universität wien

MASTERABEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Data is overrated.
An analysis of the promise of data science to drive better decisions."

verfasst von / submitted by

Mag. Michael Hafner

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien, 2023 / Vienna 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 944
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Interdisziplinäres Masterstudium Wissenschaftsphilosophie und Wissenschaftsgeschichte
Betreut von / Supervisor:	Univ.-Prof. Tarja Knuuttila, MSc M.Soc.Sc PhD

Data is overrated

An analysis of the promise of data science to drive better decisions.
Predict less, explain more.

Abstract

Data science drives better decisions. That is the promise of many data science textbooks. Other concepts see data science as the abandonment of hypotheses, models and causality - statistical correlation replaces all that.
Decisions, however, are not just at the end of the data science process; they run from the initial questions, to deciding what to consider as data, to technical issues of data modeling and analysis. Decisions here depend on purpose, question, and context.
I argue that data, as used in databases and data models, should themselves be considered as models and outcomes of modeling processes. They are created and structured artifacts that were generated under specific conditions and can provide precise answers under specific conditions.
Data create their own universe with their own rules of meaning and validity. This thesis is directed both against the notion of data as a royal road to better decisions and against the notion of a bias that subsequently distorts data and that could be eliminated by reducing decisions.

Data Science führt zu besseren Entscheidungen. Das ist das Versprechen vieler Lehrbücher über Data Science. Andere Konzepte sehen Data Science als die Abkehr von Hypothesen, Modellen und Kausalität - statistische Korrelation ersetzt all das.
Entscheidungen stehen jedoch nicht erst am Ende des Data Science-Prozesses; sie erstrecken sich von den Ursprungsfragen über die Entscheidung, was als Daten zu betrachten ist, bis hin zu technischen Fragen der Datenmodellierung und -analyse. Diese Entscheidungen hängen vom Zweck, der Frage und dem Kontext ab.
Ich argumentiere, dass Daten, wie sie in Datenbanken und Datenmodellen verwendet werden, selbst als Modelle und Ergebnisse von Modellierungsprozessen betrachtet werden sollten. Sie sind erstellte und strukturierte Artefakte, die unter bestimmten Bedingungen erzeugt wurden und unter bestimmten Bedingungen präzise Antworten liefern können.
Daten schaffen ein eigenes Universum mit eigenen Regeln der Bedeutung und Gültigkeit.
Diese Arbeit richtet sich sowohl gegen die Vorstellung von Daten als Königsweg zu besseren

Entscheidungen als auch gegen die Vorstellung eines Bias, der Daten nachträglich verzerrt und durch eine Reduzierung von Entscheidungen beseitigt werden könnte.

Introduction

Data science promises better decisions. This is the central statement of many data science textbooks. The point where these decisions are made is open. For some, decisions are the direct consequences of correct algorithms and integral to data science processes; for others, decisions are made afterward, taking into account objectives and social and other influences. Common to all positions is the idea that (better) decisions come at the end of a structured process.

I argue against this that data science is a chain of numerous decisions that have to be made long before a decision can be made about the supposedly big question at hand. From the initial consideration of which problem to look at concretely and specifically using data science methods, to deciding which data to collect for it, determining measurement criteria and scales, decisions about normalizing, structuring, and storing data, determining metadata, and applying statistics and algorithms, a series of decisions permeate the data science process. Bias, a frequently discussed problem of data science projects, is thus not so much a subsequent falsification of actually neutral data, but rather an inherent feature of data projects that is necessary to create results.

I argue that, in order to comply with this, data, as used in databases and data models, should be considered as models and results of modeling processes.

Data play the roles of objects and instruments of knowledge generation in data science processes. They are created and structured artifacts that have been generated under specific conditions and can provide precise answers under specific conditions.

I consider this with examples from the open data context, where the promise of better decisions is further complemented by the promise of transparency and controllability - although open data is often the product of a long series of decisions about what all not to publish.

In order to understand and comprehend decisions (made at the end of the data science process), the purpose and context of the issues underlying the decisions must be present. Therefore, I argue against seeing data science as a method that guarantees presupposition-free research and more neutral decisions.

I consider data as an artifact that is created in data collection and preparation processes. Only then can data analysis follow. The question of how the results of analysis can be followed by concrete actions requires knowledge of the purpose and context of data preparation.

Some data purists regard this as a limitation. However, since data can only be given meaning and authority in concrete contexts and purpose relationships, I consider it an unacceptable restriction to dispense with these purpose relationships, which determine the decisions made in the data science process.

This thesis is divided into five major sections.
In the first section, I present the promise of data science and the problems associated with it. In doing so, I also discuss open data and further promises of transparency and control associated with this term.
In the second section, I discuss different data concepts. Besides some very reduced definitions that only make sense in concrete contexts, most concepts describe data as results of decision-making processes. The concrete form of data concepts in turn has implications for how data can be analyzed and interpreted.
With data assemblages and data journeys I present two concepts that emphasize the necessary contextualization of data.
In the third section, I apply the findings from the discussion of data concepts to two concrete use cases. The EU's lobbying transparency register is a case where the purpose and context of data collection have been lost in the data preparation process - to the detriment of meaningfulness and interoperability. The concept of promissory data describes a case where the purpose alone determines the consequences of the data - without the data (or what it was meant to describe) actually existing.
Based on the reflections on the concept of data and the two examples presented, the fourth section introduces a detailed concept of multilevel decisions in data science processes and examines their influence on the emergence of meaning and authority in data analysis processes. Contrary to what the promise of better decisions suggests, results of data science processes do not unfold by themselves.
The fifth section examines how the results of data analysis gain validity and impact, drawing parallels between data as models and other modeling concepts that also address the question of how (and whether) insights gained from models can be applied to facts to be described. I argue that statistics and data engineering are central aspects of data science processes, but that epistemic value of data science processes is better decided on the basis of pragmatic-constructivist concepts of technology.

1 The Problem

1.1 The Data Problem – Bias and Banality

1.1.1 Data science as decision driver

"Big Data" was originally a term that characterized a deficit: The term was used to describe data sets that were too poorly structured, too changeable, and also too large to be processed using conventional methods (Leonelli 2020). From this poor starting position, however, the term made considerable career and for the past ten years it was hard to imagine a discussion about planning, decision-making or even research and science without it.

The term „data science" experienced a similar flight of fancy: As a combination of mathematics, statistics, database know-how, and spreadsheets, the idea was not guaranteed the attention of the entire business world from the outset. Nevertheless, the job description was named the sexiest job of the century in the Harvard Business Review more than ten years ago[1].

What makes data and the professions that process it so extraordinary? There are great promises associated with data and its analysis. Data is supposed to provide new insights, it is supposed to deliver new knowledge that people with their limited analytical skills would not have been able to discover. Perhaps data will even go so far as to take decisions away from people and make human analysis processes obsolete.

This grand claim naturally raises a number of questions: What is it that we refer to as data in the first place? How can data become so active, how can it set such far-reaching processes in motion? What systems and organizations are necessary for data to play this relevant role? And last but not least: what view do we have of knowledge, engineering and technology so that we can ascribe these properties to data and data processing?

Introductory or foundational texts on data science make grand claims indeed: "The goal of data science is to improve decision making" (Kelleher and Tierney 2018), write John Kelleher and Brendan Tierney in their introduction to the topic. Data plays a crucial role in gaining insights and information, evaluating them, and taking correct actions based on them. Two approaches are considered in this context: Methods around linear regression deal with the past and analyze which factors contributed to the creation of a data set (for data scientists, this is a paraphrase for a set of facts). Algorithmic models and methods, on the other hand, deal with predictions about what will happen in the future based on existing and analyzed data sets (Kelleher and Tierney

[1] https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century, last accessed on 2023/06/15

2018, p.18). For Kelleher and Tierney, data are the results of abstractions of entities in the real world (Kelleher and Tierney 2018, p.39), so they are proxies that contain and make recognizable relevant features of that entity. This conjecture of abstraction and reduction, and their different ways of functioning, will concern us more often.

Data partly play the role of elements from reality that can be well processed in theories and models. Partly, however, they are themselves processed elements that have already undergone a transformation process before they became data. Therefore, the question whether data are elements of theory (or abstraction, idealization, or model) or of reality cannot be answered easily. Here, data discussions sometimes show parallels to modeling discussions, for example when the role of the model between theory and object is under discussion (Morgan 1999, Weisberg 2007) or when the question is raised how the knowledge gained on the basis of models should then be transferred back to the actual target system (Weisberg 2007).

In her Stanford Encyclopedia entry, Sabina Leonelli highlights the importance of data for knowledge production: she describes data science as a way "to revolutionise the production of knowledge" (Leonelli 2020). Data science "brings together computational, algorithmic, statistical and mathematical techniques toward extrapolating knowledge from big data" (Leonelli 2020). "Big data are widely viewed as ushering in a new way of performing research and challenging existing understandings of what counts as scientific knowledge" (Leonelli 2020). From a philosophy of science perspective, Leonelli looks primarily at the impact of big data on knowledge and knowledge generation. The connection between knowledge and decision-making is not explicitly addressed here; this is more evident in the more technically and pragmatically oriented position of Kelleher and Tierney.

Both ways of approaching data science focus on solving problems with new methods. In both cases, in the technical-pragmatic as well as in the philosophy of science perspective, the generation of knowledge is emphasized. Data create new knowledge (even if the path to it or the difference between data, information and knowledge are understood very differently) and in doing so data open up paths that would not be accessible to normal human analytical abilities.

In the early days of data science and big data, authors even proclaimed the end of theory in knowledge production. Theories and models are no longer necessary; knowledge is created by data (Anderson 2008). Critics of this approach point out that data science can certainly contribute to e.g. discovering new phenomena using its own methods and without an underlying theory from the respective discipline. This, however, is firstly usually based on theories from mathematics or statistics. Second, this is a potentially helpful extension, but not an argument against theories and methods (Haig 2014).

For the philosopher of science, the possibility to make diagnoses, to create categorizations and standardizations are primarily relevant. For technically and economically oriented data scientists, decisions in economic and organizational matters are often the prime examples in which data science must prove itself. Both agree that mathematics, logic, and clarity are the defining characteristics of data; insight and deliberation are outsourced to simple, automatable processes - or reshaped to be handled within such processes.

1.1.2 Data science critique

In popular discourses, data and data science oscillate between being the savior and being a threat. Data is ascribed great power, it promotes decisions, even automates them in some cases - and thus relieves people and optimizes their work, or it threatens people, deprives them of their work and reifies people, reduces them to a mere number.

The euphoria surrounding data science and big data, which has been widespread for many years, has long been accompanied by skepticism and criticism. The two main thrusts of the criticism are directed on the one hand against the supposed advantage of data and algorithms over human analysis, and on the other hand against the nimbus of neutrality and unbiasedness of data-based decisions. The first strand questions whether data and algorithms can actually make novel decisions, whether they do not rather merely reproduce (and in doing so certainly detail and supplement) what humans have given them. This line of discussion is relevant, for example, in debates about the validity of unsupervised machine learning.

The second strand of criticism deals with bias in seemingly neutral data-based decisions: Aren't supposedly evidence-based decisions weighted in a very specific direction long before the initial decision is made? Can data and algorithms even lead to neutral factual decisions, are they not rather shaped from the start by a set of human preferences, assumptions, biases, and even errors? Data scientist Cathy O'Neill cites a number of examples where Big Data has been shown to lead to completely wrong recommendations and decisions, even when all data science processes are technically and mathematically correct (O'Neill 2017).

The bias problem is frequently discussed at present. There is often a perception of pure, unadulterated technology (i.e., data and algorithms) being contaminated and falsified by human influence. Data scientists should take a close look at themselves and their approach to data in order to do justice to the powerful tools of Big Data. O'Neill suggests data scientists should take a kind of hippocratic oath that commits them to using data science methods in the service and for the good of society (O'Neill 2017, p 205). This idea illustrates that bias is understood as a flaw applied to data and algorithms from the outside. Data itself would not have this taint.

However, the idea also suggests that it would probably be very difficult to remove bias and arrive at unbiased data: The hippocratic oath for data scientists would tend to entrench bias, marking certain manifestations of bias as desirable and others as undesirable. In this way, then, different conceptions of justice from different societies could be projected into data and algorithms. This raises numerous questions about good scientific practice and ethical procedure, especially when analyses suggest results and conclusions that contradict the underlying concepts of justice.

In addition to the problems of posterior corruption (i.e., the idea that data are fundamentally unbiased and corrupted later) and unavoidability, a third question is raised here: At what stage of the data science process does bias actually come into play? This question is equivalent with the question when the decisions that data science aims to improve are actually made.

Kelleher and Tierney represent a fairly focused perspective on data science. The data science process is technically and mathematically driven. There are points of contact with domain knowledge, but the relevant decisions are made in the domain of data science. Data scientists that focus more on technical aspects emphasize the relevance of cleaning and normalization processes in order to achieve clear results (Irizarry 2019). Other approaches incorporate social and other frameworks in addition to domain science. In some cases, even the decision-making processes that are so important are located outside the data science process (O'Neill 2014). However, data science perspectives in computer science and statistics tend to regard this as a distortion that must be avoided.

The question of posterior corruption of data is largely synonymous with the question of when and how data take on meaning. When does data become information and when does it become knowledge, and where along the way does data become guidance for action or recommendations? This will require a closer look at the phases of the data science process. Otherwise, the mere diagnosis of bias will hardly get beyond the status of simple banalities.

Technical-mathematical oriented data scientists see less problems in bias. There are several reasons for this. On the one hand, data are regarded as unadulterated representatives of fragments of reality. This may seem strange from a philosophical point of view. However, a look at papers on data driven, model-free, and hypothesis-free research shows that, especially in technical disciplines, the promise of data to deliver unambiguity is upheld. Second, bias also serves a function in the data science process: Bias is necessary to make distinctions and decisions. "Without a leaning bias there can be no learning, and the algorithm will only be able to memorize the data" (Kelleher and Tierney 2018, p.144), write Kelleher and Tierney. The learning bias helps guide decisions and categorizations and structures and simplifies the world - much in the same way that prejudices do. The learning bias is the basis for abstraction and

simplification. It is questionable, however, according to which principles learning takes place here. Is consistency in the foreground, i.e. are conclusive and preferably repeatable categorizations to be made? Are comparisons in the foreground, i.e. are results checked for correct and incorrect, as is the case in supervised machine learning? Or are algorithms supposed to "learn" to make predictions about the future, which are then very difficult to verify? Kelleher and Tierney address a relevant limitation for algorithm-based prediction and learning: "Implicit in the machine learning process of data set construction, model training and model evaluation is the assumption that the future will be the same as the past" (Kelleher and Tierney 2018, p.150). This "stationary assumption" illustrates how difficult it is for statistical mathematical methods to respond to changing conditions. This is not so much due to the cumbersome nature of these methods, which on the contrary can be adapted in a technically flexible way. The problem is rather the question of which part of the calculation should be adapted, where in the data science process the change should be mapped.

