefore, have rather the value of structuring a user's thinking process instead of creating absolute facts about a dataset. The second aspect ('gradual transitions of classes') means that data cannot always be described as either having one or another characteristic. A good example is granularity: A dataset can rarely be described as absolutely fine-grained or absolutely coarse-grained, but rather, there are different degrees of granularity persisting in datasets between both ends. Again, the framework should help to compare datasets or strategies in its configuration, and if a dataset is more or less granular, is often a point of perspective by comparing it to other datasets or strategies.

3.3 Setting label names and defining data dimensions and classes

For the framework to be meaningful in practice, the chosen data dimensions and data classes need to be clearly defined. For this, the three structural dimensions of meaning from pragmatic constructivism (Kure et al., 2017) are used. Those are abstract meaning, criteria-based meaning, and exemplary reference. Considering these criteria, this framework could be seen as a concept in its entirety; however, each data dimension and each class separately could be considered a concept as well.

Abstract meaning: Abstract meaning means that the content of the concept must be outlined. This is done by clearly defining each data class in comparison to the other classes in a dimension. Also, it is done by clearly defining the data dimensions through a specific term for each dimension to distinguish one data dimension from another, which, for instance, through the label 'data type' is not possible since all dimensions differentiate different types of data.

Criteria-based meaning: For a concept to practically work, it must overcome individual subjectivity by possessing criteria to practically work for actors (Kure et al., 2017, p. 218). This is especially relevant for the differentiation of data classes in a specific data dimension since here, the goal is to differentiate data properties based on specific criteria for actors to use and as a result to be able to identify data characteristics of datasets in use. Since data content is highly context- and environment-specific, subjectivity cannot, however, be eliminated, especially for those data dimensions, which consist of gradual merging criteria like, for instance, the degree of granularity. However, the

criteria are theoretically explained for each dimension and, therefore, can be adapted to the specific use cases at hand. The criteria to differentiate between private and public data is relatively definitive and generally valid, but for instance, for granularity, the classification depends on what the unit of analysis is with respect to the specific dataset analyzed.

Exemplary reference: Exemplary references are important for actors to establish a shared understanding of a concept, for instance, through shared collective experiences. In this framework, examples are manifold so readers can understand the classes and dimensions regarding their different terminologies.

3.4 Analysis of data dimension's locations inside the data value chain

After presenting all data dimensions and classes, the last step was a logical inquiry about identifying the location of the data dimensions inside the data value chain model from figure 1, which builds the basis of how we can understand the main processes from a signal of the world towards data practices. This means that for the data dimensions to have conceptual value they should be able to answer specific questions about those processes or areas. Only then they create visibility about the data production infrastructure for actors.

4 Data classifications

This section provides every developed data dimension and data class of the framework in no specific order. Every data dimension represents a distinct perspective towards datasets with at least two data classes inside describing specific data characteristics.

To increase the framework's readability, new data classes are **bold**, when introduced, and data dimensions, which are often are cross-referenced, are always *italicized*. The framework includes 23 data dimensions and 63 first-level data classes (Table 4).

4.1 Origin

As explained in the methodology, data dimensions should represent meta-perspectives about aspects of the data production infrastructure, which could be applied on all possible datasets. One relevant meta-perspective addresses the state of the original object, phenomena or signal captured in data as either being analog or already digital. In harmony with the computer-oriented perspective, data must always be digital, which means that a differentiation between digital and analog data here cannot be applied. Instead, the question, which needs to be answered by the data classes, is about the origin of the information stored in the dataset of either originating from the digital or analog realm. If the original information is analog, it must be digitized to be represented in data. Therefore, *origin* uses the differentiation of **digitalized/digitized** and **digital born data**, also termed 'born-digital' (cf. Snee et al., 2016, p. 68; Kitchin, 2014b, p. 7) or 'natively digital data' (cf. Rogers, 2013, p. 206) to describes the different origins of the information the data is based on of either being a representation of physical objects or analog materials, or being created directly in the digital space. Examples of 'digitized data' available today are mainly digitized books, movies, pictures, maps, music and more, which were originally created on physical paper or analog material before the digital era. Digitization makes it possible to process analog material, originally unreadable for computers, through the binary digital representation, which results in computer-readable data formats. Especially in the digitization of books, a subsequent objective lies in the datafication of the content, making the text searchable by readers and analyzable by software programs meaning that the book pages are not only represented by pictures but by genuine letters and words. Digitizing past information initiated the worldwide unprecedented fast distribution and consumption of information, where nowadays information is mainly natively born digital. This 'digital born data' is data, which is created directly online like social media posts, blogs, forum posts, E-Mails, articles, spreadsheets, etc. or in digital devices like digital video cameras, smartphones or other Internet of things (IoT) devices; but it is also 'research data', which intentionally is created by scientists applying online methods like online experiments or online surveys. This data can directly be analyzed by software and normally do not have a physical counterpart in the physical world.[1]

4.2 Sector issued

The terminology of private and public data is not used synonymous throughout literature due to different views of what those terms should describe in perspective to data. The distinction of private vs. public can either describe the difference sources, the different uses or the different access possibilities towards a dataset. This framework will use the source-based view because other terms are better suited to describe the use and accessibility perspective. Therefore, **private data** in this framework is all data which is created in the private sector, meaning private companies or individuals not representing public institutions. **Public data**, on the other hand, is all data created in the public sector, mainly governments, governmental institutions, inter-government organizations and other public institutions. However, not all

[1] A physical representation of an E-Mail, which by definition is in its original form digital, would be a printed E-Mail on paper.

institutions can be distinctly allocated into either the private or public sector, for example, in the case of non-governmental organizations (NGOs). If the use of data should be the focus of argumentation, the terms public-use and private-use data should be applied, where ´public´ here means that the data was primarily created or issued for organizations and institutions in the public domain or the general public (*Sector used*). For instance, publicly traded companies in the USA are required to issue quarterly results based on security laws enforced by the Securities and Exchange Commission (SEC), which is a government agency, to inform the public about recent financial performance (cf. Investor.gov). Based on the classification used in this framework, those datasets are 'private data' and 'public-use data'. There are arguments against this definition, describing those filing documents as 'public data' since they are available to everybody. However, these descriptions are possible by the *Accessibility* perspective, which applies terms to describe datasets as either publicly available, restricted or closed. Therefore, the quarterly results are private, public-use and publicly available datasets. Another example could be a user on Facebook posting a picture with the assumption that only friends can see it, however, using a wrong setting and, therefore, making this picture visible to everybody. This data should be described as private, private-use, publicly available data.

Table 4: Overview data dimensions and classes

No	Data dimension	First-level data classes							
1	*Origin*	digitized				digital born			
2	*Sector issued*	private				public			
3	*Sector used*	private-use				public-use			
4	*Accessibility*	publicly available		restricted		closed			
5	*Generation purpose*	administrative	scientific-use	social interaction	statistical-use	sensory	surveillance	code	
6	*Level of collecting obtrusiveness or bias*	made				found			
7	*Format restrictiveness*	open formatted		proprietary formatted		standard non-open			
8	*Processing degree*	raw				processed			
9	*Machine interpretability*	structured		semi-structured		unstructured			
10	*Measurement nature*	statistical	textual	visual	audio	spatial	temporal	network	log
11	*Sensitivity*	sensitive				non-sensitive			
12	*Scope*	population				sample			
13	*License*	non-licensed		open-licensed		restricted-use			
14	*Nature of information*	quantitative				qualitative			
15	*Reference type*	metadata				content data			
16	*Actor type*	machine-generated				human-generated			
17	*Granularity*	coarse-grained				fine-grained			
18	*Relationality*	relational				non-relational			
19	*Access method*	self-generated		direct accessed		indirect accessed			
20	*Analysis method fitness*	descriptive		predictive		prescriptive			
21	*Feature dimensionality*	high-dimensional				low-dimensional			
22	*Measurement dimensionality*	unidimensional				multidimensional			
23	*Generation velocity*	high-velocity				low-velocity			

4.3 Sector used

As explained before, the terms 'private-use data' and 'public-use data' describe the target sector for which a dataset is primarily created. **Private-use data**, therefore, is data that was primarily created to be used in the private sector, whereas **public-use** data is data that was primarily created to be used in the public sector, for example, in governmental institutions.

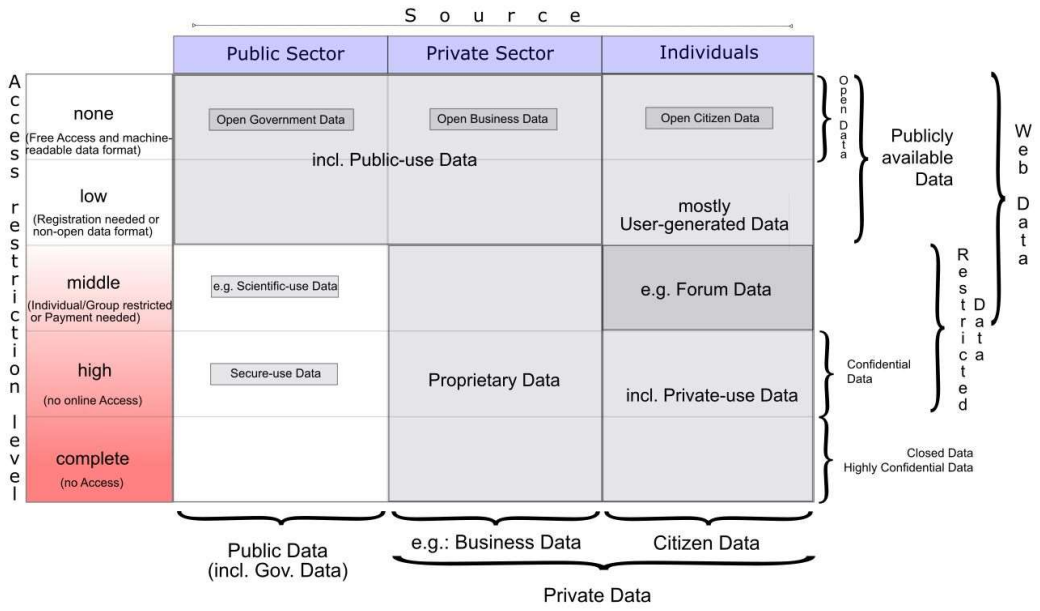
4.4 Accessibility

The Accessibility dimension describes data by the possibilities of actors to access it. **Publicly available data** are all datasets that are downloadable by anyone who has Internet access without involving financial payments or proving their identity. **Restricted data** have restrictions applied, which make the data only possible to download by actors proving a specific status defined by the data issuer or through payments. **Closed data** are datasets unavailable to actors outside a specified space around the data issuer.

Figure 2 describes data accessibility as a five-level dimension on the y-axes. The lowest restriction level describes datasets which are freely downloadable without any conditions and which are additionally in an open data format, which means that these datasets can be opened by a free downloadable and usable software program (*Format restrictiveness*). The second level describes datasets, which are freely downloadable but have either a non-open data format or registration is needed in the process. Since individuals don't need to prove a specific status e.g., a specific occupation, to gain access, these datasets are still considered to be publicly available. The third level includes datasets where individual - or group restrictions make sure only actors of a predefined type can access the data. For example, one data source for social scientists is the Consortium for European Data Archives (CESSDA) (cf. CESSDA), where social scientists can upload their research results and corresponding 'research data' making this data reusable for other studies. Here, 'sensitive data' (*Sensitivity*) where participants of a study could be identified can only be downloaded by using a request form where the requester needs to prove the exclusive scientific use of the data by declaring the specific study the data should be used for. Another example is data from online forums for users with common interests, for example, a specific car brand, or similar opinions and worldviews, for example, an online forum for people of a specific religion. These two examples show that individual - or group access restrictions can be of very different intensity, whereas faking being a scientist might be more difficult than faking the ownership of a car. However, in both cases, the corresponding data is intended to be not publicly available. The fourth level describes datasets, which, due to high *sensitivity*, are not available online but can be accessed personally by physically visiting a specified location of the data owner (On-site access) making sure the requester is personally identified.[2] CESSDA are calling those datasets "Secure-use files" (cf. CESSDA Webinar, 2018, min: 19:40). The fifth level are datasets only accessible internally inside an organization or by a small group inside the organization (e.g., the top management). Proprietary business data of that kind are often termed 'confidential data', 'highly confidential data' or 'proprietary data'. The fourth and fifth level excludes data, which is accessible over Web browser URL addresses (Web data) or public APIs (*Access method*), because that would make the data accessible to outsiders.