1.1.3 The data science process

Kelleher and Tierney describe the data science process as a pyramid (Kelleher and Tierney 2018, p.57). At the lowest and broadest level are data sources. The data aggregating step builds on this, followed by data exploration. Only then does machine learning begin, with decision making at the top. The pyramid model not only depicts the quantity ratios, but also the effort involved in the individual process phases: "80 percent of project time is spent on gathering and preparing data," while algorithms and models are by far the smaller part, in which existing knowledge can also be better reused (Kelleher and Tierney 2018, p.67).
CRISP DM as Cross Industry Standard Process for Data Mining describes six stages: business understanding, data understanding, data preparation, modeling, evaluation and finally deployment (Kelleher and Tierney 2018, p.58). Deployment is the last stage during which insights are translated into decisions and transferred into processes. Business understanding as the first phase is a less technical step during which the actual problem needs to be discussed, understood, and sharpened before data can be used in the second phase to describe the problem domain as a proxy.
Long before data can be mapped into data models that can be analyzed or that serve as a base for calculations and thus lead to decisions, a series of decisions must therefore be made at a wide variety of levels. The necessary factual knowledge, on the basis of which relevant data can be identified, is already built on models, theories and hypotheses (even if some data scientists object; I will come back to this). Deciding on specific data, data types, and metrics anticipates a

great deal what can be analyzed and, more importantly, excludes much that can no longer be analyzed if it cannot be mapped in the data. The preparation and collection of data requires the specification of metrics, thresholds, scales, and their calibration. These determinations are decisions that influence the further process, and they are in turn based on decisions made in advance that sometimes have little to do with the object under study but are shaped by social values or political criteria. (Here this is only briefly touched upon; I will elaborate in later sections: Heather Douglas, in her discussion of Carl Hempel's concept of inductive risk, has shown how, along this path, unscientific and non-technical values can very directly play a role in science and technology (Douglas 2000)). In the model-building step, numerous decisions take place, evaluation reviews these decisions and thus decides whether the decisions made in advance meet expectations.

Data that are prepared in such a way that they can serve as meaningful indicators of certain facts can also already be seen as models themselves. They are the result of design processes, they are subject to factual constraints, and they can serve as instruments of cognition and are sometimes objects of cognition at the same time - this corresponds to essential criteria of some model concepts (Morgan and Morrison 1999, Knuuttilla 2021, Suárez 2003).

Giere (Giere 2010) describes data as part of the realities under investigation and as part of models; their classification is not clear-cut. Weisberg (Weisberg 2007) distinguishes between models as constructs of indirect representation and theories as ways of directly analyzing reality. Data occur in both domains. Weisberg puts up for discussion not only which decision guidelines can be used to create models, but also how decisions made on the basis of models or knowledge gained via models can be transferred back to reality.

So before we get to the top of the pyramid, where decisions actually take place - according to the literal sense of the pyramid description - a great many decisions have already been made. This raises questions, given the promise of data science to help make better decisions: Where, at what step, are these decisions being made? And what is actually being decided on?

Analyses, bias and external influences on neutral technology give rise to the image of an independent technology that stands alongside spheres such as nature, the environment and society, and which we can hardly approach without distorting it. – This is a perspective that disregards the necessity and usefulness of abstractions and idealizations.

An answer to the question of the place and nature of decisions requires a detailed view on data science processes. After some basic considerations about what qualifies as data, I will describe specific Data Science processes as examples. I will focus mainly on Open Data processes, in which multiple decision points emerge even better and more clearly.

1.2 Data Science and Open Data

1.2.1 Open data as a promise

Data is a promise. Dealing with data promises transparency, openness, neutrality, controllability, efficiency - data are a contemporary symbol of technical rationality. With these characteristics, big data and data science are relevant in various sciences, on the one hand to optimize processes and expand knowledge possibilities, and on the other hand, as in the case of Digital Humanities, for example, the preoccupation with data defines new research fields and promises questions and answers off the beaten track.

The promise of openness, transparency and neutrality is also relevant for politics. Evidence-based decisions are readily promised, processes are supposed to be transparent, and decisions and bases for decisions are supposed to be controllable. One way to deliver on this promise of transparency is through open data. Just as other data science processes promise better decisions, open data promises openness and controllability. open data should make it possible to examine the work of authorities or governments from diverse perspectives and to work out entirely new, unexpected questions. At the same time, the publication of open data sets is also a kind of proof of quality for the work of authorities and governments, which can thus demonstrate that they have an overview of the key data of their work at all times and give the impression that they need not shy away from any questions, controls or criticism, because all information is openly and comprehensibly available on the table. In addition to the promise of openness and transparency, this proof of the traceability of decisions is the second great promise of open data. Data science thus takes on the task of relieving decision-makers. Data science frees them from the suspicion of ideology or patronage-oriented decision-making - with open data, decision-makers can claim to be guided only by facts. Anyone who contradicts them or would decide differently is acting in contradiction to facts. Open data thus may function as a camouflage for political discourse.

1.2.2 Open data critique

The expectation that data dcience will lead to better decisions is not only driven by methodological or epistemological aspects, but to some extent also by normative aspects. It presupposes that the action suggested by the results of data science should prevail over other options. If the data “say” something, this is hard to oppose (I will describe more detailed

examples here in chapter 3.2 dealing with the concept of promissory data). In open data, this normative aspect is even more pronounced: Data is open, transparent, democratic, and good; missing data is a deficit and an indication of deficiency.
The scope of this claim raises questions and thus attracts critics. Criticism of the methodology is accompanied by criticism of the political dimension: The same methodological problems can arise with open data as in other data science projects, but the question of freedom of information is often added as an additional problem. If open data only involves fragments of data sets that have not yet been published, i.e., if authorities still decide on the publication of data, it is more likely to be controlled data than open data.
Rob Kitchin, for example, is particularly explicit in his criticism of the assumptions surrounding open data. He identifies three major thrusts of open data criticism: "open data facilitates the neoliberalization and marketization of public services" (Kitchin 2014. p. 61), because the expectation is often that citizens or NGOs will use open data to build information services or services that the state no longer has to provide; with open data, public agencies are thus abdicating their responsibility. A second objection is that many open data publications lack "sustainability, utility and usability“ (Kitchin 2014. p. 61). This objection is of a more technical nature and relates to the processability of open data records. These often do not comply with technical standards and thus hinder the enrichment and productive use of data. The third line of criticism: open data "promotes a politics of the benign and empowers the empowered.“ (Kitchin 2014. p. 61). Openness and transparency become a gesture of those who still decide what will be open and transparent. Open data sets that are disclosed still raise the question "whose interests are represented within the dataset and whose interests are excluded" (Kitchin 2014, p. 63). This also makes it clear that decisions in the provision of open data sets are made on a multitude of levels - usually, however, not only after the analysis of data, but already during the collection and provision of data, in fact already during the decision which data will be collected with which methods and according to which criteria. This is why Kitchin is even more explicit in his critique: open data "propagate injustices and reinforce dominant interests; open data can function as a tool of disciplinary power" (Kitchin 2014, p. 63).

1.2.3 Open data genealogy

From the royal road to transparency, openness and controllability to a powerful disciplinary tool - this is an astonishing career for a supposedly neutral technical term.

Jonathan Gray traces the genealogy of open data to the "cornerstone of official transparency and accountability initiatives around the world" (Gray 2014, p. 1). He sees economic rather than political motives in the success story of open data.
Gray analyzes the promise of open data and contrasts it with concrete data publications and policies. His diagnosis: „The rise of open data has coincided with a focus on technological innovation, public sector efficiency and economic growth in official transparency discourse rather than on social justice and meeting the needs of citizens“ (Gray 2014, p. 1).
Transparency, information and openness are referred to as high and public goods in open data debates, however they are rather traded as commercial goods. Transparency is demanded with political and social arguments, but then processed with commercial benefit and efficiency- if not profit-oriented perspective. "Open data advocacy and policy often place an emphasis in a mix of transparency, accountability civic participation, democracy and innovation" (Gray 2014, p. 8).
The formal arguments for Open Data strive images of freedom and independence, thus the concept finds broad support in diverse political circles. However, benefits from open data arise in economic spheres; social and political spheres remain largely untouched. The services created on the basis of open data often address commercial problems and allow some entrepreneurs to create new business models. For citizens, however, the relevant difference is not whether a service is provided on a public or private basis. For users, aspects such as reliability, availability, usefulness or affordability are more relevant. For service providers, on the other hand, the use of open data, together with the argument of providing services that the public sector does not provide, can be an additional marketing argument.
The commercial services offered benefit more from the open data nimbus than citizens benefit from the openness and transparency promises of open data. Online and digital pioneers certainly welcome this; Tim O'Reilly, for example, developed the scenario of "government as a platform," which he sees positively (Gray 2014, p. 10). Governments and public authorities thus only provide interfaces and data; citizens or commercial providers do the work. On the one hand, this means freedom and accessibility; on the other hand, it means the elimination of all democratic control and responsibility: those responsible have only made data and interfaces available; others have acted, but they are not accountable to anyone (at most to their owners). Public goods such as information thus become raw materials for those who can afford to process them and market the products.
This makes it clear how Kitchin's verdict on open data as an instrument of discipline can be interpreted. For the distinction as to who controls open data also extends over several stages, and information content and freedom decline in the process. Even before the decision is made about which data sets are made available in which form, a decision must be made about how

data is processed, which technology is used for this purpose, and in what granularity the processing takes place - which means that some information details are already filtered out and are no longer recognizable for future users. And even one step before that, decisions are made about which topics are mapped in data sets based on which criteria and which section of the topic area under consideration is viewed in what detail.

1.2.4 Decisions in Open Data

The collection, preparation, and delivery of open data sets also occur along a series of decisions. The decisions informed by analytics and data science that users of open data can then make based on that data depend on those preliminary decisions.

Radical transparency, Jonathan Gray diagnoses, is not at the core of open data. On the contrary, around open data, freedom, democracy, and commercial interests in public goods intersect. Gray sees radical transparency more on the borderline between transparency and criminality. For Gray, Edward Snowden or Chelsea Manning are advocates of radical transparency who publish without having coordinated with people in power and now face persecution and imprisonment (Gray 2014, p. 22). Aaron Swartz, an early advocate of open government data, warned of open data puppets as "reactionary political movements (...) cloaking themselves in nice words" (Gray 2014, p. 22),

For Gray, open data has become more of a battle cry, claiming openness and transparency while obfuscating them just as efficiently. Open data must be seen as a comprehensive process, not just a neutral provision that gives people all the freedom to interpret it. Open data frames a certain understanding of transparency and thus pursues a goal (Gray 2014, p. 23).

Even in Gray's account, the supposedly neutral technology that maps facts via open data is anything but neutral; the repeatedly diagnosed dependencies on a series of decisions is a genuine component of the concept.

Data are not neutral or privileged proxies for fragments of reality. The promise of being able to make decisions based on data alone ignores many fundamental questions about the very notion of data. I will explore some of these questions in the next section.

2 What counts as data?

2.1 Data in discourse

So what do we mean by data when this term is given great power, when it is at once threatening, enticing, neutral, and manipulative?

Why and how should data take decisions away from humans? By which qualities and for which reason?

Minimalist attempts at data definitions are not very useful in the field of tension around data science and open data. Pietsch lists some very reduced attempts at definitions for data. Computational definitions are primarily oriented towards the machine readability of data, according to which data would be "uninterpreted inscriptions" or "anything that is stored in a symbolic form on a medium" (Pietsch 2021, p. 6) This very minimalist definition leaves unresolved the question of how the symbolic component can function without interpretation. Representational definitions see data as "representations of purported facts." The restriction to "purported facts" is meant to circumnavigate the problem of how false claims or fictions can be represented.

Lyon, in his contribution on data in the "Oxford Handbook of Philosophy of Science," opts for a diaphoric definition of data that sees data primarily as difference (Lyon 2016). This property of making differences discernible is a very relevant property of data that recurs in very many concepts.

To broaden the perspective, it makes sense to look at how have data been used. Looking back, data entered the stage with the emergence of experimental science that wanted to provide results that were independent of opinions, worldviews, theories, and religion.

"In seventeenth-century philosophy and natural philosophy, just as in mathematics and theology, the term ‚data' functioned to identify that category of facts and principles that were, by agreement, beyond argument" (Rosenberg 2013, p. 28), writes Daniel Rosenberg with reference to the organization of experimentation in the early Royal Society. Data creates knowledge that can be directly witnessed and does not require interpretation. Data are not questioned; those who question data deny the simplest principles of reality. This is a powerful principle that gives data an unheard-of special status and inevitably raises the question of the basis for arriving at that special status. Accordingly, early opinions on this were already controversial, as traced by Shapin and Schaffer, for example: In their study, the empirically experimentally oriented Robert Boyle and the logically rationally oriented Thomas Hobbes stand in contrast to each

other. Boyle's experiments provide partly unexpected data, which as facts become the foundations of new theories. For hobbesean thought, these are contingent fragments that cannot be true if only because they contradict the postulates of the logical theoretical edifice (Shapin and Schaffer 1985).

In this debate, data not only had the function of providing new information (or also of safeguarding existing information), they also mediated a different form of debate. Where there was data to agree on, there was no longer a need to debate principles and their consequences. Data took on the function of proxies; members of the Royal Society could discuss data in a restrained and polite manner and did not have to put their whole personality, convictions and honor on the line.

Data thus fulfill two essential functions: They convey a special closeness to things; they are, in a sense, ambassadors of a reality that is anchored in them. And they fulfill a very relevant rhetorical function. It is "crucial to observe that the term 'data' serves a different rhetorical and conceptual function than do sister terms such as 'facts' and 'evidence.' To put it more precisely, in contrast to these other terms, the semantic function of data is specifically rhetorical" (Rosenberg 2013, p. 26), writes Rosenberg. Facts may be theory-laden, but where they are indisputably valid even without theory, “we call them data” (Rosenberg 2013, p. 25). Whoever has data on his side is thus clearly at an advantage in disputes - until the question is asked how data come to their special position or, above all, to a concrete meaning in the first place. Facts only function meaningfully within a theoretical or value structure, and even data do not speak for themselves alone. Rosenberg considers the rhetorical component of data to be one of its central properties: "facts are ontological, evidence is epistemological, data is rhetorical" (Rosenberg 2013, p. 26).

Rosenberg's analysis of the relationship between data and facts is set in a world of scientific discourse. The objective is to create new knowledge, to secure knowledge, to convince and to be right. However, the rhetorical, formative, and fact-creating power of data is also relevant in quite different contexts - for example, in questions of political power.

2.2 Data controls information

The historian and legal scholar Cornelia Vismann analyzes how technical developments in bureaucracy, archiving and documentation affected power structures (Vismann 2000). This is evidence of how data structures and decisions are interdependent and thus create realities. Vismann’s focus is on technical artefacts such as folders, filing systems, and binders; content generally plays a subordinate role in files. Files are a form of specifying and recording

information. Files are an early carrier material for data, and they change just as data change in their formats and appearances. Early files recorded information in descriptive prose; the structure and design of a file depended heavily on the executive official, who thus had control over the file and was an indispensable authority. Censuses, as the first major data collections, began to change this picture; the description of the state recorded in files took on more and more arithmetical forms instead of descriptive ones (Vismann 2000, p. 214). With structured tables, the creative power of the official dwindled; the recording of information became more of a filling in of forms than a powerful and reality-creating activity. The relevant decisions were made elsewhere, such as where to decide on the structure of the tables to be filled in. Vismann describes how this disempowerment led to freedom for officials. Precisely because processes were clearly specified and little deviation was possible, checking every step was no longer necessary. It was no longer necessary to approve every decision about designations or classifications individually - the decision had already been made. Within these guidelines, individuals could act independently. Technical aspects of data entry and data management thus had an unplanned but all the more massive influence on legal and organizational processes. Content structuring such as tables, lists and classification systems contributed a part, material technical innovations such as binding systems or later the file folder were external influences with which information and data storage changed the organization of the administration (Vismann 2000, pp. 276). With flexible filing systems, data can be restructured at any time, they can be compiled and retrieved differently - and they thus also change their meaning. Registries and file folders, which emerged in the late 19th century, set the first steps towards the mobility of information. Written data is no longer stored in fixed or stapled form; parts of files can move from one file to the next, from one office to another. With the increasing formalization and mobilization of information, the problem areas of information security and data protection become relevant. Flexibly held data can move to other environments and mean different things there; information in data is fluid (Ess 2009). It can also become intelligible to recipients for whom it was not intended. Thanks in part to these properties, data become an important instrument of power. They are information, as such they define the scope for action, they allow images to emerge - and also in this very analogous way, more than a hundred years before the emergence of data science methods, data are an important instrument in the attempt to make better decisions. In this case, too, data has already gone a long way along the path of many decisions before the decisions actually intended are made – and the circumstances and techniques of collecting, manipulating and storing data have proven to be highly relevant.