[2] For example: the UK Data Service offers virtual research environments in a secure lab to analyze confidential and sensitive data (cf. McGivney, 2018, p. 16; Bauder, 2014, p. 4).

Figure 2: Access restriction levels (*Accessibility*) and Sources (*Sector issued*)



4.5 Generation Purpose

A classification scheme into different main types of generation purposes is difficult since many different purposes exist for why data is created, and additionally, also multiple purposes may coexist for the same datasets. Important in the differentiation is to first understand the primary use of the data, meaning the receivers' (machine or human) primary purpose of collecting or using the data and not what in the big data context often is termed 'secondary use', since the primary purpose mostly defines the process of data generation and therefore the needed or intended data quality level.

Administrative datasets are generated by administrative systems of organizations, especially for the purposes of registration (i.e. birth, death, marriages, voters (cf. Woollard, 2014, p. 49 f.)), transaction ('transaction data' (cf. Marr, 2015, p. 86 f.)) and record-keeping (cf. Woollard, 2014, p.49; Connelly et al., 2016, p. 2 f.). These datasets are generated for an organization or department to function regarding its operations, e.g., collect taxes in a government tax office, register a customer order in a private company, or obey laws which oblige organizations to keep track of specified information. Although private organizations and public agencies are both producing high amounts of administrative data, social scientists will more likely be aiming to use administrative datasets from public institutions since those hold lots of valuable social information about societies, are often up-to-date, are increasingly publicly available (*Accessibility*) especially in the US and Europe (cf. Connelly et al., 2016, p. 3) and often include information about entire populations (e.g. 'census data') instead of just samples (*Scope*) (cf. Connelly et al., 2016, p. 4; Woollard, 2014, p. 52). For practitioners like accountants, controllers and managers, the most common data to deal with daily are internally generated 'administrative data' to function as an organization, but also 'public administrative data', for example, for benchmarking.

Scientific-use data is created to serve scientific discovery. In literature, often only the terms 'research data' or 'scientific data' are used to describe datasets applied by researchers as evidence to validate research findings (cf. Sayogo & Pardo, 2013, p. 20), however, since theoretically all data can be used as evidence in research the term 'scientific-use' clarifies that this kind of data is created as a primary purpose for a scientific study. Research or scientific data would then be used to describe the content of the data (*Reference type*).

Social interaction data is generated for the purpose of social interaction and communication between humans. This means that this data is generated when some sort of information is communicated from one human to another. Especially 'unobtrusive data' (*Level of collecting obtrusiveness or bias*) like 'social media data' fall into this category. 'Social interaction data' generated through scientific studies are both 'scientific-use data' and 'social interaction data' however, its primary generation purpose is to serve as evidence in a scientific study, meaning no social interaction would have taken place without the study, and therefore, those datasets describe no natural but artificial social interaction. Due to the complexity of social interactions and the variety of communication and expressions possible in social media, many deeper classifications of 'social interaction data' are possible. 'User-generated data' is often used to describe the initiating proactive role of humans to create content online, e.g., a posted YouTube video, whereas

'provoked data' (cf. Marr, 2015, p. 86) or 'reaction data' would rather describe a responsive role of a user reacting to others like one-click opinions, comments, resharing activities or live-stream chats. Also, pictures created with digital cameras, music produced in digital audio workstations (DAW) or writing a book on a computer should be labelled as 'social interaction data' since they mostly only produce individual value when the resulting products (pictures, music, books) are consumed by others.

Statistical-use data is data created for statistical purposes. For scientists or business intelligence departments in corporations, especially public datasets, issued by governments and other public agencies, hold important social information about populations of interest. Important here is that 'statistical-use data' in its primary form is 'raw data' not yet processed (*Processing degree*) and transferred into statistical variables. The term statistical-use only describes the purpose of collecting data and not the data´s *measurement nature*.

Sensory data (in literature the term 'sensor data' is used more often) is created by sensors *"to measure physical quantities and transform physical quantities into readable digital signals for subsequent processing (and storage)*" (Chen et al., 2014, p. 181). This data is relevant in today's advanced machines like production or transportation devices guaranteeing operational functionality ('operational data' (Floridi, 2008, p. 8)). It's generally 'machine-generated data' (*Actor type*) and can further be classified into the kind of environmental signal type getting measured or monitored, e.g. space, temperature, pressure, voice, proximity, smoke, or humidity, which often gets measured as 'log data' (*Measurement nature*).[3] 'Sensory data' is generated to inform either machines or humans, and specific types of sensory data can be used by scientists or social media companies like Facebook to indicate social behavior, for example, by analyzing GPS datasets generated by smartphones sensors. The high volume of 'sensory data' nowadays is caused by the increased amount of daily used devices[4], which are equipped with sensors, and which are connected to the Internet or internal networks informing other systems, devices or networks. This development is often termed the Internet of Things (IoT), and the corresponding data produced by those devices is sometimes called 'IoT data'.

Surveillance data (or with a positive tone, also termed 'monitoring data') is generated primarily to monitor the behavior or activities of machines, products, individuals (incl. 'health data'), groups or entire societies. 'Video data' (*Measurement nature*) from cameras fall into this category, but also data from specific surveillance programs like PRISM, biometric data like fingerprints or RFID data to track the movement of products. The true reasons to monitor and collect 'surveillance data' are manifold and range from movement control of goods, traffic control, tracking of user´s behavior collected in apps to display personalized advertisement, security control, crime-fighting or social order and control. E.g. China's social credit system is an extreme example of how surveillance technologies and data are used to establish a surveillance state to totally control the behavior of citizens in conformance with those in power. An important aspect in this context is again the consideration of primary vs. secondary purpose in the data creation process. For example, the primary purpose of a video camera in a factory might be to make sure employees are working and not stealing materials. These datasets are, therefore, only generated for the purpose of surveillance. On the other hand, if a state uses social media communication data of their citizens to identify possible dissidents and their social networks, the primary purpose of the monitored data is still communication. These datasets are 'social interaction data', since they exist independent of the secondary use for surveillance afterwards.

Code data is used to define the behavior of a computer or a digital device which is handling inputs (e.g., commands, instructions or measured signals from sensors) to create outputs (e.g., information or specific behavior) according to the programmed code. Its syntax and formal logic are defined by the programming language (like C, PHP, Python, Java, etc.) it is based on. The purpose is to finally automate tasks which would be either time-consuming or unable for humans to perform. Programming code can include comments included by programmers, which are not used by the computer, but make it easier for humans to read. For programming code to work correctly and consistently, demands on data quality in the code are extremely high. Code with even tiny mistakes or flaws could lead to unintended results or computer crashes.

4.6 Level of collecting obtrusiveness or bias

Data used in research or for any data analysis task in corporations can be differentiated between **made data**, also termed 'created data' (cf. Marr, 2015, p. 86) and **found data**, also termed 'naturalistic data' (cf. 6 & Bellamy, 2012, p. 74). This distinction is aimed to describe a researcher's or collector's involvement or obtrusiveness in the data generation and collection process in a study, whereas 'made data' in science are created through traditional methods of experiments ('experimental data' (cf. Marr, 2015, p. 87)), surveys, censuses and interviews and 'found data' have no researcher involvement meaning data can be ´found´ especially online for example on social media platforms. In this context, 'scientific-use data' (*Generation purpose*) is always 'made data' since those datasets are specifically made for

[3] An overview of 15 sensor types provides Sharma et al., 2021, p. 28 f.

[4] Examples are especially all electronic devices, which are using wireless protocols to transfer data and which are increasingly using the "smart" prefix to differentiate them to older technologies equipped in those devices like smartphones, smart cars, smart refrigerators, smart watches, smart speakers, etc.

scientific purposes. In practice, marketing or business intelligence departments could ask customers specifically about their opinion of a product by conducting a questionnaire, which would create 'made data'. Extracting opinions from social media without involvement from the company itself is 'found data'. In science, the obtrusiveness in 'made data' depends on the research design and is highest in experimental studies due to highly controlled and unnatural ceteris-parisbus conditions set by the researcher. It is significantly lower in data from observational studies and nonexistent for non-scientific-use datasets since those are generated for other purposes outside the influence of the researcher or collector.
It is important to point out that obtrusiveness here purely relates to the involvement or non-involvement of the data collector, e.g., a scientist or an organization and is not a general indicator for the fitness of the dataset towards answering specific (scientific) questions. 'Found data' does not indicate that the data is created unobtrusively and, therefore, does not necessarily indicate high data quality. E.g. although datasets extracted from social media as 'found data' can be unbiased regarding the data collector, multiple other technological or non-technological biases may exist. Therefore, in relation to data quality also terms like biased vs. unbiased data or obtrusive vs. unobtrusive data could be used to describe the existence or non-existence of the influence of unintended or external forces on the measured phenomenon, which then finally will be present in the dataset.

4.7 Format restrictiveness

Formats of data vary due to the different *measurement nature* represented in datasets (e.g., tabular data, textual data, pictures, videos) and can be classified into either **open formatted data** formats, which can be read by freely available software, and proprietary data formats (**proprietary formatted data)**, whose encoding scheme is either secret or restricted by licenses or patents. There are plenty of open data formats incl. JPEG, GIF, CSV, HTML, PDF, XML, ZIP or JSON. Additionally, there are formats which are proprietary but worldwide accepted and highly used that those became de facto standard formats and, therefore, can be classified as **standard non-open data** formats. Examples here are especially Microsoft´s DOC, XLS or WMA file formats. Proprietary non-standard data formats, for example, for the researcher to analyze datasets are SAV for the statistical software SPSS, DTA for STATA or SSD for SAS.

4.8 Processing degree

A very important aspect for scientists and analysts when working with datasets is to understand the degree of processing involved in the datasets at hand. **Raw data** is considered as the initial, unprocessed data directly generated from the source, and therefore, this data is uncleaned, unrefined, unedited, unaggregated, unformatted, unmerged, uninterpreted or in any other way transformed. **Processed data**, on the other hand, is all data which is not raw anymore and, therefore, in some way manipulated to make the data useful for analytical purposes like statistical reporting and analysis, scientific interpretation, visualization or standardization for either humans or machines. Other terms similarly used for raw data are 'microdata'[5], 'source data', 'primary data' (cf. Kitchin, 2014a, p. 7) or 'unprocessed data'. While definitions of these terms are not always identical, they aim to emphasize a high degree of purity, neutrality, transparency and objectivity of raw datasets compared to processed datasets, also termed 'secondary data' or 'derived data' (cf. Kitchin, 2014a, p. 6), which are impacted by subjective choices of actors manipulating the data for different purposes. Since these purposes do change due to different requirements of varying research objectives, concepts, target systems, audiences, analytical methods or other secondary activities, the same 'raw data' can be processed in many different degrees, directions or depths, resulting in multiple processed datasets deriving from the same raw data. Standard data processing steps involve data cleaning (or cleansing), data aggregation, data summarization, data reporting and visualization.

4.9 Machine interpretability

The differentiation in structured and unstructured data is in literature normally termed as being a question of machine readability (cf. Gandomi & Haider, 2015, p. 138). This framework proposes the term of *machine interpretability* instead. When using machine-supported analysis techniques it is often not enough that computers can read and present the datasets to the user, they also need to extract useful information. For example, it might not be enough for a computer to be able to read and display videos from recorded interviews when an analyst is only interested in those parts where specific topics are discussed. Here, speech recognition tools of the audio files could be used to detect keywords defined by the user.
Another example would be to identify pictures with specific content, for instance, pictures showing a beach. Therefore, it is not a question of machine readability, because the computer can read the videos and pictures, but a question of the

[5] Microdata is often used to refer to data containing individual raw observations and responses of individuals and household acquired through censuses and surveys (cf. Bauder, 2014, p.4; Narayanan & Shmatikov, 2008, p. 1; IPUMS USA FAQ).

degree of causing difficulties for computers to extract useful information from the data. This degree depends mainly on the *measurement nature*, the information or content saved, the source of the data (*Sector issued*) and the intended analysis task on the data (*Analysis method fitness*).

In general, only 'statistical data' (*Measurement nature)* is considered as being fully **structured data** since it can be saved in spreadsheets and relational databases (*Relationality*) (cf. Gandomi & Haider, 2015, p. 138; Kokina et al., 2017, p. 51; Moffitt & Vasarhelyi, 2013, p. 5). These datasets are easily searchable and statistically analyzable. Textual, graphical, video and audio datasets are considered as being **unstructured data** except for Extensible Markup Language (XML) or Hypertext Markup Language (HTML) data, often described as **semi-structured** (cf. Chang et al., 2006, p. 1413). However, the machine-interpretability also depends on the information stored in the data and, therefore, also the source. The content of social media posts is much more difficult to interpret by machines than contractual text data, where rhetoric devices like irony or sarcasm can´t be present. The term 'unstructured data', on the other hand, can also refer to the wide variety of multiple data types, leading to difficulties for computers to relate their content to each other. This is, for example, the case when Facebook feeds from users are extracted, which can consist of text, pictures, videos, likes, links and other data at the same time. Whether datasets are considered as structured, semi-structured or unstructured also depends on the analysis tasks a user wants to get performed on the data by a computer. Chang et al., 2006, for example, describes reasons for differing views on the structuredness of XML and HTML data lying in the different research backgrounds and goals of users (cf. Chang et al., 2006, p. 1413).

4.10 Measurement nature

Considering data types discussed in literature most often, those refer to the nature of measurement applied in the recording process deriving from world phenomena to datasets representing those observed phenomena. The *measurement nature* defines the possible analytical methods available and the needed software for accessing, manipulating, analyzing or visualizing the data. *Measurement nature* relates very strongly to the *nature of information* since possible measurement techniques must fit the phenomenon desired to be represented in the data. The distinction, however, is important since methodological issues can arise, especially when quantitative/statistical measurement techniques are applied to essentially qualitative/social phenomena of the real world.

Statistical data is the most common data type scientists use as evidence in quantitative research, and companies use it for reporting and decision-making. Typical sources of 'statistical data' include public statistical-use and administrative datasets, made scientific-use censuses, samples from social surveys and data from firms' transaction systems (*Sector issued, Generation purpose, Level of collecting obtrusiveness or bias, Access method*). Depending on the methodology, the data can be analyzed through descriptive, exploratory, inferential and predictive statistical methods (*Analysis method fitness*). 'Statistical data' is commonly organized in a structured manner, for instance inside Excel spreadsheets or inside a SQL-based data warehouse system and includes especially numerical and categorical data types. The deeper classification into numerical and categorical data, however, is normally a description of a measured dimension in the dataset and not a description of a complete dataset (*Feature dimensionality*). Therefore, 'statistical data' can have numerical and categorical data aspects included especially with increasing *feature dimensionality*.

A **numerical data** dimension in a dataset is an expression or measurement in numerical terms of that dimension, e.g., monetary measurement of daily revenue in US Dollars. It can further be divided into discrete and continuous data, where **discrete** have a finite amount of data points between every two data points. Monetary measurement is an example of a discrete measurement, where cents are the lowest data units[6]. **Continuous data** is data which can have infinitive data points between any data point in relation to a measured phenomenon. Examples are height, weight, time or temperature. It should be noted that this definition of continuous data is only valid in accordance with the 'signal-oriented perspective' explained in the second chapter. Therefore, it describes the real analog world we live in and not how a computer measures it. A computer cannot measure infinite time data points between two time points because it is digitally built and has a lowest time unit on which it performs. Any signal variation between two consecutive time points cannot be captured or processed by a computer. One further differentiation of numerical data is between **interval** and **ratio data**. The difference between both is that 'ratio data' has a true logically zero point like 0 for weight or age, whereas interval data doesn't. Generally, 'ratio data' is preferred since some mathematical operators like multiplication or division are not possible on interval data (+3°C is 2° warmer than +1°C, but not three times warmer since temperature measured in °C can be negative).

Categorical data is data describing an object's attribute by allocating it into specific groups or categories. These categories can be textual, like describing a color category named 'black', or numerical by dividing sports groups into the teams '1' and '2'. However, in both cases, most mathematical operators cannot be performed because there is either no logical order between the categories ('black' is not better than 'green', '2' is not twice as good as '1') or the order is not consistently measured like, e.g., a questionnaire asking for the subjective opinion of a restaurant service between a

[6] Under the assumption that you cannot earn fractions of a penny.

rating of 1 to 5. In this example, the categorical data is ordinal (**ordinal data**), meaning that data can be ranked logically by, e.g., saying a restaurant with a rating of 5 is better than 4 but without the possibility to mathematical calculate how much better. In case of no logical ranking (eye color example) data is defined as **nominal data**. The distinction between ordinal and nominal is normally only done with categorical data since a logical order in numerical data is always existent and therefore numerical data is always also ordinal.

Text as elements of discourse and as a means to save and share information, opinions, and knowledge can support evidence-based research when analyzed either manually or machine-supported. Sources of **textual data** for social scientists and data analysts are especially online libraries, transcribed interviews, emails, publicly available social media comments, blogs, online forums, survey responses, online news, contracts, etc. (cf. Gandomi & Haider, 2015, p. 142). Especially textual 'born digital social interaction data' on social media (*Origin, Generation purpose*) is considered as a fast-growing pool of social information useful for research either for science or private organizations (cf. Warren et al., 2015, p. 399 f.; Chen & Yu, 2018, p. 11 f.). Analytical methods to extract information from 'textual data' are, especially content analysis, text summarization, text categorization, text clustering, sentiment analysis and more. Although 'textual data' are normally described as being 'unstructured data' (*Machine interpretability*), text elements like sentences can be put into tagged containers to give meaning to machines and to organize them in a structured manner. This is, for example the case in data saved in the Extensible Markup Language XML format, which can be interpreted by humans and machines. Therefore, these datasets can be interpreted as a mix of 'textual data' and 'statistical data'.

Graphical and video data are both **visual data**, which are increasingly used as a form of social communication (*Generation purpose*) and analyzed in social science research (cf. Knoblauch et al., 2008, p. 1 ff.). 'Graphical data' includes digital images, drawings, photographs, symbols, maps and paintings. 'Video data' consists of digital sequences of images normally joined with audio patterns. Graphical and 'video data' can be captured by researchers and analysts by using visual methods in their study, e.g., photo interviews or video-recorded focus groups, or unobtrusively (*Level of collecting obtrusiveness or bias*) collected from image and video sharing platforms like Facebook, Instagram, YouTube, streaming sites like Twitch or, if granted access, from public or private surveillance cameras. While in the past, images and videos could only be analyzed manually, advances in machine learning nowadays make it possible to analyze high volumes of 'visual data' automatically to extract useful information (cf. Gandomi & Haider, 2015, p. 141 f.). Those analytical possibilities, together with the heavily increased volume of 'video data' due to the massively worldwide growth and public acceptance of camera surveillance (cf. Lyon et al., 2012, p. 1 ff.) and people's increased willingness to freely share personal information holds huge potential for researchers and private organizations to observe and understand social behavior and opinions.

Audio signals like music, speech or acoustics can be captured in **audio data**. Audio files can be analyzed by machine algorithms to extract meaning and information either to detect non-verbal cues from human speech, for instance, a mental state change of an individual in a recorded interview indicated by changes in the voice frequency, or when high data volume makes manual investigation impossible. The most common analytical techniques are speech recognition to translate spoken languages into text and speech emotion recognition to classify speech utterances and emotional attributes (cf. Parthasarathy & Tashev, 2018, p. 1).

Spatial data as an umbrella term for geo-, geospatial-, geographic- and georeferenced data[7] is defined by Evans et al., 2019, as "*discrete representations of continuous phenomena over the surface of our changing planet*" (Evans et al., 2019, p. 152) and consists of either raster data created by remote sensors to mainly create geo-images, vector data to formalize relationships between objects through points, lines and polygons (cf. Evans et al., 2019, p. 154), or graph data to create digital and interactive maps. The availability of spatial datasets for researchers to study mobility, cultural and historical developments, urbanization and neighborhood, transportation, environmental changes (e.g. pollution), disasters, climate change, etc. has increased mainly due to the high amount of georeferenced information created in devices by GPS sensors today and the open data initiatives around the world to make those public datasets publicly available (*Accessibility*).[8] Sources of those datasets are, especially georeferenced social media posts, geo-databases like Google Earth, Google Maps or OpenStreetMap, which all include volunteered geographic information (VGI) from citizens, or publicly available data from governmental institutions like public transportation data. The scientific value of 'spatial data' considerably increases when it is analyzed through time. In this case the data is termed 'spatiotemporal data' (cf. Blundell et al., 2018, p. 263).

[7] There are differences between those data types. While geo- and geospatial data refers to all datasets referring to some spatial dimension or representation on earth, geographic data refers to a specific place on earth represented by a distinct coordination (Longitude and Latitude) and georeferencing meaning the adding of geospatial information to other datasets.

[8] For instance: the INSPIRE Directive from the European Union (https://inspire.ec.europa.eu/).

Temporal data in its raw form is like 'spatial data' normally generated by machines (*Actor type*) and as metadata automatically attached to the 'content data' (*Reference type*). As 'spatial data' refers to the spatial dimension of the world, temporal data refers to the temporal dimension and includes fixed time stamps or time durations.

Network data is data representing graphs, which are sets of nodes and edges combining nodes (cf. Lajaunie et al., 2018, p. 170). In its raw form, 'network data' is 'statistical data', which can be shown in structured relational tables, but since it is a very important concept in social science and often visualized as networks (sociograms) it is presented here as a separate data type. Those graphs can be of relevance when they can indicate social networks, which today are much more visible due to the worldwide use of the Internet as a communication tool. Based on the different characteristics a social network can possess (directed vs. undirected edges, weighted vs. unweighted edges, uniform vs. multiform nodes) and how it is visualized (complete vs. partial vs. egocentric), different types of network models can be constructed and analyzed (Cf. Ackland, 2013, p. 49 f.). Sources of network data can be web pages, where a community is present indicated through social interaction[9], and public sources[10]; however, the most prominent sources are social media platforms like Facebook, Twitter and LinkedIn. The rules of communication on those sites define the resulting network type.[11] Social networks can be analyzed quantitatively by many different measures, for instance size, centrality, density, inclusiveness, degree, dependence and more. Those are used in many different disciplines to understand social structure and collective behavior like social influence, collaboration, participation, social movements, communities or power. Another important type of networks used in social science are hyperlink networks extracted from hyperlinks connecting webpages and therefore information on the Internet.