2.3 Representational and relational data

Sabina Leonelli, in her reflections on model organisms, deals with the question of how, in biology, some organisms can serve as models and how models come to have their representative character. Organisms have model character when they express certain properties particularly well, which are considered relevant. There is a performative aspect to this as well, similar to the use of data as rhetoric. Models are models because they illustrate properties that are particularly relevant; these properties, in turn, are also particularly relevant because they are particularly well and clearly represented by models that are considered as relevant representatives of their species. "What defines model organisms as a specific subclass of experimental organisms is the representational power attributed to them" (Leonelli and Ankeny 2020, p. 9). Models have special representational power because this is attributed to them. Models are thus always dependent on concrete context in order to be able to unfold their meaning. Models that fulfill the role of a prototypical representative, representing something as it is, nevertheless require a number of conventions in order to fulfill this role.
The rhetorical component here also takes on normative features: something is considered a model, and thus this organism also sets specifications for others, which they have to comply with in order to be considered what they actually already are. Along these guidelines, decisions can be made about deviations, fulfilled expectations or changes in the characteristics of a model or an organism. Londa Schiebinger has reconstructed such processes using the example of the depiction of ideal-typical male and female skeletons in the representations of sixteenth- and seventeenth-century anatomists: In order to be able to draw particularly meaningful, ideal-typical skeletons that were supposed to be representative of male or female anatomy, some researchers gathered bones from different skeletons. The mere depiction of the average, on the other hand, was considered by many to be a falsifying representation guided by the wrong principles (Schiebinger 1989).

What do models and prototypes have to do with data? Both are attributed a form of special representative power, they depict facts or emphasize particularly relevant criteria especially clearly. And in both cases, they arrive at this significance by rather circuitous routes that are somewhat at odds with the assumption of this special status.
Leonelli and Ankeny describe the attribution of representational power. Suárez opposes even more clearly the idea that models become models by virtue of internal properties: He argues against similarity and isomorphism as essential criteria of the model property (Suárez 2003). Suárez sees representation as the central function of models; however, representation does not

result from similar structures or other internal features of a construct that is considered a model, but from the directed and intentional use of the construct as a model. Thus, the purpose of working with models is not to recognize similarities, but to initiate surrogative reasoning (Suárez 2003, p.229). Models must be worked with.

This autonomy of models, however, raises the question addressed by Weisberg of how and on the basis of which criteria the work with models can then be transferred back to their target systems (Weisberg 2007).

The approach of considering models both as objects and as instruments (Morgan and Morrisson 1999) allows to look for the answer to this question in the work with models themselves.

Other perspectives on models wish to remove the problem of representation from the discussion altogether, and thus focus on the creation of and work with models as those activities from which meaning emerges (Knuuttila 2021).

I will return to these aspects later, when the emergence of meaning from data is a specific topic. For now, models are not simplified representations of something that counts as a model because of internal structures; model property arises primarily from engagement with something as a model for something and for a specific purpose.

Data are not straightforward simplified representations of a reality any more than models are. They do not point directly to meaning; engagement with data cannot readily be seen as a projector or magnifying glass through which more complex facts can be deciphered. Rather, the path to meaning leads through varied relationships and choices.

In her reflections on data science and big data, Leonelli expresses a very similar idea. Leonelli distinguishes between the "representational and the relational view" on data (Leonelli 2020). These are two fundamentally different ways of looking at data and its key properties. Depending on the concept applied, data fulfill their task differently. Leonelli characterizes the "key epistemic role of data as empirical evidence for knowledge claims or interventions" (Leonelli 2020). How data perform this role depends on what kind of relationships data rely on to achieve meaning and authority. "Within the representational view, the identification of what counts as data is prior to the study of what those data can be evidence for; in other words: data are 'givens'" (Leonelli 2020). Data here stand in space as something immediately present and refer to something. This relationship already exists and is independent of who uses data and in what context. Data exists, it is independent of whether it is considered data or not.

The relational view is in clear contrast to this. "Within the relational view an object can only be identified as datum when it is viewed as having value as evidence" (Leonelli 2020). Thus, data here only becomes datum when it is viewed as such. The decision to consider something as

datum makes it so; it is not a special property of an object to be a datum. Anything can be a datum - but the datum relation applies to a specific context and purpose. The decision to consider something as a datum gives rise to meaning. Data viewing and data processing are "integral part of inferential processes" (Leonelli 2020).
The relational view of data is reminiscent of pragmatist and deflationist perspectives on modeling. No substrate can be assumed that particularly characterizes data (or models). Usefulness arises in the engagement with data (or models) and thus also depends on specific questions and environmental conditions.
The relational view is criticized because this data perspective places an emphasis on the portability and mutability of data - data must be relatable in order to fulfill its purpose of communicating something about the world. Pietsch (Pietsch 2021) criticizes that this mobility is an unnecessary addition that adds nothing to the data property. One can counter that wholly immobile data can neither acquire nor convey meaning - they do not enter into any relationship.
The need for relations focuses attention on the practice of data handling and processing. Dealing with data is an active, creative process that gives data its data status in the first place. Along the way, many decisions need to happen until there is a clear picture of what is to be considered as data. This perspective knocks data off its throne of immediacy. Data are per se subject to influences. Not only are there disruptive influences by which data are distorted, data emerge in a process and this genesis between data collection and the purpose of data interpretation is part of the property of being data. Therefore, data practice, that is, the handling of data, must also be part of the analysis of data (Leonelli 2020).

2.4 Data within their relations

Data are the result of decisions. This criticism of the supposed immediacy of data is very evident in Rob Kitchin's work. For him, data are not given but taken, consequently they should not be called data but capta (Kitchin 2014 p.2). Kitchin goes beyond considering only the data collection and interpretation processes as shaping influences on data, for him data are also dependent on "thought systems" that guide their extraction and consideration (Kitchin 2014, p. 20). Data, then, are not only subject to influences at their core; the same element may be datum in one perspective and not in another.
Kitchin elaborates this property of data as the basis of his data theory, which does not consider data in isolation, but in its history, environment, and technical preparation as so called data assemblages (Kitchin 2014). Kitchin also introduces the designed character of data as a

dimension of a data typology. This typology recognizes the common distinctions of formal aspects (qualitative and quantitative data), type (indexical, metadata) and producer (primary nd secondary data). These three criteria are supplemented by a further aspect: The distinction according to source separates into captured, derived, exhaust or transient data. Captured data corresponds most closely to the traditional image of data, but "captured" emphasizes the active part of collecting and shaping data. Derived data are calculated or extrapolated; here the creative intervening aspect is already clearer. Exhaust data is the result of data mining or other analytical processes in which other types of data are actively worked with, so that new data sets are created or data are grouped, interpreted and evaluated differently. Transient data is data that is often not processed because it is too large, too unstructured or too time-consuming to process. New processing methods and technologies, however, are constantly giving rise to solutions for processing such data sets (Kitchin 2014, p. 4).

Kitchin notes that both captured and exhaust data are often considered original, unprocessed data sets. Transient data is often not considered data at all. Only derived data is perceived as processed and subject to human influence. "There is much more to conceptualizing data than science and business generally acknowledge" (Kitchin 2014, p. 21). Part of this editing process is visible where processing is subject to strict rules. Technical data processing, for example, and its normalization processes that define rules for how data should be prepared, decomposed, and stored, must be documented as rules and simultaneously enacted as practice. They establish a way of dealing with data and a kind of data, and at the same time must presuppose that this way of dealing with data and this kind of data correspond to what data are by their very nature. "Normalizing processes have to be constantly and recursively reaffirmed through implementation" (Kitchin 2014, p. 23).

Rob Kitchin describes the idea of data assemblages. Data are not spontaneously found independent representatives of an untouched unadulterated truth, they are the result of diverse processes in which diverse influences have their effect. How can this be taken into account? Data assemblages are data infrastructures in which not only individual data elements are considered relevant aspects, but in which connections, potential links, are elementary. Data assemblages are sociotechnical systems in which the origin, purpose, processing, processability, or use and usefulness of data always play a role (Kitchin 2014, p. 47). Data by themselves are hardly tangible; assemblages provide context and contrast - and it is within their framework that data become useful.

The concept of data assemblages changes the way we look at data. It would be a mistake to view data as content and infrastructure as a vessel. Data cannot be removed from one infrastructure and transferred to the next without changing. Data and infrastructures are inseparable parts of a

whole meant to convey information and meaning; data depend on their infrastructures as a context for meaning. As a contrast to data assemblages, Kitchin criticizes mere data holdings or data dumps, which are much more likely to assume the role of mere vessels, and in which data are mere fragments that have not been cleansed and validated and lack interoperability (Kitchin 2014, p. 64). For Kitchin, data that are only accessible in dumps or holdings do not provide entry points for meaningful analysis. These deficiencies in the supply of data are considered responsible for the fact that interest in data, especially in open data, often flares up only briefly, but quickly falls flat again due to the lack of usefulness and applicability of the results (Kitchin 2014, p. 65). In the context of open data, Kitchin occasionally sees these deficits in data provision as part of a strategy: those who provide data can claim transparency and openness for themselves; if data cannot be productively analyzed, however, no new information will emerge. The original owner of the data will retain the sovereignty of interpretation and the knowledge advantage; this kind of data is not useful for the public.

2.5 Mutability and portability as central characteristic of data

Data cannot be considered as fixed, context-independent entities. This aspect of data has come to light several times and has been argued from different perspectives with different emphases. A particularly clear practice-oriented expression of this idea can be found in the concept of data journeys (Leonelli and Tempini 2020). In the course of observation, data transforms from fixed, unquestionable and unambiguous realities to flexible dynamic structures that depend on many factors and change with them. How does this change come about?

A "view of data as a fixed, context-independent body of evidence, ready to be deployed within models and explanations, also accompanies contemporary discourse on big data", criticizes Sabine Leonelli, and immediately follows: "what constitutes data in the first place is not questioned, nor is the capacity of data to generate insight" (Leonelli and Tempini 2020, p.5). Given the supposed clarity attributed to data, there is surprising disagreement about what data actually is and how it fulfills its role of generating meaning. While philosophy and philosophy of science attach great weight to the question of how meaning is created, this aspect is often disregarded in internal discourses on data science processes. Meaning is the responsibility of the domain discipline; statistics, mathematics, and data science are not responsible for it. This opens up undesirable room for maneuver: If no stringent consequences can be derived from data at the domain level, i.e., if data does not directly tell us what to do, how can it be data that fulfills the promise of data science to help us make better decisions? How does the perspective

have to change in order to avoid a missing link, to prevent the return of arbitrariness in the interpretation of data?

Leonelli criticizes insufficiently reflective data science positions on two levels: First, there is a need to analyze the decisions that researchers use to identify and use some things as data. This is a strongly practice-oriented reference. Second, a "critical framework" needs to be developed for the analysis to examine "how data come to serve as evidence and the conditions under which this does or does not work" (Leonelli 2020a p.5).

In contrast to static, self-contained and closed data, Leonelli identifies "two key factors" for data that can also be used in the context of open data: mobility and interoperability (Leonelli 2020a, p. 19). Mobility opens up new contexts, interoperability establishes relationships with other datasets, allows linkages to new environments, and helps data function in networks and dependencies where meaning can emerge. "Data are mobile entities, and their mobility defines their epistemic role," writes Leonelli (Leonelli 2020a, p. 23). Data, then, can only create cognition and meaning as mutable and transportable elements. For Leonelli, this property even becomes the central criterion for the identification of data: "for any object to be identified and recognized as datum, it needs to be portable" (Leonelli 2020a, p. 23).

If mobility and linkability are relevant characteristics of data - what requirements do these characteristics place on the way data is viewed? Interoperability is the less surprising of these two properties. Databases are made up of links, new records must be easily added, normalization and standardization are well-established processes that seek uniqueness, identifiability, and clarity. In a technology-dominated perspective, these processes serve to remove data from the messy, unclear world of its environment and move it to safe places where it can rest fixed and unaltered. Leonelli criticizes this notion of immutable data and instead of immutability postulates precisely the mutability of data as a relevant factor in their property of allowing cognition. Strict normalization processes are also criticized elsewhere in discussions of digital humanities: Where, for example, data describe things that go back a long way in history, or deal with phenomena that are no longer directly accessible or comprehensible to us today, no strict normalization can be carried out. Normalization processes distort data, introduce inappropriate criteria, and lead to unacceptable truncations (Flanders and Fotis 2016).

Leonelli relies on "relational and historicized understanding of data as lineages", which has to be considered historically in a twofold way (Leonelli 2020a, p. 18). Data are created in processes and manipulated in processes. Data are used in different projects for different purposes, so they can be used over longer periods of time and adapted in the process. Different methods and techniques, external factors such as legal or social norms, and tools and instruments change perspectives on data (Leonelli 2020a, p. 21).

2.6 Focus on data practice

The data science process itself, from data collection to decision making, is a dynamic journey in which different standards are applied to data. While data dcience likes to claim to provide direct and unbiased insights, the shift in perspective from static data to data journeys shows how much this claim of data is linked to intensive mediation work: "using data as evidence is everything but straightforward, and sophisticated methods, resources and skills are required to guarantee the reliability of the empirical grounds on which knowledge is built" (Leonelli 2020a, p. 26). Numerous steps of data planning, collection, and processing follow concrete plans and strategies, theoretical frameworks (be they technical or directed to the specific domain) ensure consistency and traceability. Working with data, viewed as a dynamic historical process, thus always follows theoretical guidelines; data does not order itself without history. In this respect, Leonelli explicitly emphasizes, the "idea of data journeys (...) is a direct counterpoint to the distinction between 'hypothesis-driven' and 'data-driven' models of research" (Leonelli 2020a, p. 26). Data are produced; theories, expectations, tools, practical circumstances, recording methods, database systems, documentation, expertise, and technical know-how are just some of the factors to which data are exposed during their production and which must be kept in mind if data are to fulfill their function as carriers of evidence and meaning.

Material properties of data, data-holding systems, and data-recording methods also play a role; thus, data directly interact with material properties of their objects or their environment: data can change their environment or their objects and give rise to new behaviors.