Log data (or logging data) is like 'network data' "*not an evident category of data separated from other data types*" (Ørmen & Thorhauge, 2015, p. 337), since it can include textual, spatial, statistical, temporal or even graphical data. However, due to its unique and new digital event-based tracking capabilities and, therefore, the behavioral observation opportunities arising for research and practice, 'log data' is here presented as a separate data type. Log datasets consist of information from log messages created by computer systems in response to stimuli, which can be specific activities (i.e. events) or are triggered based on time intervals (cf. Chuvakin et al., 2013, p. 2 f.; Ørmen & Thorhauge 2015, p. 337). Types of log messages can be differentiated into informational, debugging, warning, error or alert types (Chuvakin et al., 2013, p. 4). While 'log data' historically was used to analyze machine behavior with the goal of creating audit trails to understand and improve software behavior, nowadays, it is increasingly used to identify and analyze human behavior. In this case 'log data' is often termed 'trace data' or 'digital trace data', because logs here can describe the interaction of a human with and social communication through a digital device. Examples are logs about how website users interact with a specific website ('clickstream data') and especially sensory and usage 'log data' from smartphones, which can unobtrusively (*Level of collecting obtrusiveness or bias*) log human behavior throughout the day. This is possible due to the increased integration and everyday usage of smartphones by humans, which makes it possible to for example, map human movements by smartphone's sensory GPS data logs or to analyze mood, personality traits and social relations through application usage data, which is often saved automatically in form of metadata (*Reference type*) (cf. Ørmen & Thorhauge, 2015, p. 338 f.).

4.11 Sensitivity

The question of data sensitivity and privacy is not new to social scientists or analytics departments; however, it might get increasingly important and challenging to especially small private organizations with limited digital expertise and resources. In science, research guidelines have been developed already decades ago describing ethical research principles for obtrusive research methods like experiments, interviews or surveys. Those principles include the researcher´s obligations to inform participants (informed consent) about the methodology of the study, the use of the collected data and the responsibility to protect the participants against harm (cf. Little, 1973, p. 79 f.). Harm in this context includes the invasion of privacy done by external disclosure of sensitive information after the study. Sensitive information is not objectively delimitable against non-sensitive information since perceptions and interpretations of privacy vary worldwide and change over time. However, for individuals, it includes mainly specific 'sensitive personal data', for private organizations confidential 'business data' and for the public sector 'state security data'. The EU General Data Protection Regulation from 2016 defines in paragraph 75 a long list of general privacy and freedom rights of citizens which can be violated by data processing and through this indirectly defines categories of 'sensitive personal data'. Those include information about "*racial or ethnic origin, political opinions, religion or philosophical beliefs, trade-union membership, … genetic data, … sex life … criminal convictions, …, work, economic situation, health, … location or movements …*" (Council of the European Union, 2016, p. 14). Therefore, datasets including this personal

[9] This interaction however can be most minimal and unilateral for example one follow click on Twitter.

[10] For example: Stanford University (http://snap.stanford.edu/data/index.html).

[11] Ackland, 2013, distinguishes between four main types: affiliation, information, communication and social network (cf. Ackland, 2013, p. 73 f.).

information on an individual basis ('raw micro data' – *Processing degree*) can be defined as sensitive. Since these datasets can hold lots of scientific value for researchers and private companies, anonymization is one standard method to make individuals or attributes of individuals unlinkable to outsiders even when data is publicly published (cf. Torra, 2017, p. 8). However, these anonymized datasets are still considered as 'sensitive data', since they are keeping individual records ('raw attribute data') of sensitive individual information. Additionally, already in 2008, Narayanan and Shmatikov have shown in their Netflix case that anonymization of 'micro data' does not prevent of being detected since high data dimensionality (*feature dimensionality*) made cross-correlation with other publicly available datasets possible (in this case with IMDB data) and therefore individuals re-identifiable (cf. Narayanan & Shmatikov, 2017, p. 1 ff.). Therefore, **sensitive data** can be summarized as data holding individual (raw) sensitive information independent of anonymization. Individuals in this context can mean persons, organizations, departments, companies, institutions, etc. This data sensitivity obliges scientists and private companies with access to these datasets to protect them against unwarranted disclosure. **Non-sensitive data** is data without sensitive information or data with highly aggregated sensitive data without individual records.

4.12 Scope

Both in quantitative and qualitative research, scientists must consider to which degree internally generated or externally acquired data statistically or conceptually represents a population on which inferences and statistical or analytical generalizations are intended to be drawn. If observations or characteristics of the complete population of interest are present in a dataset, this is termed **population data**. The most common population datasets are created through censuses, which are conducted with the objective to contacting every single unit of a population if those are either single persons or single institutions (cf. Bauder, 2014, p. 5). Sources are especially public datasets from governmental institutions (*Sector issued*) with statistical, administrative or scientific purposes (*Generation purpose*). Despite these sources and the promises of data inexhaustibility in the big data era, 'population data' is rarely available leading to datasets that usually only represent a subset of a defined population of interest. These datasets are termed **sample data** and differ between sample size and sampling scheme. The sample size is evaluated against the size of the population and can be as small as one case or one observation, for example in single case studies. Although not an inevitable condition of qualitative research designs (cf. Onwuegbuzie & Collins, 2007, p. 287 f.), sample sizes are generally smaller in qualitative studies since the in-depth understanding of a phenomenon often is more relevant than generalizability of research findings. Sampling schemes can be differentiated between probabilistic (random) and non-probabilistic (non-random) schemes (cf. Onwuegbuzie & Collins, 2007, p. 285). Probability samples have the advantage that statistical sampling errors and, therefore, sampling bias are measurable; however, due to convenience, non-relevance of generalizability or when important characteristics of the population are unknown, making randomization impossible, non-probabilistic methods are also widely used in social science. The question of sampling method is relevant not only for scientific-use data created by scientists but also for publicly available data, especially when sampling methods are hidden or changing over time (*Access method*).[12]

In private organizations, scope can be equally important. For instance, a retailer could offer customers to use loyalty cards with discounts to connect their buying behavior to personal data. Here, sample data is collected to indicate the general buying behavior of all current customers or even the buying behavior of the whole population.

4.13 License

Due to the constantly increasing availability of externally generated datasets, questions on what rights researchers or organizations have on analyzing, quoting, republishing, copying, modifying and sharing are of great importance. Hereby, the idea of data licensing is that the data owner grants specific rights or assigns specific obligations to a data licensee. The question of data ownership however is not always easy to answer since ownership can be interpreted differently[13], many actors in the data value chain may claim ownership simultaneously, data ownership legislation is often either non-existent or unclear[14], and national privacy laws are varying worldwide (cf. Jaakkola et al., 2014, p. 34).

---

[12] E.g., the logic behind APIs of social media sites is often unknown or changing over time effecting what kind of datasets can be extracted by queries (cf. Felt, 2016, p. 3).

[13] E.g., the question if personal data is always owned by the individual or even if ownership in data in general actually exists (cf. Van Asbroeck et al., 2017, p. 4 and p. 22 ff.).

[14] E.g.: Despite the General Data Protection Regulation of the EU from 2016, Van Asbroeck et al., 2017 conclude in their white paper on EU data ownership that there are numerous legislations in the EU that impact issues of control, access and rights in data, however no EU legislation that specifically regulates data ownership (cf. Van Asbroeck et al., 2017, p. 22).

Despite these uncertainties, data controllers and data processors[15] can assign licenses on datasets under their control to inform data consumers about the rules of use. Information on those rules can be placed online next to the download links of the datasets and additional as metadata (*Reference type*) inside the datasets. The possible ways to save license information in the metadata depend on the *measurement type* and have the advantage that users can identify those license information independent of the download (e.g. property information of an excel file) (*Access method*), or that the license type is machine-interpretable (*Machine interpretability*) (e.g. in case of XML data) allowing software tools to execute errors when specific data tasks are not allowed like e.g. merging data (cf. Ball, 2014, p. 12). When there is no information available (at the access point or in the metadata) from the data issuer about conditions of use, actors can contact the issuers to clarify possible restrictions. Until clarification, these datasets should be treated as **non-licensed data**, which is very different to **open licensed data**, where the issuer specifically declares zero use restrictions. 'Open licensed data' (also termed 'free content data') is one of the core conditions of data being considered as 'open data'. Use restrictions for data consumers can be of different degrees. The lowest restriction only constitutes an obligation to reference the data issuer/creator when data is used for example, in a scientific study. The strongest restrictions do not allow any external use at all, meaning all rights are reserved. To classify **restricted-use data** a few licenses schemata has been developed by different organizations, whereby the Creative Commons (CC) licenses classification scheme is one of the most respected and used ones (cf. Heath & Bizer, 2011, p. 52; Jaakkola et al., 2014, p. 34). CC-licenses differentiate between the rights and obligations of attribution needed vs. no attribution needed (BY), share-alike required vs. remixing allowed (SA), commercial use allowed vs. only non-commercial use allowed (NC) and derivatives works allowed vs. derivatives works not allowed (ND).

4.14 Nature of information

The distinction between qualitative and quantitative data can be confusing because in methodology-related literature, differences of both quantitative and qualitative research designs and quantitative and qualitative analysis methods are discussed often without explicitly defining either quantitative or qualitative data. This leads to contradicting interpretations of what properties a dataset needs to possess to be characterized as being quantitative or qualitative. This is very good visible in relation to discussions about big data being described as "*necessarily quantitative*" (Cowls & Schroeder, 2015, p. 470), but also considered to refer to "*enormous amounts of unstructured data*" (Cuzzocrea et al., 2011, p. 101) like 'social interaction data' from social media (*Generation purpose*) (cf. Richins et al., 2017, p. 65), which consists of mainly qualitative information like comments, discussions, video content, etc. The reason for these diverse interpretations is that, when dealing with large datasets in the private sector and in academia, increasingly quantitative analysis methods are used to analyze basically qualitative information. One example is to use sentiment analysis on a dataset consisting of comments on a specific topic from social media with the objective of making sense of public opinion of either being positive or negative. Since a computer cannot and probably never will be able to understand human conversations, it can only apply quantitative methods, in this case through counting words predefined as either positive or negative, to then estimate a sentiment. However, data should be described independent of possible analysis methods (Exception: *Analysis method fitness*). In this framework, the distinction between 'quantitative data' and 'qualitative data' depends on the nature of information in the data and the measurement method applied to create this data. If there is some form of natural law or human-made and generally accepted logical mathematical concept underlying a measured phenomenon, it is quantifiable. By a representation of that phenomenon through a quantitative data measurement process underlying these laws or concepts, 'quantitative data' is produced. There are underlying natural laws for the weight of a person (gravity), and a quantitative measurement method, therefore, is the logical choice. Revenue or costs are based on a consistent mathematical human-made concept, i.e., money, and therefore can also be measured quantitatively. Therefore, **quantitative data** possess numeric information (quantities, percentages, statistics, values) with the goal of clearly measuring information grounded in natural laws or human-made concepts on distinct scales (*Measurement nature*) and metrics like monetary units, lengths, temperature, sizes, amounts, etc. Due to these consistent measurements defining the data, computers can analyze the information statistically. This is the reason 'quantitative data' is often considered to be automatically also 'structured data' (*Machine interpretability*). This one-to-one relation, however, cannot be made to 'statistical data' (*Measurement nature*), since this includes 'categorial data' (e.g., a survey answers to favorite artist), which in its raw form is non-numeric and therefore qualitative. **Qualitative data** is all data, which is not quantitative data like narrative texts, descriptive observations, pictures or videos. Simple quantitative methods applied to this data can, for example, answer questions of how many percent of people in the dataset have a specific favorite artist. For other questions the underlying qualitative phenomenon however cannot be described quantitatively without applying qualitative and, therefore, subjective assumptions of how to describe the phenomenon numerically.

[15] Organizations providing data analytical services are an example of data processors. Data brokers on the other hand are data controllers normally with the goal of selling data (cf. Van Asbroeck et al., 2017, p. 20 f.).