Leonelli's way of detailing this view of data leads to engagement with case studies, data practice, and practitioner observation. The anthology "Data Journeys in the Sciences" (Leonelli and Tempini 2020) describes, among other things, the work of marine biologists grappling with how different types of fishing nets affect the plankton samples collected with those nets and the data read from them, of physicians using anecdotal medical histories as data sources in their data sourcing, or of genome researchers having to deal with the influences of computing infrastructure on their results. The latter is not only a topic for scientists in the field or philosophers of science; the influences of infrastructures on computational results and thus on data are also increasingly discussed in depth in computer science. Long runtimes of computational processes, overloaded storage capacities or data tables that become confusing are some problem points that lead to results being distorted by premature aggregation or by gaps in loading the basic data. One research problem in such cases is whether these changes in data structures are due to desired learning processes in the computations (i.e., learning bias, which

Kelleher and Tierney, among others, emphasize), or to technical or conceptual flaws in the data science process. In highly computationally intensive processes such as genome sequencing, for example, the runtime of cloud infrastructures has been found to negatively impact their performance and thus the precision of the results; trivial interventions such as regular restarts could remedy the situation (Hao Tong et al. 2021). In the analysis of open data projects, data history and metadata describing that history are relevant factors in the evaluation of data. Data sets often provide clear and concrete results in the direct context of their initial collection. However, when the same data are examined in only slightly different contexts or in the context of questions with divergent research interests, they quickly lose their validity. Examples include undocumented questions and response options in data collected via surveys, or missing information about whether missing data are due to refused information, inapplicable questions, or other reasons. A renewed focus on metadata is intended to provide clarity - but it is worth questioning whether this clarity can be achieved at the data level or whether other criteria need to be included (Blank 2019).

2.7 Data as processes

Lisa Gitelman and Virginia Jackson address another dilemma in their introduction to "Raw Data is an Oxymoron": "At first glance data are apparently before the fact: they are the starting point for what we know, who we are, and how we communicate" (Gitelman and Jackson 2013, p.10). In light of the previous considerations around data, it may seem surprising to position such a naturalistic view of data as an adversary. However, in data science practice or management consulting, such monolithic data concepts are common. Other science and technology researchers also diagnose "naturalistic" data concepts (Longino 2020, p. 398). In computer science studies, more critical data concepts are fighting for space in courses on technology and society.

In technology- or business oriented perspectives, data claim a special position; they are postulated as a fixed point. Thus, they are waymarks in an otherwise indistinguishable set of circumstances that would not allow orientation without these waymarks. "Like events imagined and enunciated against the continuity of time, data are imagined and enunciated against the seamlessness of phenomena" (Gitelman and Jackson 2013, p.11). This is a process applied to phenomena, which on the one hand structures and thereby creates clarity, but on the other hand, through these interventions, also removes ambiguities, thus producing the supposedly represented phenomena in the first place: "When phenomena are variously reduced to data, they are divided and classified, they are processes that work to obscure – or as if to obscure

– ambiguity, conflict, and contradiction" (Gitelman and Jackson 2013, p.17). Gitelman and Jackson declare the supposedly objective data coming from the thing itself to be the expression of pure individuality and perspective: "Thus the history of objectivity turns out to be inescapably the history of subjectivity, of the self, and something of the same thing must hold for the concept of data" (Gitelman and Jackson 2013, p.14). A techno-mathematical perspective on data would see this as the ultimate corruption of the idea of data - but can such a judgment be argued with an almost animistic notion of technology that considers technology and data unperturbed by humans and society?

Daniel Rosenberg reverses the perspective altogether: "It may be that the data we collect and transmit has no relation to truth or reality whatsoever beyond the reality that data helps us to construct" (Rosenberg 2013, p. 45). Does the data discussion, then, contrast a notion of technology as neutral tool detached from people and society and a notion of data as technical constructs that represent nothing but the preferences and prejudices of people and society?

The legal scholar Viktor Mayer-Schönberger is one of those theorists who see very great power in data - and who view data as an independent component, at least not one that is socially determined. Mayer-Schönberger views data like commodities and even suggests taxing corporations according to data ownership or allowing them to pay tax debts in the form of data releases. (Mayer-Schönberger and Ramge 2017)

In addition to a number of potential privacy issues, this also raises the question of whether data can be traded on exchanges and retain its value for different owners in a reasonably trackable way, just like other commodities or money. Whether this value can then also be stored or transferred to third parties is another open question.

Most of the other data concepts mentioned here would doubt that. Data exist only in and from context, they lose their meaningfulness in isolation, and data alone are not a quantity that could be considered on their own. Kitchin and Leonelli seldom consider data alone; data must be considered in the context of its creation, application, and processing in order to be analyzed meaningfully. Both authors develop their own concepts for data analysis, which emphasize the importance of context, history and purpose. Vermaas and Kroes highlight the decisive elements in data and technology, with Dewey's and Hickman's pragmatist view von data we are reminded of the relevance of purpose and context (Vermaas and Kroes et al 2011, Hickman 1990). I will get back to these concepts later.

None of these views supports a substantialized concept of data. Instead, creating data, working with data, analyzing and interpreting data are processes along many stages with many decisions. I want to highlight the practical details of these processes in the next sections.

3 Cases – Data Problems in Practice

3.1 The Lobbying Transparency Register of the European Union

The company should be deleted from the register immediately, no connection should remain traceable. - It was the sister of Eva Kaili, the former Vice-President of the European Parliament, who approached the EU lobbying register with this urgent request. Her company, ELONTech, had had to cease its activities and should be removed from the register. That was in mid-December 2022, around the time of the blowing of the corruption scandal surrounding Eva Kaili, who had taken bribes from lobbyists in Qatar and Morocco, Politico magazine reports[2]. The entry has been removed. However, traces can still be found on websites of databases operated by NGOs.

The European Union lobbying register was introduced in 2011 after long discussions as a milestone of transparency. Lobbyists and lobbying organizations were to be recorded so that it would be clear for exposed politicians of the European Union and for journalists and the interested public, who lobbies on behalf of others and who represents their own interests. The directory grew quickly and now includes over 13,000 entries of registered lobby organizations. But how useful is a database in terms of openness, transparency and controllability, i.e. the great promises of open data, if traces in it can be covered so easily?

Discussions about deleting data are a governance issue that can be resolved fairly easily. However, the lobbying register database and open data publication has been and continues to be criticized for many other problems[3].

3.1.1 The collecting process

One group of criticisms concerns the rules of data collection. Registration is only mandatory for a few organizations. The wording of the rules leaves much room for maneuver. The range of

[2] https://www.politico.eu/article/eva-kaili-sister-mantalena-europe-lobby-register-transparency-qatargate/, last accessed 2023/06/15

[3] The current version of the register can be accessed online here: https://ec.europa.eu/transparencyregister/public/homePage.do, last accessed 2023/06/15

organizations represented in the register is correspondingly wide. Classic interest groups such as trade unions, consumer protection associations, industry associations, chambers of commerce and industry can be found there, as well as large companies, universities or even small businesses that have sought contact with members of parliament for their concerns. In addition to these organizations, which primarily represent their own interests, there are numerous agencies and other professional lobbyists who represent the interests of their clients. A rough categorization of the registered organizations should help to maintain an overview, but often raises questions in detail.

The information in the register is entered by the organizations themselves using an online questionnaire. Master data such as addresses and branch offices are requested, as well as information on the number of employees and the turnover achieved, on the issues lobbied for, on objectives, on contacts with EU institutions. Details of clients, any EU funding received and other information complete the data sets.

In theory, this results in a great wealth of data, which makes much more transparency possible. In practice, though, numerous problems become apparent.

This relates to the next set of criticisms that are often leveled against the transparency register. For the user of this database, the process of data collection is opaque. It remains unclear which of these disclosures are mandatory, which are optional, and which may be tied to certain thresholds (e.g. turnover or employee numbers).

For these two key figures in particular, it is also unclear whether the total revenues (and employee numbers) or only the revenues (or employees) generated with (or employed for) lobbying are to be reported, or whether the figures should perhaps even be broken down specifically to revenues and employees with reference to the institutions of the European Union. This leads to strange constellations: Among the largest lobbyists are often medium-sized industrial companies (which indicate their entire corporate turnover) or universities (which indicate all employees engaged in public relations or public affairs); lobbying agencies, on the other hand, which interpret the criteria for data collection differently, make themselves small.

Key topics are drawn from a predefined list of categories, while information on clients and interests is collected in an unstructured way. This again creates several problems: customers are sometimes not recorded at all, sometimes with different designations (for example, Google or Alphabet, Facebook or Meta), and the input (and later output) takes place in a single unstructured text field. There are also apparently no specifications for the input for interests and content focus. Prose is often found in this field, sometimes even in different languages.

3.1.2 Messiness in structured publications

These content-related and process-related weaknesses lead to numerous weaknesses of the register as an open data publication. For some years, the data was stored as csv and xml files in EU open data repositories; since 2021, the data can also be retrieved via a more easily accessible web application, which, however, cannot be used to create more complex queries - which is mainly due to the weaknesses of the underlying data.

The lack of information on mandatory and optional data, for example, raises the question of how to deal with null values. When should they be interpreted as "not applicable," "0," or "no information"? The unstructured free text fields such as interests and focal points are practically impossible to analyze, and structural problems are exacerbated by the use of different languages. Different names for customers and their unstructured recording (all customers are shown in one text field) make it practically impossible to formulate simple queries such as the most frequent customer or to relate customers to topics and countries. Moreover, in the free text fields where multiple entries can be made, different separators are used - making it difficult to apply text mining methods to structure the raw data.

The mixture of structured information retrieved via categories, metrics that give the impression of precision, and unstructured data richness that gives the impression of being able to find a great deal of information is a very common manifestation of open data publications. The raw data of the transparency register is a prime example of those poorly structured data repositories that Kitchin distinguishes as "dumps" or "holdings" as opposed to data assemblages (Kitchin 2014, p. 64). The deficiencies reduce the usefulness of the data and limit their interoperability. Data can be poorly processed and they can be related to other datasets only to a limited extent. Before these problems become apparent, however, there are still some issues to overcome with handling the data itself. How can a data processing or data science process look like in a specific case, how can data help to make better decisions? Such a decision could be: We need to keep an eye on organization X if we want to understand how decisions are made in topic area Y. Or: In issue area Z, organizations from country X lobby most frequently, so we need to keep a close eye on awards made to companies from country Z. These are examples of simple mathematical tasks. Of course, some domain knowledge is necessary to formulate such questions and to model possible solutions. Nevertheless, the decision criteria seem simple at first glance: Whoever frequently lobbies on a certain topic is relevant for decisions.

The first step of the data science process, obtaining data, is simple in this case. Data has been collected and published, and the semi-structured formats enable initial quick orientation

attempts. We can quickly see how many organizations are registered, in which countries particularly many organizations are active, and which topics are particularly common.
In order to move from an initial orientation to valid statements, specific normalization, cleaning and structuring steps should now follow - and this is where technical and factual problems now mix. At first glance, metrics can be easily processed; sums, averages or sequences are formed according to the rules of mathematics. Without the information about what the metrics refer to (Is it the total revenue? Only that generated by lobbying? Can sales figures be related to the most common issues?), these arithmetic rules, seemingly independent of external influences, cannot be applied. They do not make sense.
Similar questions arise when analyzing lobbying issues that result from selection from a given list of issues. If several topics are mentioned, are they all to be considered equally important? Are there indicators for priorities? Or should the respective combination be included in the analysis instead of individual topics?
Free text fields and those data fields that allow multiple entries without form specifications pose a similar problem: Should they be considered as a unit? If so, each individual field is likely to be unique, and there will be little overlap. Should we try to split the fields into multiple fields using separators (such as commas)? This doesn't always provide meaningful information and again raises questions of possible prioritizations. Or is this a use case for artificial intelligence to make structuring and prioritization suggestions? That only shifts the question.

3.2.3 Can Metadata fix this?

Depending on how these few questions are answered, very different results are obtained. The data can be processed with clear, mathematically comprehensible and well arguable processes and will lead to completely different results, depending on which detailed decisions were made one step before. None of this needs to be wrong. Technically and mathematically, all paths are verifiable.
And yet, the decisive criterion by which an ultimate judgment could be made is missing.
For example, if one calculates which of the largest (by headcount) lobbyists advocate for which issues, then climate issues are the most relevant issue in the world - because they all have climate issues on their agenda, even if banking and agriculture are the larger industries (which almost all lobbyists also have on their agenda). If one calculates the most relevant issues of a country based on the most relevant issues of the biggest lobbyists, then the result always

depends on the applied prioritization of the issues in case of multiple responses. When customers are taken into account, an additional complexity dimension comes into play[4].
In none of these questions has a substantive debate oriented to factual issues been started; so far, it has been technical questions of data handling that have raised problems. Substantive questions about appropriate evaluation criteria can hardly be answered on the basis of the available data; here, political or other value-oriented criteria would have to be used as a substitute. Finally, in order to treat the data in a meaningful way, it would be necessary to know their context of origin, i.e., to know what mandatory fields were, how individual questions were formulated in concrete terms, and what metrics refer to exactly.
Ultimately, then, the lobbying register is a publication that tempts one to expect a great deal of information and a high degree of transparency. It also suggests, on the basis of the abstraction of relationships in data, that data science methods can be applied here. However, data science methods fail in the attempt to provide significantly more information than could be generated by simple counting. The poorly prepared data refuse to be analyzed automatically in more depth. Individual data sets can be looked at, but as soon as data is aggregated to achieve higher information density in order to identify trends (i.e., clearer differences)), the fuzziness also increases. With increasing uncertainty, delimitations become blurred, and ultimately the information content decreases. To actually know something concretely, we must return to the individual data set. The expectation of using data and data science methods to get more information out of an open data set here has not been fulfilled.
The data sets provide little new information, and this new information cannot be applied to the subjects under study. The results are abstract and mathematical in nature (such as: "The most frequently mentioned lobbying topic is X"), but due to the unclear data structure, this result can hardly claim practical significance in relation to lobbying organizations or clients. Data science methods can be applied only with many caveats that always put the results obtained with them into perspective.
These shortcomings of the lobbying register need not be understood at all as political agitation or demonstrations of power using data in Kitchin's sense (Kitchin 2014, p. 63). Rather, the idea of being able to draw more and better information from such open data inventories is based on a problematic understanding of technology.

[4] I have published details and further examples of this in the context of another research on dataanalyst.at. Similar criticisms are also voiced by organizations such as Transparency International EU or lobbyfacts.eu.

Approaches to solutions that are frequently discussed at open data conferences and among open data managers generally involve metadata in particular. Metadata fulfills a kind of commentary function that supplements the actual data content. In the case of image databases, for example, metadata is information about the photographer, the location and time of the photograph, or the technology used. In the case of the transparency register, metadata could be used to identify mandatory and optional information, to map prioritization of topics, for example, or to classify metrics on employee and revenue figures. Metadata, which is always included in the analysis but plays no role in the calculation, could even contain the specific question from the questionnaire, the answer to which is the current data field, so that relevant context can be quickly established if necessary.

The more ballast in the form of metadata or other annotations data carry with them, the more they lose those characteristics that distinguish data as the preferred object of research and computation. Data are then no longer more abstract and easier to reuse. They lose their meaningfulness if annotations are needed to interpret them. Enriching data with metadata and additional information improves the comprehensibility and information content of data, but it makes them more difficult to handle. We are approaching a scenario as criticized by O'Neill in the vision of Mayer-Schönberger and Cukier: Everything is in the data, data do not map the world, they also have no relationship to the world, they are the world (Mayer-Schönberger and Cukier, 2013). But then we have not actually gained anything, but on the contrary we have moved back. We face similar problems here as do adherents of representational theories of modeling, who demand to see in models as accurate and comprehensive a representation of their objects as possible (Weisberg 2007). We end up back at the starting point.

What ways out are possible?

3.2.4 Not all Open Data Publications are useful

The attempt to transform data into information and knowledge encounters two major problem areas in this example of the application of open data: One is technical and concerns data quality and interoperability. Here we are dealing with data formats, separators, unstructured data fields and data gaps. The second area is political in the case of open data in the broadest sense. In the case of scientific data science projects, it would be issues of demarcation and legitimacy, and issues of power quickly come into play.

In the context of the supposedly neutral analysis process that is supposed to lead to a decision, a number of assumptions and decisions have to be made on different levels. These must either be explained and argued or enforced by other means. This is necessary so that the result

obtained can claim validity and is not just one interpretation among many others. Simple mathematics that follows clear rules thus becomes a complex framework that must take social and political dimensions into account and needs numerous extensions on the technical level in order to be able to defend the results of analyses.
From this perspective, Kitchin's criticism of open data as a disciplinary tool becomes clearly understandable (Kitchin 2014, p. 63). But this tool can also be applied from the bottom up by pointing out gaps and weaknesses in open data publications. Deficiencies in open data can be addressed as part of an extended analysis.
Metadata is often addressed in the course of self-reflection and self-criticism by open data providers as a means to address ambiguities or inconsistencies in datasets (Blank 2019). Metadata can be part of datasets, but as additional information they are usually not part of the analysis process. They create constraints and conditions that must be carried along like a carryover in calculations and ideally can be resolved at the end. Moreover, different standards for metadata or different interpretations of diverse standards create new problems in the interoperability of metadata and in their resolution (Haslhofer and Klas 2010). Metadata thus often become a slowing down ballast in the analysis process.
Viewed as a framework for a snapshot, however, metadata can help give analyses conciseness and significance. Metadata, along with other assumptions and pre-decisions, create moments in a data journey that can now be considered separately. For the moment under examination, relationships of the data under consideration to the baseline situation and to the interests of the observer can be clearly stated. Questions become specific and thus clear answers become possible.
Is this concretization a limitation, a deficiency? Does it mean a weakening relativization of data science and analysis compared to data absolutism?
If analysis results are presented without reference to their assumptions and limitations, the discussion of these only shifts in time, it does not disappear, it takes place later. Doesn't it therefore rather make sense to take up the supposed lack right away and to give this lack its place in data science and philosophy of technology?

I will explore this question in the last two sections of this thesis. Before that, I would like to look at the inherent flaw in data (not only in data sets) from another perspective.

3.2 Promissory Data

3.2.1 Data as deficiency

"Data as promise are a more powerful resource than data as evidence" - this is the pointed conclusion of the Danish medical sociologist Klaus Hoeyer after his discussion of open data using the example of personalized medicine (Hoeyer 2019, p. 539). Contrary to the promise of providing clarity and ensuring decisions without political baggage, data turn out to be tactical and rhetorical tools, used mainly by their absence, not by their statement, to impose opinions. Hoeyer characterizes this type of data use and data as "promissory data."

Hoeyer examines the debate over personalized medicine and open data in medicine, using Denmark as an example. The promise of proponents of personalized medicine is that once we have enough publicly available data, we can develop personalized diagnoses and therapies that are far more efficient than conventional procedures. Critics, on the other hand, note that no data at all are yet available on the basis of which this statement could be meaningfully made. Between them, a debate develops about the usefulness and necessity of data, which Hoeyer analyzes in detail.

In this analysis, data is characterized by a deficiency. There is never enough data to make a decision. In one interpretation, this leads to a call for more data to support a thesis. In another interpretation, it leads to a degree of skepticism both about the thesis (which is not supported by data) and about the continually repeated call for more data.

This lack of data, which does not lead to a decision or is not yet sufficiently available to lead to a decision, leads to two shifts: One shift operates into the future. Decisions are not made now, they are made when there is enough data to make evidence-based decisions. The data will then, as data science promises, decide for itself. In this promise, the second shift makes itself noticeable: Decisions are shifted into the future and toward the data, thus relieving the burden on whoever is expected to make the decision. Hoeyer diagnoses an "ability of quantitative data to underpin systems of responsibility" (Hoeyer 2019, p. 531). Responsible parties are relieved of multiple burdens: They don't have to make a decision now. They do not appear weak when they do not make a decision, but matter-of-fact- and evidence-driven. And when they do make a decision, they are not actually making a decision at all, but are guided by the facts. What is decided upon, however, instead of the point-by-point decisions originally called for, is a framework of rules, thresholds, data sources, options under discussion, and other frameworks that impose far more constraints and conditions than a decision based on insufficient data would have imposed.

This kind of data discourse follows its own rhetoric. For Hoeyer, "data intensification is becoming a way for public authorities to respond to a wider set of problems related to budget constraints and demographic challenges, problems for which they basically have no solutions" (Hoeyer 2019, p. 531). The first step in this type of problem-solving is to refer to the need for more data. This reference, however, is not supported with data that should give it the authority sought by seeking more data. The reference to the need for more data itself becomes the new authority. In his analysis of debates on data and personalized medicine, Hoeyer notes that reports from management consultancies and trend researchers in particular are becoming a genre of their own, less in the nature of data-based documentation and more becoming a self-referential source of authority, making recommendations that function as self fulfilling prophecies (Hoeyer 2019, p. 534). Burdens of proof are shifted to the future. In the present, arguments are not data but narratives (Hoeyer 2019, p. 535).

This tactic has several advantages over limiting the analysis to available and predictable data. It is not limited to available data but can also speculate about data that may become available in the future. It can thus postulate computational paths and evidence that are not currently apparent. The biggest difference: It is not yet possible to argue meaningfully against data that are not yet available and against their consequences. The promise to bring about better decisions remains untouched. In terms of political or value-based argumentation, this is useful: "data-as-promise, or what I call 'promissory data,' is - politically speaking - a more powerful resource than data-as-evidence" (Hoeyer 2019, p. 539).

When data are used in this way, they help to defer public accountability and provide leeway for those in charge (Hoeyer 2019, p. 539). This is a markedly different outcome than transparency, clarity, and control as promised by data and open data in particular. Hoeyer's remarks can be used as starting points for ethical or political reflection. Other effects are also relevant for the knowledge-oriented perspective on data. Here, data from the future that does not yet exist generates effects in the present. The reference to data is almost as strong an argument as results and recommendations calculated on the basis of data. Data works in frameworks and is used in a value-driven way. Data do not have to exist, the facts do not even have to exist that could one day be processed into data in order to have an effect.

These strange gaps are one more starting point from which to ask the question about the essential properties of data. I also take the gaps as an indication that other concepts of data and technology will be necessary than the view of data as natural representatives that can be reckoned with unconditionally.

In Hoeyer's view, relying on data creates imprecision. It leaves very wide margins. The imprecision, actually indeterminacy, is created by giving great importance to data that do not yet

exist. They are supposed to decide in the future, however, it is supposed to be decided now that data will be decisive (and mostly also that data will be right). Thereby it remains open, how this power of data can be decided, if these data are not yet available at all. Where shall they refer to judgments, arguments, measurements and other usual analytical procedures?
We encountered similar uncertainties in the previous section with respect to open data. In this case, however, inaccuracies and uncertainties arose not from a lack of data, but from a plethora of different data sets that can be analyzed and interpreted in very different ways. The shortcoming in this case was noticeable in the frame of reference to be applied. open data publications such as the lobbying register lack a framework or network in which impact and meaning could emerge. Open data plans, as in the case of promissory data, in turn consist virtually only of this frame of reference, which predetermines a great deal of meaning before any data is even available.
In the case of the lobbying register, data generates structure and structure generates disorder - disorder then prevails everywhere outside the concrete structure (Boumans and Leonelli 2020, p. 94). Promissory data circumvent this as well. They do not create a concrete and tangible structure, thus they do not create specifically identifiable disorder. But if one leaves the framework they provide or contradicts the narrative, then there is no basis for discussion.

3.2.2. Data as creative force

Promissory data describe particularly clearly how data first create the universe in which they are then supposed to function as the most conclusive evidence. This is problematic for their role as preferred objects of science.
In considering fundamental questions about data, theories of truth are often useful as guides. This is all the more true when the question of the emergence of meaning is also discussed. Concepts of data as representatives are reminiscent of theories of correspondence; data represent something, their value arises from this mapping relationship and from their property of providing a particularly clear mapping for certain features or purposes, which illustrates certain facts particularly well.
Relationship-oriented data concepts, as Leonelli contrasts them with the representation model, replace this representational connection between data and objects with relations of data to one another (Leonelli 2020). Data form networks within which they give feedback to each other, which can be processed into information and meaning for their users. Getting behind the network, behind the data, is of less interest. Even there, behind, is always only data after al.

Both notions reach their limits when we dispense not only with the objects "behind" the data, but also with the objects themselves. This is exactly what happens in the case of "promissory data": data that do not yet exist are supposed to provide evidence for something that also does not yet exist; and this connection is at the same time considered as evidence that both the missing data and what they are supposed to prove are necessary. Here, the relation to something outside the data is missing, but we also cannot establish relations between the data.

3.2.3 Data Science as long chains of decisions

Computer scientists and mathematicians often portray the data science process as a path lined with clear instructions. Perhaps some iterations are necessary, but the path leads in a clear direction and deals with data. At best, data is matched with domain knowledge, but otherwise follows its own mathematical and technical rules. They can be made to speak with statistics and computer science, and are close to the truth, provided they are not corrupted by confounding factors.

I think this scenario is not tenable even as a theoretically desirable ideal scenario. Data alone do not succeed in producing concrete and unambiguous meaning. They rely on a whole set of relationships, extensions, conditions, and reassurances. Attempts to ascribe more authority to data and data science recall deterministic conceptions of technology that see technology as the heir to magic and algorithms as the descendants of oracles (Marenko 2019).

Data can take a variety of forms and be viewed in very different ways. Data and the data science process are often given great authority - this is difficult to justify. After all, the data science process is only the preliminary end of a long series of decisions that must be made very early on, far earlier than the data science promise would suggest.

If we keep the two examples from this section in mind, does this help to find out, where decisions happen in the data science process? And how do they happen?

4 Data, Decision Making and Open Issues

4.1 Decision Levels

The two examples from the previous section show the many levels at which decisions are made when working with data. The individual steps can be grouped into several stages.

Essentially, working with data is divided into two major areas. The first area is the collection and preparation of data. Here, questions are identified that are to be processed in more detail with the help of data. This step of data work ranges from the identification of a subject area to the collection of data, data engineering in the narrower sense, and the publication of prepared data sets.

The second major area deals with the analysis of processed data sets and the attempt to derive specific recommendations for action based on even more specific questions.

Open data occupies an special position with regard to this distinction: In many cases of work with data, the separation between these two phases is a smooth transition. Data is collected and processed in order to be able to analyze it. In organizations, these activities are often performed by different departments or teams of specialists, but the boundary between these steps is permeable.

In the case of open data, the boundary is clearer: data is collected, prepared and published within an organization to be made available to the public. Data preparers are no longer part of the analysis process; on the contrary, the point of open data publications is to let others work with one's own data to address new questions that were not even on the horizon and to produce new results.

At first glance, this division sounds like a simple and clear process. First, data is collected, sorted and processed, then it is detailed, analyzed and finally a decision is made - just as the ideal image of the data science process promises. However, we have already seen in the first sections of this paper that the boundaries cannot be drawn so clearly - and that, above all, they do not draw themselves.

I further distinguish each of the two major steps - preparation and analysis - into yet another five stages each in order to get a more specific picture of the data process. Each of these stages focuses on different data handling priorities. The big picture shows how data creates its own universe and its conditions of validity.

4.1.1 Data preparation and data engineering

4.1.1.1 Subject area

The first major phase begins first and foremost with a decision: What should it be about? On which topic should data be collected and processed? Who will do it, with which methods and with which objectives? The selection of a topic is the first major abstraction. The decision to deal with one topic (and not all the others) creates distinctions that make specification, ordering, structure, or categorization possible. This decision has an economic and an ontological dimension.

At the economic level, resources are allocated through decision-making and prioritization. This can bring about an ordering, but it can also mean an either-or decision - with the consequence that other issues cannot be dealt with. Social and political decisions are made here as well. Particularly with regard to open government data, the specification of a data and topic area is a decision that determines what is published and documented and what is withdrawn from the public domain. In scientific or technical subject areas, domains are constituted by such decisions; focal points such as age- or gender-specific prioritizations are also potentially far-reaching decisions.

The decision to make a subject area the object of a data process contributes to the further delineation and sharpening of that object; it is the start of a process that is intended not only to procure and prepare data about a subject area, but in the process also to specify the object itself. That's the ontological dimension.

This selection is a first decision, which could practically always have turned out differently. Subject areas are not predetermined or self-constituted - even the very first basis requires a decision.

In the specific case of the transparency register, for example, relationships between politicians, business and interest groups are the broadly defined subject area, which is further narrowed down and specified by assumptions. A hitherto poorly defined relationship is specified in order to describe it in more detail. With the specification come valuations - or are these represented by the specification? This question must be left open at this point.

In the case of Danish health data, the issue is even fuzzier. Public and networked health data is both a threat and a promise. How and whether health data can be used to advance research and therapies depends on the extent to which it can be harnessed. The subject area is therefore only just being constituted, and it thus also creates the framework within which further questions can be posed in a meaningful way. Due to the mobility of the topic, which is currently still being

discussed openly, definitions have further-reaching consequences. And they anticipate what, according to their own claim, data should first prove.

4.1.1.2 Questioning

A selected topic needs to be specified by more concrete questions. These are decisions that further narrow the perspective. This excludes even more potential data, but the section of the topic area that remains in focus can be covered more precisely. That sounds trivial. However, in the wake of data science and big data, inductive methods and the hypothesis of abandoning hypotheses are celebrating a comeback. Representatives of these positions even argue for new data science-specific forms of induction (Pietsch 2021).

Part of the specific research question are initial hypotheses. Specific assumptions may be tested, revised, or discarded; they may also turn out to be untestable or inappropriate to the topic area. The same is true for the idea of hypothesis-free and purely data-driven research. Freedom from hypotheses generally means fuzzy hypotheses that nevertheless put certain relationships in place. Explorative data work in the case of unsupervised machine learning also follows a hypothesis. Tasks like clustering or ordering do not necessarily imply a concrete hypothesis about the criteria according to which clustering or ordering should be done. But they do imply the hypothesis that clustering or ordering makes sense with respect to this problem.

Initial hypotheses develop differently, depending on whether data are already available (even if they may not yet be considered data, more on this in the next point) or whether the collection of data must first be organized. In the first case, questions can be very open-ended - "What do I see here?" is an initial question, for example, which may well guide initial exploratory analyses. In the second case, hypotheses must be oriented more towards the future. Ideally, they must try to anticipate how the collected data sets will be worked with in the future, in order to collect data with which these work steps can also be completed.

Accompanying the specific question are considerations for operationalization. Thus, in order to answer questions, consideration must already be given to what can be accepted as a potential answer, so that in the data preparation, elements can be sought that are suitable for this purpose. Here, too, there is a strong indication that data can hardly be regarded as simply found things, givens or privileged representants. Their creation is based on purposeful processes.

As a representative of an inductivist perspective, Pietsch (Pietsch 2021) argues against this assumption. His example: The pixels of a digital image are data, no matter for what purpose they represent something. I would counter that pixels are not data representing the image - they are the image. If, on the other hand, it is the content of the image that is to be captured, then a

number of further specifications are necessary. For this, I return to the example of the lobbying transparency register.

In the case of the lobbying transparency register, no data are yet available at a level that would be ready for analysis. The facts exist, but they have not yet been systematized and logged. Relationships between politics, business and interest groups are to be documented in order to make possible influences visible or to show the absence of influences. To this end, further delineations need to be created: Which companies, which interest groups are considered lobbyists? Which economic metrics are relevant? Which contacts and activities should be recorded? How can they be operationalized, in what ways should they be able to be linked to meaning later, possibly only in the analysis process? Many of these questions already refer to the next point, to the central question of every data process, which deals with what, which reality fragments, metrics or values can now actually serve as data for the specific case. From a political point of view, however, it is also very relevant who decides on these questions and thus decisively shapes the benefits and scope of open data publications.