Often the difference between quantitative and qualitative data is compared to the difference of the underlying activities of `describing` vs. `counting` (Cf. 6 & Bellamy, 2012, p. 82), which however is misleading, since numeric representations of things or concepts are also descriptions. For instance, a dataset representing the number of YouTube users pressing the Like button under a video. This data is clearly numeric and quantitative in respect to answering the question of how many users have clicked on the button, but this data cannot answer the question of why they clicked without making qualitative and subjective assumptions. It describes the number of clicks made; the underlying phenomenon of clicking, however, is a social one and is not based on natural laws. In this perspective, the same dataset could be quantitative data for answering the question of how many clicks have been done (objectively measurable) and qualitative data to answer questions of sentiment, which cannot be concluded without subjective assumptions. Probably, most scientists and practitioners would disagree with this interpretation and always equalize quantitative and numeric data. However, I believe it is useful to consider the nature of the underlying phenomenon and the questions we want to get answers to instead of just considering the measurement method alone (*Measurement nature)* because if we apply quantitative methods to measure a qualitative social aspect of reality, there are threats to validity which needs to be considered. Often, direct measurement of a (social) qualitative phenomenon is not available (we don't see people's reaction to a YouTube video, nor can we talk to them), but instead, available data (YouTube likes) is used as a proxy to measure that phenomenon. If a YouTube video has 80% likes (proxy), we say that 80% of users liked the video or agree to the statements made (phenomenon). The measurement method, therefore, in this example is quantitative and already defined in dimension *Measurement nature* as statistical data, whereas the underlying nature of information intended to gather out of the data is, however, qualitative.

4.15 Reference type

Reference type is the dimension differentiating between **metadata** and **content data**. If a dataset consists of a reference or multiple references to a specific resource through a distinct descriptive statement, that statement is considered metadata (cf. Pomerantz, 2015, p. 26). In literature normally these resources are equated only as being 'other' data, which makes the standard metadata definition ´data about data´ probably the most used one (cf. Chen et al., 2018, p. 302; White & Breckenridge, 2014, p. 332). However, with the advent of the semantic web, these statements can also refer to real objects, e.g. a city or a person, with the goal to generate queryable facts about the world for machines to process (cf. Färber et al., 2016, p. 1; Gartner, 2016, p. 89 ff.). Metadata statements consist of so-called triples, which include a subject, a predicate and an object (i.e. the resource). Through this, information (i.e., the subject) can be stated about an object, where the sort of information is defined by the predicate. Due to many sorts of information, which can be stated through predicates, metadata types can be classified, most commonly between descriptive, administrative and structural metadata. Descriptive metadata provides information about a resource to be better findable either by humans or machines (cf. Kitchin, 2014a, p. 9). Administrative data includes technical metadata, which is often saved automatically in the background of digital devices like the GPS location and time of a digital picture made on a smartphone, rights metadata e.g., a Creative Commons license (*License*), and preservation metadata (cf. Gartner, 2016, p. 6 ff.). Structural metadata are descriptions of the organization and structure of an object. 'Content data' is the remaining part of the dataset after excluding the metadata; therefore, it is intended to represent the measured real-world phenomena.

4.16 Actor type

In discussions about the causes of increased data volume nowadays the term **machine-generated data** is sometimes used as an argument to explain this increase. This is due to the limitations of data production growth triggered by human activities since time constraints and limited population growth will finally also limit human-generated data growth. Therefore, the main reason for the current exponential data growth must be generated automatically from computers and digital devices without the explicit activities of humans. Although this is a reasonable argument, it doesn´t always make a clear distinction between machine-generated and human-generated datasets possible. Indeed, many problems arise in defining the actor type of a dataset. For example, if a person installs a camera to surveillance his house, the resulting datasets are clearly triggered by an explicit human action, although automatically gathered surveillance video data is considered to be machine-generated. However, if this argument is carried too far all data would become human-generated since humans decided to program devices to generate data. Therefore, **human-generated data** should be defined as data which is a direct primary result of an intended and manual human interaction with a digital device. Important are the terms 'direct', 'manual' and 'intended', which emphasize that only those datasets are human-generated, which are inevitable consequences of explicit human choice. This means that the raw image data generated by a human pressing the trigger on a digital camera is human-generated, the related technical metadata that is created in the event as well, is machine-generated. **Machine-generated data**, therefore, is data which is created by machines automatically and independent from direct activities triggered by a human actor or is a data byproduct of such an event. When discussing implied characteristics of datasets like objectivity of 'machine-generated data' it is important to exclude data byproducts like 'technical metadata' from the human-generated data class since the

human intention during such an event is not to generate, observe or modify those byproducts. 'Machine-generated data' mainly relates to 'sensory data', 'log data' and 'surveillance data' (*Generation purpose*) and is often unprocessed/raw (*Processing degree*) and unobtrusive (*Level of collecting obtrusiveness or bias*). 'Human-generated data' relates mainly to 'processed data', 'social interaction data', 'scientific-use data' and most of 'administrative data'.

4.17 Granularity

The term *granularity* or resolution is used to express the level of proximal measurement a dataset is holding with respect to a phenomenon or unit of analysis (cf. George et al., 2016, p. 1494). Historically, many technical advances over time have enabled data capturing on deeper granularities in respect to the measured objects, like higher camera resolution, increasing the number of pixels in graphical or video data, or by sensors capable of temporally measuring in faster frequencies, for instance, in temporal or log data. In relation to social behavior, data granularity often refers to the extent to which detailed individual behavior is visible in the data compared to showing only aggregated information grouping individuals together on specific criteria (**coarse-grained data**). The generation of **fine-grained data** on individual levels is one key aspect of the success of social media platforms enabling advertising companies to target and monitor individuals (called microtargeting) through personalized ads instead of advertising mostly undifferentiated without clear indication about the number of ad recipients or their characteristics making measuring of ad success speculative guesswork. Because fine granularity therefore can be a characteristic of social media datasets, big data proponents emphasize the opportunities of 'social media data' not only due its high scope (*Scope*), and high amount of sensitive information (*Sensitivity*), but also its fine granularity enabling social scientists and marketing and business intelligence department to answer a much wider range of research questions by identifying subgroups, hard-to-reach individuals or rare events typically hidden in sampled data or coarse-grained population data (cf. Staff et al., 2016, p. 11).

4.18 Relationality

"*Relationality concerns the extent to which different sets of data can be conjoined and how those can be used to answer new questions. Relationality is at the heart of relational databases…"* (Kitchin, 2014a, p. 74). Traditional relational databases are normally only discussed in the context of highly 'structural data' (*Machine interpretability*), where unique labelling and identification of spreadsheet rows of one table through Ids and keys ('indexical data') enable one-to-one or one-to-many interconnections to information ('attribute data') located inside other tables. Databases like this are easily queryable through SQL queries (*Access method*), where the data relationality enables a central and single point of data access to information about an object whose data is distributed between multiple tables. Due to its high demands on predefined structuredness and data quality, data in traditional relational databases must pass a so-called ETL process (Extract-Transform-Load), which includes data transformations like data cleaning or reformatting, making those datasets, on the one hand, highly predictable but on the other hand not 'raw' anymore (*Processing degree*) (cf. Kitchin, 2014a, p. 86). Since this inflexible reliance on 'structured data' cannot keep up with the increase of mostly 'unstructured data' like textual and visual data in the big data era, non-traditional databases (NoSQL databases) are getting more important nowadays supporting larger volumes of dynamic structured and unstructured datasets (cf. Chen et al., 2014, p. 186). Since these datasets, therefore, cannot be saved in traditional SQL relational databases, they are not considered **relational data** in the traditional technical view; however, deep granularity (*Granularity*) and embedded metadata (*Reference type*) make them highly relational on fundamental criteria since these characteristics (fine graininess and metadata) enable interlinking to other similar equipped datasets. Interlinks can, for instance, be built based on individual information which are unique to a person, group or institution, like social security numbers, central index keys, emails or social media account Ids; or they can be based on embedded metadata created automatically like time stamps, geo locations or IP addresses creating reference points by which these datasets become "*fundamentally networked*" (Boyd & Crawford, 2011, p. 2) and linkable. Although the term 'linked data' as a description might be useful here, it mostly refers to best practices enabling the machine-interpretability (*Machine interpretability*) of previously **non-relational** information published on the Internet in the context of the semantic web paradigm (*Reference type*).

4.19 Access method

The method with which data is accessed is highly critical for those scientists not creating their own internal 'research data' through case studies, observations, experiments or surveys and for companies accessing externally generated datasets ('external data' or 'externally generated data'), since the access method can influence the content and characteristics of resulting datasets in relation to its original true form. **Self-generated data**, 'first-party data', 'internal data' or 'internally generated data' correspond to 'made data' (*Level of collecting obtrusiveness or bias*) if not generated by other scientists or collected from scientific databases, whereas 'found data' as a defining feature of big data (cf.

Connelly et al., 2016, p. 10) needs to be accessed in technical processes, which either can be characterized as direct or indirect describing the relation between data receivers and data issuers or data controller.

Direct accessed data is data which has been accessed by an external actor without any limitations, restrictions or boundaries regarding data content and characteristics and, therefore, is a one-to-one copy of the requested dataset. This can be the case when actors have full data access to databases or systems, which themselves are not in direct control of those regarding system design, data flow, content, etc. This could be a scientist conducting a study in an organization and having full data access for the time of the study. However, for private organizations, this kind of access possibilities to external databases will rarely be the case. If, however, 'web data' is the data of interest, also web exports will belong to this class since those can create 1:1 copies of the presented 'web data' (*Accessibility*).

Indirect accessed data is data which is created within a process which contains some form of technical or logical filters, boundaries, limitations or moderations lying in between the data collector and the physical location of the requested datasets, preventing direct access. This means, that the resulting datasets, in most cases, are not 1:1 copies of the requested or intended data regarding content, *scope*, *format*, *processing degree* or other characteristics. For the accessing party it is of fundamental importance to understand the logic behind those filters, boundaries, limitations or moderations to elaborate in which form or degree accessed data is varying compared to the real data to estimate data quality and, therefore, finally, the risk of using the data in decision making or the validity in a scientific study.

Many indirect data access possibilities are possible. Three important ones are further explained here:

Second-party data is data which is acquired from partner organizations. Compared to 'first-party data', which is 'internally generated data' and therefore 'self-generated data', 'second-party data' is provided directly by another organization. Data content, *format*, *processing degree*, delivery form and frequency are normally defined in a data contract between the issuer and receiver. This means that although the receiving party does not have direct access to the original data or database, the direct relationship to the data provider and concrete conditions of delivery makes resulting data quality relatively easy to assess, normally high and stable. One example would be a retailer providing customer sales data to a client regarding the client's product line.

Third-party data is data that is acquired from third-party organizations, also called data aggregators (cf. Deloitte, 2012, p. 3). Their business is to collect or acquire data from multiple sources, process and compile them ('compiled data' (cf. Marr, 2015, p. 87)) in some form like clean, normalize, aggregate, merge or analyze it (Processing degree), and sell the 'processed data' and other data-related services to organizations. To elaborate on accuracy, reliability, validity or data quality on those datasets might be difficult since many processing steps might be involved and since the third-party provider's self-interest is not to provide full information about how or where data is collected and how the data is technically processed.

API data is data generated by accessing databases or services through an API (Application Programming Interface). APIs here are interfaces controlling the outflow of data defined by the API provider, which is the organization in control of the database. Therefore, APIs provide an access point through which external organizations can query the database (cf. Snee, 2016, p. 19). Since the API provider wants to control what and how much data a user can collect and how often a user can query the database, limitations and restrictions are normally in place preventing direct and full access (cf. Olmedilla et al., 2016, p. 79). Social media platforms like Twitter for instance offer different API access points to search for publicly available tweets (cf. Twitter API). Those can differ for example, in how many tweets can be extracted in a period or how many days back a user can extract tweets (e.g., 7 days vs. 31 days). Important for users of APIs is that limitations and restrictions designed into the API can lead to data collection bias, especially regarding *scope* (cf. Morstatter & Liu, 2017, p. 2 ff.). When a query delivers more results than the API gate is allowed to pass back to the requesting system or user, just a data sample is delivered. This is especially problematic when there is no built-in logic in place to make sure the sample is completely randomized. If the behavior of the API in these cases is unknown to the requesting user, the data quality of the retrieved data might be impossible to estimate, since the relation between sample and population is unknown (*Scope*).