In the case of networked health information, security and privacy issues must be discussed before specific questions can be addressed. It also remains to be seen which aspect of the topic should be prioritized: personalized therapy, prevention, documentation and research are different focal points with different issues. The lack of data means that very concrete hypotheses have to be formulated, for the proof of which corresponding data are then to be sought. In his remarks, Hoeyer illustrated how this takes the project quite far away from an ideal-typical data process. However, it becomes all the more clear on this path that data processes cannot function without specific questions, progressive specification and thus a series of successive decisions, and that in early stages.

For only after these first two preparatory steps of data preparation can one of the central questions be asked: What should actually count as a datum in the current case? And how do we find and recognize these fragments of reality?

4.1.1.3 What is considered a datum?

Data is the given, which we simply find. Data speaks for itself. These are the quick prejudices and the great promises that are readily associated with data. But it is also the prejudices that we should abandon, as Helen Longino for example argues (Longino 2022 p. 391). The decision to regard something as a datum must be actively made - and it rests on a set of assumptions and entails a set of consequences.

The essential characteristics of data are described in many different ways. Data have been described as privileged representatives, as indicators that can acquire and convey meaning only

through relations (see chapter 2). The definition of data that appears suitable often depends on the specific context and objectives. This is one of the few points on which many different perspectives on data almost agree.

Specific data practices must therefore be as much a part of the analysis of data processes as the consideration of algorithms, queries, and data models. Data brings a history with it. Data comes from processes in the past, and it is often tasked with helping shape knowledge and decisions for the future. Data comes into being with a purpose, and the decision to consider something as a datum gives an object or its fragments a purpose-bound property that can take different forms depending on the specific purpose.

What elements are considered data in the case of the transparency register? In addition to the master data that identifies an organization, it is metrics on economic strength and content focus. Even in the first attempts to deal with the collected data, however, it becomes clear that it is not so much the data, its categorization and structuring that is the relevant variable. Often, analyses fail to produce meaningful results because the purpose of specific data can no longer be discerned, because their context and conditions of origin can no longer be conveyed. It is no longer clear, which decisions in the data gathering process were left to the discretion of the individual officer and which were to follow strict procedures.

The less the decision that makes an element a datum for a specific purpose is addressed, i.e., the more reliance is placed on this decision being self-explanatory, i.e., not actually a decision, the more leeway is created when working with data, and the more unstructured and unsystematic these decisions are made. This has major implications for data quality and consistency.

In the case of public health data, due purpose orientation becomes even more apparent. On closer inspection, there are two kinds of data here: First is general health data, which is collected and linked. Privacy and data protection issues are relevant in these decisions, as are technical questions about storage capacities, interoperability and specific processes in doctors' offices and hospitals. The second type of data, which is the one actually relevant to the data science promise in this context, does not yet exist: there is a lack of data on whether and how widespread collection of health data actually helps to improve specific health issues or therapies. However, the effectiveness of this still missing data is postulated and set as the driver for the need for further data collection. Here, the decision of what is now a datum is, in a sense, reversed: datum is what helps fulfill the promise. This is a reversal of the data science promise: data does not lead to decisions, data helps to argue decisions after the fact. If the promise is not fulfilled, then it is the wrong data, too little data, or in some other way insufficient data. The

property of being datum is clearly tied to a specific purpose. This binding is so strong and so purposeful that it already seems exceedingly questionable in this case.
Two essential basic questions come up again at this level of decision-making: Once it has been decided what counts as a datum, how does this datum acquire validity and meaningfulness about the object described? And how does its meaning become specific? These two questions are always at the center in the second major step, in the evaluation and analysis of data. Before that decisions have to be made on two more levels in the process of data collection and preparation. The next level of decision deals with the issues of measurement, scaling and categorization.

4.1.1.4 Measurement, collection, categorization

It has been clarified what is to be considered a datum, and there is also a plan for where this data is to be found. Next, decisions must be made about how to collect the desired data, how to structure and store it, whether and when to subject it to categorization.
Decisions at this level also work in two directions: They specify what can be recorded as data and how detailed which properties and functions of data are to be preserved for evaluations and analyses in the future.
This becomes particularly clear in quantitative surveys of the question of threshold values: Should all expressions of a characteristic be documented? Or are only those characteristics considered relevant that exceed certain thresholds? In the first case, data components are lost; moreover, in order to determine the threshold values, it must already be clear which limit ranges are useful and what these mean. In the second case, there is a risk of excessive data volumes and falsification by outliers, which have to be cleaned up afterwards. However, in this way, the dataset as it has been recorded is still available if the thresholds for outliers have to be recalculated.
This is a matter of idealization and necessary tradeoffs. Weisberg (Weisberg 2007) describes various types of idealization and various representational ideals. Ideals focusing on simplicity or monocausality are types of idealization that would probably rather favour some reduction in data. Completeness or maxout are other representational ideals from Weisberg. Data processes following these principles would rather build data models that summarize or aggregate as little as possible. All ideals and all types of idealization come with certain tradeoffs. In the case of data collections, the big temptation is to collect as many data as possible and avoid any sampling. This is done at the expense of processability, if at all possible: technology and computing capacity quickly reach their limits.

After the content modeling, which became more specific in the previous decisions, the level of specific factual and technical modeling is reached with the questions of structuring and categorization. In data science or modeling textbooks, this step is often treated as a formal matter that must satisfy technical principles. Certain data formats are intended for certain content and bring with them technical advantages and disadvantages. Cardinalities and their notation follow strict guidelines. Principles such as the rules of normalization of data over several levels give the impression that these steps follow a necessity.

In data-oriented disciplines such as Digital Humanities, other approaches in the preparation of data take hold, which are directed against strict normalization processes, for example (Schreibman et al. 2004). However, decisions in categorization, structuring and further abstraction steps are not limited to technical and mathematical points of view. The definition of scales, their calibration, and the determination of measurement parameters are technical steps through which values and evaluations find their way into data science processes in a very specific and measurable way.

Many of these conditions apply to all areas of science. Data science, according to some accounts, has stepped in to address them. Data science wants to create clarity and help to make automated decisions (Kelleher and Tierney 2018, Irizarry 2019), other perspectives on data science emphasize that the necessity of ordering hypotheses fades into the background when using computational methods (Pietsch 2021). However, it is precisely the decisions that supposedly follow only factual necessities that are often and necessarily based on value judgments. Continuing Hempel's reflections on inductive risk, Douglas has shown how on the surface technical decisions reflect value attitudes and questions of responsibility. Central to this is the question, "What if I am wrong?" Returning to the threshold example: Data collection without thresholds allows for constant recalculations, so thresholds could also be recalculated. On the other hand, once thresholds are induced, it is difficult to revise them; not every data collection process can be repeated. Thus, to keep the inductive risk low, thresholds can only be used where the impact of a possible error is judged to be low. So a data model considered in this way also maps value decisions. (Douglas 2000)

Finally, specific decisions about the data model also have a significant influence on how data can acquire meaning and how data becomes information. On the technical level, number and text formats can be sorted differently and related to each other in different ways. No direct comparisons are possible between non-numerical values. Numbers can be summed or calculated as average values, texts or date formats can at best be aggregated to maximum or minimum values. Thus, the data format as a manifestation of the modeling and structuring

decisions determines which options are available in the subsequent phase of evaluation and analysis.

In the case of the transparency register, the lack of decisions is particularly noticeable. In the course of data collection, apparently very few specifications were made as to the form in which data should be provided, which data fields represent mandatory information and which specific information needs should be satisfied with individual data fields. The value "two million", for example, can be found in the formats "2M", "2000000", "2,000,000", "1.5m-2.5m". This is an obvious problem in the specific case; however, in a historical analysis that has to deal with different currencies, measures or exchangeables, this openness can be useful. Free text fields are always very difficult to evaluate when a certain number of records is exceeded, multiple mentions always lead to the question whether sequences represent prioritizations, with which separator system multiple mentions within a field can be distinguished from the next data field without making the later evaluation unnecessarily more complicated, and how and whether individual mentions can be related to metrics. I have already explained the shortcomings of these decisions in the previous section (chapter 3.1).

In the case of transparent health data, it is difficult to write about structuring and categorization because the desired data are not even available yet. Nevertheless, there are already some questions in the room: According to which criteria should commonalities be established in order to be able to transfer findings from a patient group to a specific patient? How will the lack of information be documented (for example, about training status or nutritional behavior), when are such information gaps grounds for exclusion for certain data sets, when does the lack of information play a less important role? How should decisions be made about such questions when the relevant data and experience about the effects of working with these data are not even available yet?

This leads to the last major group of decisions. In the process of data collection and preparation, many things are in motion, decisions intertwine and can be revised in several iterations. For data to serve its purpose, it must be made available in the form of data publications. In a research or business context, the boundaries are often fluid, and contacts between data preparers and data users are possible. In the case of open data, however, the circle of potential data users is basically unlimited and unpredictable. All the more relevant are those decisions that are made around the publication of data sets.

4.1.1.5 Publication

Data depend on context. They are not immobile, immutable, they can be analyzed for different periods of time, be stored, reactivated, and mobilized. For Leonelli, this mobility of data is one

of its central characteristics (Leonelli 2020). She specifically sets this characteristic against the theses of other analysts who set the immobility of data as its essential characteristic (Leonelli and Tempini 2023, p. 23, Pietsch 2021, p. 10). Yet, for Leonelli, data are also immobile, despite their mobility, in the sense that they can perform their task in different environments and time horizons. This, in turn, is linked to specific purposes, conditions, and limitations; "using data as evidence is everything but straightforward, and sophisticated methods, resources and skills are required to guarantee the reliability of the empirical grounds on which knowledge is built" (Leonelli and Tempini 2020, p. 26).

Data publications are, on the one hand, data-intensive visualizations, animations, and graphics that seek to explain facts based on data. On the other hand, and especially relevant in the context of open data, this refers to data repositories in which data sets are accessible in a citable and processable form. Machine readability, interoperability and, for better traceability of past preparation processes, descriptive metadata are considered quality criteria for data publications. Data publications are the result of the decisions made so far in the data process and those "sophisticated methods, resources and skills". Data publications should anticipate what questions can be asked of them, ideally they should be able to serve a variety of questions to the data, but they must also define boundaries. Not every data publication can serve every purpose - publications can only work for those purposes that were addressed at the previous stages of the data process and carried along to this point. Unless, in exceptional cases, it is precisely gaps in the documentation that open up new possibilities and new interpretations.

At this final stage of the data preparation process, data can no longer get better (unless through multiple iterations over the previous stages), but it can get worse.

The broader and more diverse the potential audience of a data publication, the more relevant become the ability of data publications to be understandable and applicable even without ongoing and detailed commentary by the creators.

In the case of the transparency register, the many quality criteria are not met. Within the publication different standards for data formats and data structuring are used – the publication can only be evaluated with high effort and can only be linked with other data sources poorly and in a very limited way. There are also hardly any descriptive metadata.

Especially in the case of open data, these characteristics of data publications are extremely relevant. In addition to technical criteria, they also touch on social and political issues such as freedom of information, access to data, privacy, information security and information rights (Gray 2014). – These are also criteria that would be extremely relevant for transparent health data but cannot yet be discussed in this specific case due to the lack of specific publications.

In the process of data collection and preparation, a multitude of decisions have to be made on a number of different levels. The result of this process is what is often referred to simply as data: a data publication, which is already the result of elaborate processes. This can now be followed by the next level: The process of data analysis can also be broken down into several stages regarding the necessary decisions.

4.1.2. Analysis

4.1.2.1 Questioning

We have already encountered questioning as the first stage of decision-making at the level of data collection as well. The very first clarification deals with the choice of topic and its concretization. This decision was made, why is it here again?

At the level of analysis, the question is further specified. In addition, a change of perspective takes place. During data collection and processing, decisions were made about the topic and focus; ideally, the result was data published in such a way that a number of questions could be addressed to it. Thus, specific questions were always future ones, potential ones, and rarely one's own. At the level of data analysis, on the other hand, a specific question becomes actual and manifest. Data collection presents a collection of data to which acute questions have not yet been asked. As long as the engagement with data is at this stage, it is tempting to collect as much data as possible and to avoid categorization and aggregation. This reduces the inductive risk and it broadens the promise of data science to be able to provide many answers.

However, with actual and manifest questions, it quickly becomes apparent that without categorization and aggregation, it is hard to get meaningful answers - otherwise the answers just repeat the input of the data collection. Thus, to understand the answer, the whole data set would always have to be kept in mind. This theoretically improves the scope of the answers, but reduces their quality. They are hardly understandable for humans. Conversely, specific questions or working with aggregated data reduce the scope of the answers, but this gives the answers more concrete meaning and thus quality.

Analysts look to data publications for as many answers as possible that are as interconnected and interrelated as possible and as unambiguous and clearly assignable as possible. For data to do this job and help provide these kinds of answers, analysts must decide on one specific question at a time. Individual questions can certainly be aggregated into more complex metrics, but data must be specifically interrogated in order to provide comprehensible answers.

In the case of the transparency register, possible specific questions include: Who are the most active lobbyists? Which lobbying topics are most important for clients from which country?

These are specific questions, but they also need to be further broken down and operationalized so that answers are comprehensible and as direct as possible. This necessary operationalization leads to the next decision level.

4.1.2.2 Indication, meaning, interpretation

Initial results to specific questions often raise the further question, "So what does this mean now?" or "Now where is the evidence and the robust indicators for this initial assessment?" Just as decisions had to be made at the level of data collection about what is considered a datum, for what reason, and for what purpose, the process of analysis must further flesh out these decisions. The datum-object-purpose relationship is extended to the datum-object-purpose-question relationship.

This is often done experimentally and by reassurance in the network. First results of explorative analyses have to be validated or rejected by controlled changes of parameters. This can be used to validate which indicators actually stand for which properties. Repetition with repeatable results creates reliability and accuracy.

Ideally, at this stage, analysts can draw on the decisions made at the previous stage by having documentation, metadata, and definitions of terms accessible to them. Without these frameworks, data would have to speak for itself - and data is very bad at that, as we have now seen many times. A difficult trial and error phase begins for analysts, who face uncertainty on two levels. Unannotated and undocumented data publications are unstable on a technical level and on a content level. It is not clear which elements can be linked and how, and which relations are functionally permissible. It remains a question of experience to assess initial results. If this experience is lacking, results must be compared with information, if possible outside the data publication, already in this early analysis phase in order to be able to perform meaningful plausibility checks.

The decisions at this level are closely related to the decision of the specific question; the exemplary questions are also similar. If – to take up the transparency register again – questions were asked about the most relevant lobbying topics of each country, then the relevant dimensions and metrics must now be clarified. Does the frequency of mention of a topic apply? Should lobbyists or clients be scored, especially if they are headquartered in different countries? Should criteria such as turnover and number of employees of the lobby organization or expenditures of the client organization also be evaluated? Can an indicator be calculated that weighs expenditures or revenues per issue? Is it even possible to make meaningful statements without such indicators?

Technical know-how, expertise and, if possible, broad access to documented or metadata-based preliminary decisions of the data preparation phase are indispensable here - all the more so the more abstracted and automated the analysis is to be. Data are context dependent. Their reference to an object is borne by a purpose, so that meaning becomes ascertainable. And in order to be able to decide on the meaning, a specific question must be in the room.
So there are four interrelated aspects that together influence whether data provide answers and how their truth is judged. For representatives of inductive data science approaches, there can be no wrong data; wrong is at best the interpretation (Pietsch 2021). However, it must be countered that data without interpretation have no meaning and that there may very well be false data, in the sense of unsuitable data for specific questions. Not all data from a specific topic are suitable for answering all questions on that topic, even if they seem to provide answers.

4.1.2.3 Compatibility, interoperability

At the level of indication and interpretation, the relevant questions are primarily semantic. Here, content and the emergence of meaning are negotiated.
When decisions have to be made about compatibility and interoperability, technical details are up for discussion. Now it has to be clarified how different tables, nodes or data sets can be related to improve information quality and to be able to draw new conclusions.
In part, these questions can be clarified on the basis of technical criteria. This applies to aggregation and linking methods, but the prerequisite here is also clarity about the available data formats and, in case of doubt, sufficient documentation or meaningful metadata.
Documentation and standardized metadata become imperative when links to other data sets from other data publications are to be enabled. Established standards allow links even if different data fields are labeled differently, even different content can be merged using transformation rules. Common examples of this are classification schemes for media data, where fields labeled as "author" or "creator" mean the same thing. Strictly normalized datasets that store first names and last names in separate fields can be matched with less strictly normalized datasets via transformation rules.
Thus, it is not the data itself that determines the extensibility of data sets, but the form of its storage, structuring, comprehensible standardization and the relationships that can be established to these frameworks. None of these are mandatory consequences that result from the existence of data; they are decisions that must be made on the way from data to information.
In the case of the transparency register, for example, such decisions have to be made when the published information on lobbying organizations is to be matched with the MEPs'

documentation on their lobbyist appointments. Other use cases of decisions on interoperability are comparisons of financial key figures (turnover, number of employees) with industry key figures in order to be able to make classifications on size ratios. Forming time series, whereby changes in relationships between clients and lobbyists or developments of individual lobbying organizations can be made visible, are further use cases in which the compatibility of different historical versions of the lobbying register itself must be checked in particular.

4.1.2.4 Communication, Argumentation

The previous three stages served to make decisions about which specific issue to address and which content and technical criteria to use to create and justify meaning.

The last two steps deal with the evaluation and dissemination of results. At the level of communication and argumentation, results from the data process are available; specific questions have found answers, they have been related and can claim to be information. How does information become relevant, how does it gain specific meaning that can be understood as a call for decision?

Results and content must be distinguished from other relevant results and content in order to be put into relation. This results in specific evaluations and classifications; only here can data ultimately actually become rhetoric and fulfill the purpose of convincing others. In connection with a specific purpose, recommendations for action can be derived from data at this stage, which are valid as long as the specific purpose is valid.

In the case of the transparency register, framework conditions that help to arrive at specific evaluations are, for example, the comparison of the real documentation behavior of MEPs with legal documentation obligations, the comparison of country-specific lobbying priorities with country-specific interests and conflict potentials, or the classification of metrics in comparison with other metrics, for example in order to be able to compare the economic strength of lobbying organizations with other public affairs or public relations organizations.

These comparisons leave the specific data level, they do not take place on a technical-mathematical level, but entirely on a level of rhetoric and argumentation. Their persuasiveness depends on the decisions made in the phases of data collection and preparation, which metrics are documented how precisely and according to which criteria - in short: data-based arguments depend on the inductive risk taken by data collectors and preparers, even if data scientists emphasizing inductive approaches aim to eliminate these tasks and present them as following simple technical necessities.

4.1.2.5 Action

A series of decisions have been made, in an ideal data science process, data has been loaded with meaning via specific relationships and has become information, information can be used to argue in specific contexts and for specific purposes, recommendations for action can be derived from the arguments. These recommendations for action are the last major stage of decisions, they are those decisions that data science wants to improve.

How does something follow from data? And under what conditions do these consequences acquire authority, i.e. how does the chance increase that these consequences will also have an effect? This question sounded simple at the beginning. Data show the right way, it is simple and only logical and rational to follow what data say. The only thing is that data basically say little, they do not speak for themselves, and they do not speak unambiguously.

In part, this is due to the contextual and purpose-dependent nature of data.

In part, this is also due to the question of where we locate data. Are data part of the reality under study? Are they representatives that are by virtue of a property related to a reality? Are they indicators that produce new knowledge through connections, interactions, and reactions among themselves?

Weisberg addresses a similar question in asking whether the study of models requires a third step after the work of creating the model and actually working with the model, so that the results of working with the model can then be applied back to reality (Weisberg 2007). For Weisberg, this question arises from the distinction between theory and model. The theory deals directly with the system under study, the model creates something else. Other authors approach this question by considering whether working with models is a distinct form of knowledge generation that does not necessarily need to be transferred back to a reality that is different from the model in order to be valid (Knuuttila 2021). Some authors also make explicit reference to data in the context of such model discussions. In this context, data sometimes take on the role of a link, as if they could switch between model and reality and were thus the actual carriers of those insights that are needed in reality and are conveyed by models (Giere 2010).

Data approaches that emphasize the possibility of purely inductive findings or view data-driven research as hypothesis-free skip these issues. According to these approaches, data lead through processes and to decisions that are valid primarily because they come from data. Pietsch describes some such approaches (Pietsch 2021).

Context, purpose, and concrete conditions of data practice are disregarded. From the point of view of data absolutism, this focus on context and conditions would lead to an unacceptable and unnecessary relativization, whereas data could in principle answer all open questions (Mayer-Schönberger and Cukier 2013). However, this neglects a central feature of arguing on the basis

of technical criteria: Technical arguments are decisions. They aim less at deduction than at design and definition (Vermaas and Kroes et al 2011).
This turns data-driven necessities into decisions that have to be taken actively. Decisions improve the quality of data processes and data-driven recommendations, but they reduce their potential scope. This contradicts the universal promise that data science would like to make; it undermines the pose of comprehensive transparency that open data likes to claim for itself.
After having addressed decisions in data processes, I want to discuss how this focus on many decision levels in data science affects the emergence of meaning and authority through data science processes.

5 Create meaning and effects – Data, information quality and authority

5.1 Data will fix it?

Many public and political expectations about data science and open data can be traced back to two influential essays. Data is helping inductive methods to make a comeback, theories are no longer necessary for knowledge production, postulated Chris Anderson, then editor of the technology magazine "Wired": "Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all" (Anderson 2008). Thus was born the idea of a kind of automated decision-making that could be applied to science, politics, and many other areas of life. Data processes, so the promise, deal exclusively with facts and thus deliver the right decisions.

This idea found its continuation in the diagnosis that with the optimization of technical processes, which could process ever larger amounts of data, this shift from hypotheses and models to unconditional empiricism had been taken to a completely new level. With big data, distortions that could arise from sample selection would be reduced, and it would no longer be necessary to speculate about causality, which would be replaced by correlation, Mayer-Schönberger and Cukier wrote (Mayer-Schönberger and Cukier 2013).

Despite frequent criticism, these approaches had great impact; "hypothesis free" is still a frequently used catchphrase today. The most effective objections to the theses of the end of causality and hypotheses focus on two thrusts: First, causality has never been an uncontroversial concept in philosophy. At least since David Hume the inductive conclusion from observations to laws is questionable. Moreover, causality in philosophy is to be separated from teleological concepts of causality, which seek not only causes, but meaning, ends and reasons. The second important objection relates to the role of induction. Induction starts from observations and draws conclusions. Particularly in data science processes, these inferences must be aligned with a target conception. This goal conception does not necessarily specify content goals. The specification to form clusters or to draw clear boundaries based on the expression of various characteristics is also a goal. Data science processes (particularly evident in the case of machine learning processes) approach these goals iteratively. Conclusions are tested and optimized or discarded. This requires at least abstractly specified

goals - ultimately hypotheses. Pietsch calls this type of induction "variational induction" and sets it as a typical type of induction for data science and machine learning algorithms, with which both the requirements of hypothesis-driven science and those of data-driven exploration, which is as presuppositionless as possible, could be fulfilled (Pietsch 2021).

5.2 Authority as a matter of scope

These questions lead into the heart of the issue of how the results of data processes acquire meaning and validity. The question arises from the idea of being able to achieve different or better insights from data than from hypotheses and theories. In this context, data should not only play the role of validating or challenging theories, they should provide their own insights - which may also lead to new theories.

Data are thus in a certain sense detached from the reality about which they are supposed to say something, but at the same time they are connected in a special way.

This changing relationship of data as object and instrument shows how far data are exposed to very specific practical contexts and circumstances if they are to be a meaningful source of knowledge.

Data processed in data science processes using mathematical and computer science methods are abstract entries reduced to numeric or logical form that are intended to clearly represent certain properties. Captured data refer to individual entities, aggregated data calculate probabilities for groups of entities. These probabilities can be used, for example, to calculate predictions for behaviors. At a descriptive and analytical level, the processes here are clear. The statements take something like the form, "This is how it might be at probability X."

The situation becomes more problematic when it is no longer just analysis that is required, but when concrete results are to be achieved with actions based on these findings. The question, "How can I increase probability X?" requires, first, a hypothesis about goals associated with the change in that probability. Second, at least a minimal concept of causality is necessary. If outcomes are to be changed, it must be possible to identify the factors relevant to the change. A rationale about the meaning and purpose or the precise "how" of this relationship is not yet necessary, yet the necessary relationship goes beyond mere correlation in that the controlled characteristics not only co-occur, but one characteristic does not occur without and responds to changes in the other.

Both the hypothesis (as a focus on a specific expected change) and the isolation of specific characteristics simultaneously lead to a narrowing and a broadening of the scope of data

results. The narrowing concerns the quantitative scope and generalizability of the results. Despite the reduction to mathematical and logical features, data lose the character of universality if they are supposed to answer specific questions. The extension concerns the qualitative aspect. With reference to concrete environments and tasks, data can provide clarity and improve decisions.

Unlike the original promise of data science, however, this improvement does not take place as a genuine part of the data science process itself. In "Doing Data Science" (O'Neill 2014), mathematician Cathy O'Neill strives for a very broad concept of data science.

O'Neill's Data Science process has multiple stages and iterations and, unlike the process outlined by Kelleher and Tierney, options and forks (O'Neill 2014, p.41). The initial stages of data collection, data processing, and data cleaning proceed in a straight line. After that, possible paths differ depending on how far the original problem can be addressed at that point. If no more questions are open, then visualization, reporting and communication can follow. If no results are visible yet, then explorative analysis is the next step: data is tested against various suspicions to find new correlations or clusters that can be used to revise models or that reveal errors in data mining or cleaning with unexpected results. New phases of data collection often follow this phase, and sometimes the findings are transferred to machine learning phases in which algorithms perform the exploratory work. However, machine learning phases can also immediately follow data cleaning. The preliminary endpoints of data science processes at O'Neill are the communication of results via visualizations and reports or the creation of data products.

Unlike Kelleher and Tierney, O'Neill's decisions are no longer a direct part of the data science process. Decisions are the goal, but they are part of the data science process to the extent that the outside world is also part of the data science process as a universal data provider. Without data found in the world, there is no data science process, but the world is not exhausted in its role as a data provider. Decisions can be based on the findings of data science processes, but they do not automatically follow from them.

This is very different from the data science process outlined by Kelleher and Tierney or Irizzary. There, decisions are integral elements of the process; data science is something that is applied to decisions and makes them better. Conversely, according to this description, data science processes without decisions are incomplete. For O'Neill, the data science process ends with the creation of reports or data products. Decisions can follow, just as data products can be released into the world as applications. But neither are mandatory consequences. When, if we take this separation seriously, are decisions made based on data and when are they not?

What rules do abstraction and reduction steps follow to decompose "the world" into data science-appropriate inputs? So how does data fulfill its role as a carrier of meaning at both ends of the data science process?
Various perspectives on data processes have in common that they are characterized in detail by different shifts in their orientation (to the future or to the present) and in their scope (narrowing or extension).

5.3 Tradeoffs

In considering the wide range of decisions in the data science process, I have noted the shift from the mathematical-logical to the concrete and contextualized in two places in particular: With the transition from data collection and preparation to data analysis, potential and future questions become concrete and specific tasks. Data preparation creates possibilities, data analysis exploits them. Data preparation must keep contingencies and potentials in mind, data analysis wants to focus on concrete givens. Data analysis must know that the analyzed givens are the result of changeable processes - but during data preparation we must pretend that this limited view of the world is the result of unchangeable authority so that the prepared data are meaningful and unambiguous and clarity can be achieved.
In the further course of the analysis process the next restriction takes place. Interpretation and concrete meaning generation are iterative processes. Machine learning processes in neural networks are often used to illustrate this iteration. Analyses pass through neural networks multiple times, individual neurons can learn and change in the process, thus the result changes. Data is related to an object for a specific purpose - as such, data enters databases in the course of data preparation or collection processes. The purpose of data is to represent a property. They achieve concrete meaning through the connection with an unambiguous question, in which their special quality (from a computational perspective: the reduction to logical-mathematical calculable aspects) can contribute to giving particularly clear answers. These two limiting processes concern the scope and thus the authority of decisions made based on data. As long as data science processes insist on the universal claim of being able to provide decision recommendations based on technical analyses, their recommendations are very abstract and at the same time very specific. The generalized consequence of the idea of being able to answer all questions particularly well with a population of n=all (as Mayer-Schönberger and Cukier advocate) are answers in the form of, "If you want more X, do more X." This answer is rarely wrong, but rarely useful. It lacks information about which composite

goal X is part of and what role it plays in the composition (In the reverse case, if X is already the composite goal, it lacks information about which individual components can be used to achieve it - because that would again necessitate the minimal concept of causality).
In order to be able to create more concrete benefits, the scope of the answers must be limited. The strength of data in decision making is therefore not that as much as possible can be answered, but that many variations of questions can be answered as efficiently as possible.

5.4 Technology is not about truth but about decisions

One cause of this misunderstanding about the actual strength of data processes is unclear underlying concepts of technology. I consider data science here as both technology and scientific method, and focus a bit more attention on the technology aspects because it is in the technology domain that misunderstandings of how to prioritize data are particularly pronounced.
Technology concepts map a wide range between technological determinist and social determinist perspectives; other dividing lines of different technophilosophical perspectives run along the question of how far technologies directly influence reality or society and how far this influence requires mediation and moderation. Deterministic positions, for instance, can be well reconciled with the idea of replacing causality via data by correlation - data would be right, the analyst would have no choice here. Technology as a directly structuring force can be found, for example, in Winner's theories; for Feenberg, on the other hand, mediation strategies are always necessary to "reworld" "deworlded" technology and thus make effect possible in the first place (Coeckelbergh 2020).
Diverse technology concepts share an essential aspect that is often overlooked in the application of data science as a technology and method: Technologies are not a means of finding the truth. Technologies can be used to bring about decisions, technologies can provide arguments. However, the focus is always on the purpose. Technologies do not provide reasons, technically driven explanations do not provide meaning. They focus on questions like, "Does it work?", "Is it useful?", or "What does it consist of?" (Vermaas and Kroes et al 2011).
Such questions are only useful if they are addressed in a clearly delineated context. Usefulness needs a beneficiary and a concept of usefulness; the notion of functioning needs a

specific mode of functioning as a goal to strive for. While natural sciences understand, describe and explain, the goal of technical sciences is to change and create for practical purposes (Vermaas and Kroes et al 2011 p. 56).

How can we use this special position of technology in the business of reasoning to better understand the emergence of meaning and authority in data science processes?