4.20 Analysis method fitness

In the field of data analytics and data mining often the methods applied to datasets are differentiated by the type of questions the analysis procedures are aiming to answer and therefore by the types of techniques applied. In literature, analytics are mostly differentiated between descriptive, predictive and prescriptive methods (cf. Kokina et al., 2017, p. 51; Phillips-Wren & Hoskisson, 2015, p. 88; Delen & Demirkan, 2013, p. 361) and sometimes also additionally between diagnostic (cf. Kokina et al., 2017, p. 51), exploratory (cf. Chen et al., 2018, p. 308), argumentative (cf. Chen et al., 2018, p. 309) or inferential methods (cf. Chen et al., 2014, p. 191). This framework sticks with the most common top-level types: descriptive, predictive and prescriptive.

Descriptive analytical methods aim to describe, understand, aggregate, summarize or visualize key information inside datasets, often through simple mathematical operators or statistical modelling. It is the first step in data analytics and constantly applied in business reporting for operational control or to indicate performances. While descriptive methods aim to create an unbiased and authentic picture of past events, predictive methods use datasets to predict or

forecast future events, data outputs and values. Especially machine learning processes aim to automatically build generalized models which fit the used data to calculate the likelihood of certain future events. Prescriptive analytics is an extension of predictive analytics by calculating through possible alternative input values to identify the best possible data outcomes. It is based again on predictive models, which, however, should deliver recommendations about the best possible course of action to decision-makers.

This dimension is termed *Analysis method fitness* because for datasets to be useful for these kinds of analytical methods and, therefore, to be valid to answer descriptive, prescriptive or prescriptive questions, they need to fulfill certain criteria. However, for data to be labelled as **descriptive data** no specific criteria need to be fulfilled except that the dataset doesn't exclusively exists of random meaningless values since, in harmony with the definition of data as a 'representation of evidence', it always should describe something; otherwise, it has no value for machines or humans. **Predictive data and prescriptive data** are always also 'descriptive data', but also need to fulfill additional criteria to be useful for predictive and prescriptive analytical purposes. These criteria are highly context-specific with respect to what data output exactly should be predicted and which accuracy is acceptable or needed for the analyst or decision maker. On the positive side, it could mean that a predictive model has proven that specific visible data variables inside a dataset could be used as leading indicators to predict specific output variables in sufficient accuracy to be operationally implementable. On the negative side, it could mean that the number of observations/samples in a dataset is too small for reliable predictions when *feature dimensionality* in the dataset is too high. To summarize, 'predictive data' is data fitting predictive analytical purposes, 'prescriptive data' is data fitting prescriptive analytical purposes and 'descriptive data' is all data containing meaningful content for humans or machines.

4.21 Feature dimensionality

Dimensionality is a concept adapted from mathematics and statistics, which in relation to data characteristics, can either describe the total amount of dimensions, interpreted as features, a dataset possesses (from a low to a high number of features) or the number of underlying measurements scales a dataset possesses to measure concepts (between unidimensional to multidimensional (*Measurement dimensionality)*). Data dimensionality normally relates to structured datasets (*Machine Interpretability*), whereas unstructured datasets, e.g., 'textual data', first need to be methodologically converted into a structured spreadsheet-kind of form to consist of clear observations/samples represented as rows and attributes/features/dimensions as columns (cf. Gupta, 2013, p. 2602 ff.).

High-dimensional and **low-dimensional data** can be interpreted as a relative term comparing the number of dimensions/attributes/features of one dataset to other datasets or to a general average data dimensionality not further defined, or as absolute characterizations describing data as high-dimensional when it consists of more dimensions/features/attributes than observations/samples and as low-dimensional vice versa. This could mean that datasets consisting of two observations/rows and three attributes/features/columns could be considered as high-dimensional, but since those kinds of datasets are practically rarely useful, high-dimensional normally means that the number of dimensions is absolute and relatively high. High feature dimensionality can be a problem for data mining and machine learning tasks like classification due to the increasing amount of computational power needed to train a model and due to the exponentially increasing numbers of observations needed with increasing dimensionality required to generalize a model. This problem is known as the 'curse of dimensionality' and can be addressed by dimensionality reduction techniques like principal component analysis to identify dominant patterns and correlations in the dataset.

4.22 Measurement dimensionality

Measurement dimensionality does not relate to the technical computer-oriented definition of a dimension, which would be just a column in a spreadsheet representing a feature/attribute of an observation/row, but to the number of measurements scales the data possess to describe or measure constructs or concepts we use to socially understand the world. When a dataset uses only one single measurement scale to measure an attribute or phenomenon this data is **unidimensional data**. Evidence measured in physical quantities is, in general, expressed in unidimensional datasets, e.g., meter for height or kilogram for weight, where complex phenomena like pain, poverty, well-being, etc., often can only be described or measured in multidimensional. Therefore, **multidimensional data** is data in which concepts or phenomena are measured on multiple measurement scales. However, the term 'multidimensional' must be used with caution since this term also often is used to describe *feature dimensionality*, meaning just more than one feature existent in a dataset. Unidimensional datasets, in general, are easier to process and interpret than multidimensional datasets, but, dependent on the measured phenomenon, they might not be able to fully represent the phenomena under investigation.

4.23 Generation velocity

In relation to data, velocity in general, is mostly not being considered to be a characteristic of a dataset itself but rather a quality of a sensory or monitoring system to capture, observe or measure signals of the world in high frequencies and in case of 'real-time data' to save those observations simultaneously as values in datasets. Velocity as an expression can

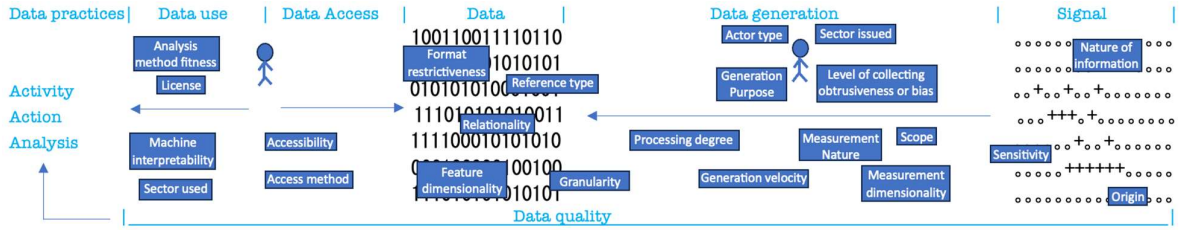
also additionally refer to the speed at which these data values automatically get analyzed or/and reported for decision-making (cf. Arnaboldi et al., 2017, p. 764). Therefore, **high-velocity data** is data, which is created in a high frequency sensory or monitoring system enabling up to real-time availability of current measured values either to machines or humans. This means that both the time intervals between consecutive measurements are very low (for example milliseconds), and the time lags of measurements to measured values availability are very low as well. It could be a sensory system, which constantly oversees the performance of a machine and reports key metrics of it to a monitoring screen close to real-time or a company monitoring social media activities close to real-time by reporting relevant posts as soon as available to social media managers. **Low-velocity data** on the other hand is data, which is created in a sensory or monitoring system where either the time intervals between recurring measurements of a phenomenon are high or the system does not allow the fast availability of measured values to machines or humans. It is obvious that a system creating high-velocity instead of low-velocity data allows faster response times by data consumers, and those faster responses can have significant value for example in the financial sector.

5 Data dimension references

After illustrating all 23 data dimensions, we can now think about where they do fit in the overall data value chain model explained in the introduction. The data value chain model depicts the key steps from some form of signal or signal variation towards conducted data practices in the form of activities, actions or analyses. In general, we can depict five reference areas to which the data dimensions can relate to, the signal, which is intended to be represented in data, the data generation process, the data accessing process, the data uses or the involved actors.

Figure 3 shows the rough locations of the explained data dimensions inside the overall data process. The references are only sometimes clear-cut, and data dimensions can relate to multiple areas (Table 5).

Figure 3: Data dimensions inside the data value chain



Signal: *Nature of information*, *origin* and *sensitivity* are located here. *Nature of information* points towards the importance for data users to understand the underlying nature of phenomena intended to be represented in data and not to confuse it with the *measurement nature*, which describes different types of data generation and measurement methods. *Origin* describes if the signal in its original state is already in a digital form or not, and *sensitivity* points towards the phenomena considered to have a sensitive character or not.

Generation: Most of the data dimensions are located in the data generation area. Part of those can be used to describe technical characteristics of the resulting datasets like format characteristics, *feature dimensionality, granularity, relationality* etc. More closely located to the signal area is *scope*, which especially indicates the importance of understanding the relationship between what is measured and what the population of interest (signal) constitutes. Also, *measurement nature* and *measurement dimensionality* are located here since specific types of phenomena to be collected require specific recoding procedures to create valid representations. The way how things are measured can sometimes be indicated by the phenomena itself.

Actor-related: In the data generation process, we have four dimensions close to the actor initiating the data generation. *Actor type* points towards the importance to understand characteristics of the actor triggering the data generation and *generation purpose* towards the actor's primary purpose or objective to generate the data. *Level of collecting obtrusiveness or bias* points out that data can be artificially created by an actor ('made data'), leading to bias or be unobtrusiveness be collected by an actor ('found data'). *Sector issued* differentiates public and private data pointing towards the importance to understand the sector in which data is generated. Different institutions and organizations can hereby be understood as different actors in which data is generated.

Table 5: Data dimensions key area references

No	Data dimensions and references	Signal	Generation	Accessing	Using	Actor
1	*Origin*	x				
2	*Sector issued*		x			x
3	*Sector used*				x	
4	*Accessibility*			x		
5	*Generation purpose*		x			x
6	*Level of collecting obtrusiveness or bias*		x			x
7	*Format restrictiveness*		x			
8	*Processing degree*		x			
9	*Machine interpretability*				x	
10	*Measurement nature*	x	x			
11	*Sensitivity*	x				
12	*Scope*	x	x			
13	*License*				x	
14	*Nature of information*	x				
15	*Reference type*		x			
16	*Actor type*		x	x	x	x
17	*Granularity*		x			
18	*Relationality*		x			
19	*Access method*			x		
20	*Analysis method fitness*				x	
21	*Feature dimensionality*	x	x			
22	*Measurement dimensionality*		x			
23	*Generation velocity*		x			

Access: *Accessibility* and *Access method* are located here, emphasizing the importance of understanding the relationship and possible content difference between the recorded dataset and the accessed dataset, having implications for the accessing actor regarding the validity of their use cases.

Use: *Analysis method fitness*, *license*, *machine interpretability* and *sector used* are located here, which describe use qualities or limitations of datasets.

All data dimensions are located in the data value chain and, therefore, have an influence on data practices. Reasons for non-functioning data practices can lie in misconceptions of actors about data qualities along one or more data dimensions. These misconceptions can for instance relate to having a wrong perception about the *nature of*

information stored in the dataset or wrong ideas about who the data generation actors are and what motivates them. It can relate to being unaware of the degree of processing already involved in the data generation processes, it can relate to an unfitness between *measurement nature* or *dimensionality* and the type or content of signal intended to be represented. It can relate to a *granularity*, which is hiding details of the measured signal. It can relate to access and use restrictions for the actors being unable to see the whole picture. These examples show that along the data production infrastructure, multiple areas exist where data quality might significantly be affected.

In information systems literature, data quality is generally defined as the "*fitness for use*" (Attard et al. (2015), p. 402; Wang & Strong (1996), p. 6); and it depends on the actual use cases and intended practices (cf. Färber et al. (2016), p. 3). In pragmatic constructivism functioning practices need the existence of facts constructed based on sound evidence (cf. Nørreklit, 2017, p. 35 f.). Therefore, considering data as a 'representation of evidence', data quality can be interpreted as the fitness of that evidence for creating valid facts and it can be low when the evidence is "*vague and inconclusive regarding what precisely generated it*" (Nørreklit, 2017, p. 35). That means that invalid reality constructions, failing practices and confusion in participating actors can arise, when those rely on increasingly digitally produced evidence through complex technical systems, which can hide existing biases, defects or inconsistent production procedures. In those cases, actors increasingly will question the validity of the used datasets by pragmatically identifying what is or is not working. A learning circle between actors develops. Here, pragmatic constructivism emphasizes the importance of language establishing "*an integrative learning theory of truth that enables us to theoretically point to problems of validity*". This leads to the question if the data language used by actors creates functioning reality constructions or not? Does it establish learning, cooperation and understanding?

This framework provides data classifications with the goal of establishing a shared understanding of universal data concepts, which finally can lead to a functioning data language between actors aiming to identify risks inside the data production infrastructure. This is because, similar to accounting, which serves as the universal financial language in a capitalistic world, data nowadays serves as the universal medium of information. Since the history of digital data practices is much shorter compared to the one of accounting practices a universal standardized data language might evolve in a similar way over time. Currently, however, concepts with low conceptual quality consume lots of space in literature, i.e. big data, which clearly do not contribute to producing successful practices. Like accounting where the financial value of companies can be represented through accounting numbers established by a shared understanding about what those accounting numbers mean, a data language could also develop, establishing a shared understanding about the value, validity and quality of datasets. By providing 23 views on how we can differentiate, compare and evaluate datasets, this paper therefore contributes to this objective.

As an example, Table 6 shows how the framework provides the possibility to concretely describe and conceptualize the different variations and elements of the big data definitions shown in the introduction by the respected data dimensions and classes introduced in this paper.

6 Conclusion

The hype of many fuzzy data-related concepts in literature and business is unbroken. Big data, artificial intelligence, digitalization or being data-driven narratives are everywhere. Due to their multidimensional and vague definitions the practical use of those is limited to actors aiming to succeed in a more complex and less understandable environment. Data practices have been controllable and more understandable to actors in the past since limited technical possibilities restricted users to the use of a limited number of purely internally generated datasets. However, it is argued here that going forward in an era of datafication and dataism, where "*the attempt to capture everything as data*" (Kitchin, 2014a, p. 130) is the growing ideology and primary driving force to generate new scientific knowledge or to produce facts for governmental or organizational decisions, we need new and more sophisticated concepts of how we understand data. When the drive to datafy everything that exists, which as a realist perspective is equal to the world, as data, then the data space in which humans live and act is expanding and getting more complex over time. To deal with this increasingly data-centric environment, too simple or vague concepts cannot capture this complexity and, therefore, can produce reality constructions which do not work in practice. The high number of data dimensions and classes in this framework capturing some of this complexity offers an alternative to the vague V-concept of big data or the non-specific concept of being 'data-driven' by providing a possibility to exactly describe and compare used datasets or data strategies.

The framework provides a distinct data language by offering 23 different data dimensions with 63 first-level data classes to specifically describe, identify and compare certain data characteristics. It can produce precise communication between data experts like business intelligence managers, accountants or scientists managing the datasets and those decision makers using data-dependent conceptualized facts to create successful actions like managers or politicians. The dimensions relate to different areas inside the data value chain, which is a model to show the different stages between the 'analog' signal (world), the stored digital data and data practices. Data is not just there; it needs to go through different processes of data generation and data accessing before anybody can use them. Those processes are set up by actors, which, visible or not, define the data infrastructure, which, to a high degree, might be opaque, especially when

the data is not generated internally. Therefore, mindful actors need to be aware of the different areas of risks involved, when using data produced in such a complex environment, which especially relate to other involved actors and the data generation, accessing and using procedures. In this play of technical and human interactions they need a data language based on a set of specific criteria, which is able produce meaning and understanding. Therefore, the framework aims to provide some light into a field of confusing concepts and terminologies and, at the same time, calls for a more universal data language in theory and practice.

Table 6: Big data definitions and relating data dimensions and classes

Data dimension	Based on definition	Data class
Origin	"…collected through devices and technologies such as … increasingly, WiFi sensors, electronic tags." (Chua, 2013, p. 10)	Digital born data
Measurement nature	"…text…" (George et al., 2016, p. 1493)	Textual data
Measurement nature	"…videodata…" (Calvard, 2016, p. 67); "…videos…" (George et al., 2016, p. 1493)	Video data
Measurement nature	"…audio…" (Chen et al., 2014, p. 173)	Audio data
Measurement nature	"…digital trace…" (George et al., 2016, p. 1493)	Log data
Machine interpretability	"…size beyond the ability of typical database software tools to capture, store, manage and analyze." (Moffitt & Vasarhelyi, 2013, p. 4) "…multi-structure data…" (Moffitt & Vasarhelyi, 2013, p. 5), "…unstructured data…" (Chen et al., 2014, p. 171; Gandomi & Haider, 2015, p. 138), "…NoSQL…" (Ward & Barker, 2013, p. 2; Akoka et al., 2017, p. 106), …unstructured in nature…" (Kitchin, 2014a, p. 68)	Unstructured data
Machine interpretability	"…structured…in nature…" (Kitchin, 2014a, p. 68), "…structured…" (Ylijoki & Porras, 2016, p. 74)	Structured data
Generation velocity	"…high velocity capture…" (Akoka et al., 2017, p. 106), "…created in or near real-time…" (Kitchin, 2014a, p. 68), "…speed at which new data is generated…" (Marr, 2017, p. 87), "…speed and immediacy of data creation…" (Calvard, 2016, p. 66), "…speed at which data is generated." (Phillips-Wren & Hoskisson, 2015, p. 90)	High velocity data
Generation purpose	"…social media content…" (Calvard, 2016, p. 67), "…social media…" (Chua, 2013, p. 10), "…social media data…" (Chen et al., 2012, p. 1165)	Social interaction data
Generation purpose	"…sensor … data…" (Chen et al., 2012, p. 1165), "…from sensors…" (George et al., 2016, p. 1493)	Sensory data
Generation purpose	"…administrative data…" (Arnaboldi et al., 2017, p. 764)	Administrative data
Scope	"…exhaustive in scope, striving to capture entire populations or systems (n=all)…" (Kitchin, 2014a, p. 68)	Population data
Data quality (not one of the defining data dimensions, but rather a result)	"…challenge of managing data quality…" (Buhl et al., 2013, p. 68), "…inconsistencies…" (Japkowicz & Stefanowski, 2016, p. 3), "…messiness or trustworthiness of data." (Marr, 2017, p. 87), "…data accuracy and reliability of data…" (Janvrin & Weidenmier Watson, 2017, p. 3), "…quality of data…" (Japkowicz & Stefanowski, 2016, p. 3), "…uncertainty surrounding data integrity and trustworthiness…" (Phillips-Wren & Hoskisson, 2015, p. 90), "…unreliability…" (Gandomi & Haider, 2015, p. 139)	Low quality data
Relationality	"…integrated from different sources and joint together." (Sparks et al., 2016, p. 33), "…embeddedness…" (Chen & Yu, 2018, p. 17), "…uniquely indexical in identification; relational in nature…" (Kitchin, 2014a, p. 68)	Relational data
Actor type	"…automatically machine obtained/generated…" (Moffitt & Vasarhelyi, 2013, p. 4 f.)	Machine-generated data
Analysis method fitness	"…big data is about predictions." (Mayer-Schönberger & Cukier, 2013, p. 11), "…draw inferences from correlations not possible with smaller datasets." (Moffitt & Vasarhelyi, 2013, p. 5)	Predictive data
Format restrictiveness	"…cannot be managed by standard software…" (Chua, 2013, p. 11)	Proprietary data
Reference type	"…embeddedness, i.e., the places, the spaces within … social interaction are operated." (Chen & Yu, 2018, p. 17), "…temporally and spatially referenced…" (Kitchin, 2014a, p. 68)	Metadata
Granularity	"…fine-grained in resolution…" (Kitchin, 2014a, p. 68)	Fine-grained data

References

6, P., & Bellamy, C. (2012). **Principles of Methodology Research Design in Social Science**. London, UK, Sage. Ackland, R. (2013). **Web social science: concepts, data and tools for social scientists in the digital age**. London, UK, SAGE.

Ackland, R. (2013). **Web social science: concepts, data and tools for social scientists in the digital age**. London, UK, SAGE.

Akoka, J., Comyn-Wattiau, I., & Laoufi, N. (2017). Research on Big Data – A systematic mapping study. **Computer Standards & Interfaces**, 54, 105-115.

Arnaboldi, M., Busco, C., & Cuganesan, S. (2017). Accounting, accountability, social media and big data: revolution or hype? **Accounting, Auditing and Accountability Journal**, 30(4), 762-776.

Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. **Government Information Quarterly**, 32(4), 399-418.

Ball, A. (2014). How to License Research Data. **DCC How-to Guides**. Edinburgh: Digital Curation Centre. Available at: https://www.dcc.ac.uk/guidance/how-guides/license-research-data

Bauder, J. (2014). **The Reference Guide to Data Sources**. Chicago, USA, American Library Association.

Blundell, D., Lin, C., & Morris, J.X. (2018). Spatial Humanities: An Integrated Approach to Spatiotemporal Research. In S. Chen (ed.), **Big Data in Computational Social Science and Humanities**, 263-288. Taipei, Taiwan, Springer.

Boyd, D., & Crawford, K. (2011). Six provocations for Big Data. **SSRN Electronic Journal**, Sept. 2011, 1-17.

Buhl, H., Röglinger, M., & Moser, F. (2013). Big Data - A Fashionable Topic with(out) SustainableRelevance for Research and Practice? **Business & Information Systems Engineering**, 2/2013, 65-69.

Calvard, T.S. (2016). Big data, organizational learning, and sensemaking: Theorizing interpretive challenges under conditions of dynamic complexity. **Management Learning**, 47(1), 65-82.

CESSDA. Available at: https://www.cessda.eu/

CESSDA Webinar (2018). Data in Europe: Poverty. Available at: https://www.youtube.com/watch?v=sk6DCg-Z_Gk

Chang, C., Kayed, M.R., Girgis, M.R. & Shaalan, K.F. (2006). A Survey of Web Information Extraction Systems. **IEEE Transactions on Knowledge and Data Engineering**, 18(10), 14,11-1428.

Chen, H., Chiang, R.H.L., & Storey V.C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. **MIS Quarterly**, 36(4), 1165-1188.

Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A Survey. **Mobile Networks and Applications**, 19, 171-209.

Chen, P., Cheng, Y., & Chen, K. (2018). Analysis of Social Media Data: An Introduction to the Characteristics and Chronological Process. In S. Chen (ed.), **Big Data in Computational Social Science and Humanities**, 297-321. Taipei, Taiwan, Springer.

Chen, S., & Yu, T. (2018). Big Data in Computational Social Sciences and Humanities: An Introduction. In S. Chen (ed.), **Big Data in Computational Social Science and Humanities**, 1-25. Taipei, Taiwan, Springer.

Chowdhury, H. (2024). **Google just can't seem to catch a break**. Available at: https://www.businessinsider.com/google-cant-catch-break-despite-having-ai-gemini-to-showcase-2024-2

Chua, F. (2013). **Big data: its power and perils**. ACCA. Available at: https://www.accaglobal.com/bigdata

Chuvakin, A.A., Schmidt, K.J., & Phillips, C. (2013). **Logging and Log management: The Authoritative Guide to Understanding the Concepts Surrounding Logging and Log Management**. Waltham (Massachusetts), USA, Syngress.

Connelly, R., Playford, C.J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. **Social Science Research**, 59, 1-12.

Council of the European Union (2016). **Council Position (EU) 6/2016 of the council at first reading with a view to the adoption of a Regulation of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)**. Available at: https://eur-lex.europa.eu/legalcontent/EN/TXT/PDF/?uri=CELEX:52016AG0006(01)

Cowls, J., & Schroeder, R. (2015). Causation, Correlation, and Big Data in Social Science Research. **Policy & Internet**, 7, 447-472.

Cuzzocrea, A., Song, I., & Davis, K.C. (2011). Analytics over large-scale multidimensional data: The big data revolution! **14th International Workshop on Data**, 101-104.

Deloitte (2012). **Open growth: Stimulating demand for open data in the UK**. Available at: https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/deloitte-analytics/open-growth.pdf

Dixson-Decélve, S., Gaffney, O., Ghosh, J., Randers, J., Rockström, J., & Stoknes, P. E. (2022). **Earth for all: A Survival Guide for Humanity**. Gabriola Island, Canada, New Society Publishers.

Delen, D., & Demirkan, H. (2013). Data, information and analytics as services. **Decision Support Systems**, 55, 359-363.

Evans, M.R., Oliver, D., Yang, K., Zhou, X., Ali, R.Y., & Shekhar, S. (2019). Enabling Spatial Big Data via CyberGIS: Challenges and Opportunities. In S. Wang & M.F. Goodchild (eds.), **CyberGIS for Geospatial Discovery and Innovation**, 143-170. Dordrecht, The Netherlands, Springer.

Färber, M., Ell, B., Menne, C., Rettinger, A., & Bartscherer, F. (2016). Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. **Semantic Web**, 1, 1-45.

Felt, M. (2016). Social media and the social sciences: How researchers employ Big Data analytics. **Personality and Social Psychology Bulletin**, 3(1), 537-548.

Floridi, L. (2008). Data. In W.A. Darity (ed.), **International Encyclopedia of the Social Sciences**, 2nd edition. Detroit: Macmillan. Available at: https://philpapers.org/archive/FLOD-2.pdf

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

Gartner, R. (2016). **Metadata – Shaping Knowledge from Antiquity to the Semantic Web**. Switzerland, Cham, Springer.

George, G., Osinga, E.C., Lavie, D., & Scott, B.A. (2016). Big data and data science methods for management research. **Academy of Management Journal**, 59(5), 1493-1507.

Gupta, V. (2013). Extracting Facts And Dimensions From Unstructured Data For Business Intelligence. **International Journal of Engineering Research & Technology**, 2(7), 2602-2606.

Heath, T., & Bizer, C. (2011). Linked Data: Evolving the Web into a Global Data Space. **Synthesis Lectures on the Semantic Web: Theory and Technology**, 1:1, 1-136.

Henriksen, L.B. (2016). Change, concepts and the conceptualising method. **Proceedings of Pragmatic Constructivism**, 6(2), 29-33.

Investor.gov. **Form10-Q**. Available at: https://www.sec.gov/fast-answers/answersform10qhtm.html

IPUMS USA FAQ. **Basic concepts – What are microdata.** Available at: https://usa.ipums.org/usa-action/faq#1

Jaakkola, H., Mäkinen, T., & Eteläaho, A., (2014). Open data: opportunities and challenges. **Proceedings of the 15th International Conference on Computer Systems and Technologies**, 25-39.

Janvrin, D. & Weidenmier Watson, M. (2017). "Big Data": A new twist to accounting. **Journal of Accounting Education**, 38, 3-8.

Japkowicz, N., & Stefanowski, J., (2016), A Machine Learning Perspective on Big Data Analysis, In N. Japkowicz & J. Stefanowski (eds.), **Big Data Analysis: New Algorithms for a New Society**, 1-31. Cham, Switzerland, Springer.

Kitchin, R. (2014a). **The data revolution. Big data, open data, data infrastructures & and their consequences**. London, UK, SAGE.

Kitchin, R. (2014b). Big Data, new epistemologies and paradigm shifts. **Big Data & Society**, April-June 2014: 1-12.

Knoblauch, H., Baer, A., Laurier, E., Petschke, S., & Schnettler, B. (2008). Visual Analysis. New Developments in the Interpretative Analysis of Video and Photography. **Forum: Qualitative Social Research**, 9(3).

Kokina, J., Pachamanova, D., & Corbett, A. (2017). The role of data visualization and analytics in performance management: Guiding entrepreneurial growth decisions. **Journal of Accounting Education**, 38, 50-62.

Kure, N., Nørreklit, H., & Raffnsøe-Møller, M. (2017). Language Games of Management Accounting – Constructing Illusions or Realities? In H. Nørreklit (ed.), **A Philosophy of Management Accounting**, 211-224. New York, USA, Routledge.

Lajaunie, C., Mazzega, P., & Boulet, R. (2018). Health in Biodiversity-Related Conventions: Analysis of a Multiplex Terminological Network (1973-2016). In S. Chen (ed.), **Big Data in Computational Social Science and Humanities**, 165-182. Taipei, Taiwan, Springer.

Little, K. (1973). Ethical principles in the conduct of research with human participants. **American Psychologist**, 28(1), 79-80.

Lyon, D., Doyle, A., & Lippert, R. (2012). Introduction. In A. Doyle, R. Lippert, & D. Lyon (eds.) **Eyes Everywhere – The Global Growth of Camera Surveillance**. New York, USA, Routledge.

Mari, L. (2007). Measurability. In M. Boumans (ed.). **Measurement in economics**. Elsevier.

Marr, B. (2015). **Big data: using smart big data, analytics and metrics to make better decisions and improve performance**. Chichester, UK, Wiley.

Marr, B. (2017). **Data Strategy: How to profit from a world of big data, analytics and the internet of things**. London, UK, Kogan Page.

Mauro, S. G., Cinquini, L., Malmmose, M., & Nørreklit, H. (2024). University research by the numbers: Epistemic methods of using digitized performance measures and their implications for research practices. ***Financial Accountability & Management***, 40, 58–84.

Mayer-Schönberger, Cukier (2013). **Big data: A Revolution That Will Transform How We Live, Work and Think**. London, UK, John Murray.

Mazzocchi, F. (2015). Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science. **EMBO reports**, 16(10), 1250-1255.

McGivney, C. (2018). Sifting Through the Sets: The Significance and Availability of Open Data. In Perry, S. M. (ed.), **Maximizing Social Science Research Through Publicly Accessible Data Sets**, 1-22. Hershey PA, USA, IGI Global.

Mergel, I., Edelmann, N., & Haug, N. (2019). Defining digital transformation: Results from expert interviews. **Government Information Quarterly**, 36 (2019) 101385, 1-16.

Moffitt, K.C., & Vasarhelyi, M.A. (2013). AIS in an Age of Big Data. **Journal of Information Systems,** 27(2), 1-19.
Morstatter, F., & Liu, H. (2017). Discovering, assessing, and mitigating data bias in social media. **Online Social Networks and Media**, 1, 1-13.

Narayanan, A., & Shmatikov, V. (2008). Robust De-anonymization of Large Sparse Datasets. **2008 IEEE Symposium on Security and Privacy (sp 2008)**, 2008, 1-5.

Nørreklit, L. (2017). Actor-reality construction. In H. Nørreklit (ed.), **A Philosophy of Management Accounting**, 23-71. New York, USA, Routledge.

Olmedilla, M., Martínez-Torres, M.R., & Toral, S.L. (2016). Harvesting Big Data in social science: A methodological approach for collecting online user-generated content. **Computer Standards & Interfaces**, 46, 79-87.

Onwuegbuzie, A.J., & Collins, K.M. (2007). A Typology of Mixed Methods Sampling Designs in Social Science Research. **The Qualitative Report**, 12(2), 281-316.

Ørmen, J., & Thorhauge, A.M. (2015). Smartphone log data in a qualitative perspective. **Mobile Media and Communication**, 3(3), 335-350.

Parthasarathy, S., & Tashev, I. (2018). Convolutional Neural Network Techniques for Speech Emotion Recognition. **2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)**, 2018, 1-5.

Phillips-Wren, G., & Hoskisson, A. (2015). An analytical journey towards big data. **Journal of Decision Systems**, 24(1), 87-102.

Pomerantz, J. (2015). **Metadata**. Cambridge (MA), USA, MIT Press.

Richins, G., Stapleton, A., Stratopoulos, T.C., & Wong, C. (2017). Big Data Analytics: Opportunity or Threat for the Accounting Profession? **Journal of Information Systems**, 31(3), Fall 2017, 63-79.

Rogers, R. (2013). **Digital Methods**. Cambridge (MA), USA, MIT Press.

Sayogo, D.S., & Pardo, T.A. (2013). Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. **Government Information Quarterly**, 30(1), 19-31.

Shah, T. H. (2018). Big Data Analytics in Higher Education. In Perry, S. M. (ed.), **Maximizing Social Science Research Through Publicly Accessible Data Sets**, 38-61. Hershey PA, USA, IGI Global.

Sharma, A., Sharma, S., & Gupta, D. (2021). A Review of Sensors and Their Application in Internet of Things (IOT). **International Journal of Computer Applications**, 174 (24), 27-34.

Snee, H., Hine, C., Morey, Y., Roberts, S., & Watson, H. (2016). Part II – Combining and Comparing Methods – Introduction to Part II. In H. Snee, C. Hine, Y. Morey, S. Roberts & H. Watson (eds.), **Digital Methods for Social Science – An Interdisciplinary Guide to Research Innovation**, 67-70. New York, NY, Palgrave Macmillan.

Sparks, R., Ickowicz, A., & Lenz, H.J. (2016). An Insight on Big Data Analytics. In N. Japkowicz & J. Stefanowski (eds.), **Big Data Analysis: New Algorithms for a New Society**, 33-48. Cham, Switzerland, Springer.

Staff, C., King, H., Roberts, M., Pannell, S., Roberts, D., Wilson, N., Mann, R., & Cooper, A. (2016). **Using social media for social research: an introduction**. Social media research group. Available at: https://dera.ioe.ac.uk//26600/

Torra, V. (2017). **Data Privacy: Foundations, New Developments and the Big Data Challenge**. Cham, Switzerland, Springer.

Twitter API. https://developer.twitter.com/en/docs/twitter-api

Van Asbroeck, B., Debussche, J., & César, J. (2017). **Building the European Data Economy: Data Ownership**. Available at: https://sites-twobirds.vuture.net/1/773/landing-pages/white-paper-form.asp

Verma, P. (2023). **The rise of AI fake news is creating a 'misinformation superspreader'**. Available at: https://www.washingtonpost.com/technology/2023/12/17/ai-fake-news-misinformation/

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

Wang, R.Y., & Strong, D.M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. **Journal of Management Information Systems**, Spring 1996, 12(4), 5-34.

Ward, J.S., & Barker, A. (2013). Undefined By Data: A Survey of Big Data Definitions. **ArXiv**, abs/1309.5821.

Warren, Jr., J.D., Moffitt, K.C. & Byrnes, P. (2015). How Big Data Will Change Accounting. **Accounting Horizons**, 29(2), 397-407.

White, P., & Breckenridge, R.S. (2014). Trade-Offs, Limitations, and Promises of Big Data in Social Science Research. **Review of Policy Research**, 31(4), 331-338.

Woollard, M. (2014). Administrative Data: Problems and Benefits. A perspective from the United Kingdom. In A. Duşa, D. Nelle, G. Stock & G.G. Wagner (eds.), **Facing the Future: European Research Infrastructures for the Humanities and Social Sciences**, 49-60. Berlin, Germany, SCIVERO Verlag.

Ylijoki, O. & Porras, J. (2016). Perspectives to Definition of Big Data: A Mapping Study and Discussion. **Journal of Innovation Management**, 4(1), 69-91.