Larry Hickman has developed an interpretation of John Dewey's pragmatic notion of technology that provides a very helpful platform (Hickman 1990). Dewey's starting point is as well that technology solves problems, as opposed to representing reality (Hickman 1990, p. 106). For Hickman's Dewey, research is a productive skill whose outcome is knowledge. Knowledge is thereby a means of finding one's way in the world and being able to respond appropriately. Therefore, the goal is not unconditional certainty, but a reliable method of solving problems or bringing about results. Success monitoring aims at usefulness and problem solving, not at matching with a reality to be mapped. This pragmatic approach operates on many levels: Meaning, for Dewey, emerges through application. This applies to assumptions, but also to linguistic entities: "objects in a language get their meanings 'in and by conjoint community of functional use' (and not, for example, because they 'picture' reality). But functional use is an ongoing project" (Hickman 1990, p. 40). Thus, the appropriateness of meaning is decided by use and by reactions within the community. Because of this flexibility, meanings can also be adapted, but the decision is up to the community of users. To be consistent with the method of scientific inquiry, this decision is the product of a specific procedure. Science and technology provide rules within which useful decisions can be made; the specific purpose for which the decision is made or what is considered useful is not decided on the basis of these rules (Hickman 1990, p. 161). Rather, feedback processes that improve the quality of the rules and their interpretation are decisive for this and are an indicator of the extent to which purposes and actions are connected according to appropriate criteria. When these conditions are met, the system works. Goals are achieved, the appropriate behavior contributes to being able to move successfully in the world, the actor encounters fewer inexplicable obstacles or confusing information that cannot be classified.

These ideas can also be read as if they were written for the analysis of unsupervised machine learning processes such as in neural networks. Processes run according to defined rules, they deliver first results, which can be adjusted by changing various parameters. The concrete processes in the hidden layers of the neural network can be disregarded as well as the question of what the processes actually represent. The machine learning process does not represent reality, it is looking for a certain perspective on reality. Adjustments in the controlling

parameters therefore do not react to gaps that arise in comparison to a reality to be mapped, but to purposes and goals that are to be achieved.

This becomes clear in current discussions on large language models: These basics of so-called generative artificial intelligence analyze texts and learn to reproduce statistical frequencies. The results are average values that can reproduce what has been read by the model most frequently. This applies to both content and linguistic form. Artificial intelligence often fails on fancy expressions or on content that argues against widespread consensus (especially in languages other than English, since most Large Language Models have been trained predominantly in English). Since artificial intelligence does not evaluate or judge, all errors will be adopted if only they are found frequently enough. It will be equally difficult for artificial intelligence when, for example, ethical issues are touched upon. As easily measurable criteria, frequency and efficiency are the most relevant parameters according to which large language models optimize themselves.

Developers counter this with the concept of reinforcement learning from human feedback (Christiano and Leike et al 2023). Large language models are not only controlled by the factual correctness of their results, but also by feedback on the social or ethical appropriateness of their results. For example, generative artificial intelligence should learn, among other things, not to formulate insults or not to be harnessed for research used to plan crimes.

This approach also concretizes and limits purposes. The scope of the results is reduced, and their desired quality is optimized. Underlying this are decisions about what is considered desirable, who determines it, and how the outcome can be controlled. Such concretizations are often portrayed as after-the-fact constraints that alter an unaffected reality through arbitrary selection. The image of reality thus distorted leads to the idea that it is possible to approach it and that it would only be prudent to dispense with concretization.

Dewey's pragmatic perspective on technology is a helpful proposition that suggests that this concretization is necessary. Without it, we have no concrete question and no way to assess and improve results. Since technical processes do not represent reality, but rather make decisions and thus tend to shape reality, these decisions also do not follow presuppositionless objectifiable choices. For Hickman's Dewey, insisting on the notion of necessity that these decisions and concretizations must follow is mere superstition (Hickman 1990, p. 117).

Since decisions cannot be dispensed with, it is all the more relevant to make them transparent and to take them consciously. In the case of machine learning, this is a sometimes difficult task that is not complete without technical details. In the case of providing Open Data

publications, however, this is often simple information that can be provided via metadata. Examples include information on the timeliness of data, on the circumstances of its specific acquisition, or on the extent of the subject area covered by this data collection and what data is missing (details have been discussed in chapter 3).

Thus, technology does not represent, technology decides. Decisions establish specific contexts, which means constraints. Some decisions subsequently exclude other decisions. For Dewey, this is less problematic than the tendency to forget that decisions have been made - sometimes even those to reduce the number of possible further decisions (Hickman 1990, p. 162). Traceable and reconstructable decisions help to move from abstractions and idealizations back to the actual objects of inquiry.

Similar questions occupy modeling discussions. What is for Dewey generally the necessity of a purpose and a concrete context, are for Weisberg inclusion and fidelity rules. These specify the framework within which a model may idealize (Weisberg 2007). Just as necessity is superstition for Dewey, different inclusion and fidelity rules open up possibilities for designing a wide variety of models. Different rules have different effects on the applicability of models designed on their basis. Models created according to rule A may not be interpreted as if they had been created according to rule B. This is formally possible in many cases. In many cases this is formally possible, but it leads to wrong results.

Applied to data questions, this means that metadata or other elements that help to keep the relationship between data and their environment or the genesis of data transparent must also allow conclusions to be drawn about the underlying inclusion and fidelity rules.

Giere clearly expresses the context-dependence and purpose-orientation of model building in a different way. Instead of representation or isomorphism, Giere finds the essential properties that make models models in the intentions of those who consider something as a model. Crucial to suitability as a model is the intention to use something as a model for something else. Specifically, this intention is expressed in processes that Giere calls interpretation and identification. In this process, a specific concept is selected as a model in order to gain knowledge about another specific fragment of reality. (Giere 2010).

While Giere opposes representation and isomorphism as criteria for modeling, he also does not take a pragmatic position that focuses on usefulness or functionality as criteria for success. In his analysis, despite the emphasis on intention, there is an additional requirement for a certain fit. For Giere, this fit fulfills a similar function for models as truth decisions do for linguistic entities (Giere 2010, p. 273).

How can models demonstrate this fit? Does fit save the detour via purpose and context? So can models with fit regain more and more context-independent meaning via some structural aligments?
Giere's solution approach is not particularly convincing for this specific question, but it leads right into the discussion of how models or data acquire meaning.

5.5 Create meaning from manipulating data

In order to be able to defend a concept of fit without requirements such as isomorphism or representation, Giere looks for ways in which meaning emerges without being derived from presuppositions. He finds hints for this in Michael Tomasello's concept of usage-based linguistics, which tries to explain the emergence of linguistic meaning from interaction (Giere 2010, p. 275). The central concern here is the renunciation of the assumption of an underlying universal grammar or of deduction from superordinate rules. Usage based linguistics builds on the skills of "intention reading" and "pattern finding". Intention reading clarifies to what extent agents interacting with each other here act within a common meaning horizon and whether they also want to interact with each other. Pattern finding identifies regularities, repetitions and contexts from which it can be reconstructed what a certain utterance or action might mean.
Similar concepts to explain the emergence of meaning can be found again and again. The applied scale changes from the emergence of language in general to the formation of concepts in children to correction loops in machine learning to fundamental analyses of information and meaning. Glasersfeld developed radical constructivism from his examination of Piaget's developmental psychology of the formation of concepts in children and also introduces criteria similar to Gieres fit: "The function of cognition is adaptive, in the biological sense of the word, and aims at fit or viability; cognition serves to organize the subject's world of experience and not to 'know' an objective ontological reality" (Glasersfeld 1997, p. 96). Radical constructivism builds an entire theory of knowledge on the idea of concept formation and understanding through adaptation. Interesting in terms of data and modeling are Glasersfeld's observations that knowledge is not only an outcome, but also an activity, and that Radical Constructivism is not a metaphysical hypothesis that can be judged as true or false, "but a conceptual tool whose value is measured only by its success in use" (Glasersfeld 1997, p. 55). This also brings pragmatist echoes into play. Giere opts for a less pragmatist turn

in his search for fit, but there are some more accounts of representation and modeling that require activity, intentionality and rather presentation than re-presentatio; to address problems of fit in pragmatist contexts seems promising.

In his analyses of representation, Suárez places great emphasis on interaction and intention, rather than on structural properties of the representative. Suárez rejects substantive theories of representation, yet not everything can be used as a model for everything. Two essential conditions are intentionality and the possibility of deriving knowledge or arguments from the representative. Suarez calls this surrogative reasoning. Thus this kind of inference is not as arbitrary as, for example, linguistic reference, so it cannot be entirely presuppositionless just by convention or usage. Intention and directionality are aspects that emanate primarily from the acting agents; the possibility of inferring surrogative reasoning from the representative involves the agent, the representative, and the target system. Intention and directionality, that is, the agent's activities, are still prerequisites, but they are not sufficient. Issues of appropriateness are co-determinants of whether the representational relationship works. (Suárez 2004)

Suárez' concept contains strong pragmatic components: Representative powers, which Suárez does not describe in detail, the term stands for the interplay of intention, directionality and functioning surrogative reasoning, exist only if they are used. Without intention as the first stage, there is no representation, so the representation relation has to be created. It is not something that can simply be found. And it works only in the application.

An open question for many of these concepts remains how "success in use", "fit", or "surrogative reasoning" can be established as success criteria.

If one considers data as something different from the target system being analyzed, then the question arises as to how results from data processes can be translated from the numerical-logical form in which they usually exist into concrete instructions for action.

This is the declared goal of data science processes, regardless of where the concrete decision-making or the subsequent action is set. They should create information, even wisdom. "Wisdom is acting on knowledge in an appropriate way" (Kelleher and Tierney 2018, p. 56). It is noteworthy that criteria such as relevance or truth do not play a role in this pyramid for quite a long time. Only on the last level of wisdom a trait appears via the idea of appropriateness, which can be interpreted as a truth criterion in a pragmatic context. Appropriateness is more a matter of effiency than of correspondence. What is the relationship between information and appropriateness?

In his theses on Philosophy of Information, Floridi develops a concept of information in which appropriateness is an essential element: "So information is not about representing the world: it is rather a means to model it in such a way as to make sense of it and withstand its impact" (Floridi 2011, p. ix). This is a very pragmatic approach that locates information in the sense of concrete meaning in the interaction in networks.

Floridi's analysis of symbol or data grounding is a non-representationalist approach that relies on interaction instead of mappings. While mappings as reference givers face the problem of ultimately never coming to an end - there is always another reference, which again needs another justification - interactions as meaning givers get by without this ever further regression. Interactions can be continued ever further and in the course of this change meaning. Openness in this direction is not disturbing.

Problematic is the incompleteness of the genesis of meaning in the direction of the past. Floridi's criterion is the so-called z-condition. Z stands for zero and the z-condition is met when something has meaning without having magically inherited it, without drawing it from pre-existing semantic structures, and without reproducing semantic structures learned elsewhere (Floridi 2011 p. 134). A hitherto meaningless object (which may not have been an object in its own right because, in the absence of meaning, it was not yet sufficiently delimited to be perceived as an object in its own right) acquires meaning as we interact with it and other objects. Floridi calls this kind of genesis of meaning action based semantics (Floridi 2011 p. 165).

For responses to emerge from interactions, concrete questions are necessary; levels of abstraction decide which differences, questions, and interactions are useful and thus have meaning. For Floridi, there is nothing but information in this matter ("it is information all the way up and down" Floridi writes, alluding to the turtles that carry the world (Floridi 2011, p.40)). In the interplay of purposes, networks, and interaction, however, this is not a problem - meaning becomes evident nonetheless.

In none of these approaches have mappings, correspondence, or other structural understandings of representation played a role. Purpose, context and interaction are essential to the authority and validity of data projects; they determine the emergence of meaning and knowledge.

This is another indication that theses of hypothesis-free research and modeling or of purely inductive knowledge production through data fall short. Without the framework of concrete applications and purposes, in which at least a minimal version of hypotheses is contained, data do not find concrete meaning. They remain arbitrarily interpretable indicators.

The decision as to whether data can still establish this reference after the analysis phase is already made in the first phases of collection and processing. Consistency, traceability, metadata, transparently applied standards, and interoperability are some of the techniques that ensure these framework conditions.

Including purpose and context in the consideration of data processes has the effect of softening the technical-mathematical promise of data, which puts neutral, unadulterated precision on the line. Concepts such as data journeys (Leonelli 2020) or data assemblages (Kitchin 2014) attempt to counteract this.

These approaches insist on seeing data in a broader context; a key aspect of this is practicality. This sometimes gives the impression that data journeys and data assemblages are trying to bring back into the debate exactly those aspects that data science actually supposedly rightly wants to get rid of.

However, the attempt to get rid of context and purpose orientation reinforces the problem of how meaning and authority should be attributed to the results of data science. Points of connection are helpful in this regard.

This is illustrated by concepts that deal with practical engagement with models. The artefactual approach suggests that models should not be viewed in terms of representational properties, but should be viewed as artefacts in their own right. Knowledge should not have to be translated, knowledge is gained by working with models (Knuuttila 2021). Nevertheless, even in this very practice-oriented approach, models cannot be created arbitrarily; models are subject to constraints in their creation, which result from the target system and the research question. The question of the constitution of meaning thus arises only as a so-called internal representation, through which linguistic meaning is conveyed in a narrower sense. The external representation is a question of the concrete work with the model or the data, thus more presentation than re-presentation. The decisive arguments do not come from one side or the other; “distributed justification” (Knuuttila 2021, p.64) as a main way of reasoning involves all aspects.

5.6 No sense in data without context and purpose

Information as a relational problem in networks is very close to the kind of information that databases provide: data are stored in elementary components, their relationships are clearly defined. Rules define which relationships are valid, which are technically possible but

meaningless, and which are technically impossible to implement. The technical impossibility does not necessarily say anything about the meaningfulness of this restriction. It is quite possible that technical restrictions prevent meaningful or even factually necessary relationships. In these cases, a redesign of the database must be considered, but often there are underlying problems with the data collection. Often, common key elements are missing, through which relationships can be established in the first place (this is what we have seen in the examples described in chapter 3). Databases provide more information than simple lists only if relationships can be established via common key elements that were previously not visible or not so precisely calculable. In relational databases, very clear structures must be created for this purpose. New database concepts such as graph databases promise schema free modeling and thus the possibility of making relationships visible even if common key elements are missing; the relationship itself thus becomes the key element. Graph databases are particularly suitable for questions in which relationships exist but cannot be easily traced because of multiple intermediaries. In investigative research such as the Panama Papers, graph databases were used to structure and visualize relationships between companies, banks, law firms, politicians and authorities. At the data level, relationships exist between two sides, such as a bank and a law firm. At the data level, there may be another relationship between a law firm and a politician. New information would be the potential relationship between politician and bank, which can be suspected, for example, above certain thresholds in the frequency of the two relationships, supplemented by other criteria such as temporal proximity. Depending on selected level of abstraction, this information about the newly visible relationship receives different meanings, which can range from a mere diagnosis of concurrency to a concrete suspicion of corruption.
Different data management systems handle elementary relationships in different ways, but the rules in the network are always relevant. The rules are partly defined by technical criteria, and partly a matter of content. The technical aspect is sometimes seen as fixed, as if there were only one way to handle data. New data management models (such as just Graph or Nosql databases), but also other perspectives in data issues, for example from the field of Digital Humanities, are softening this dogma (Schreibman and Unsworth 2004).

Working with data creates new knowledge. Data science processes have the potential to bring forth new aspects that would not have been accessible without technical support. Data science methods are tools that can create new perspectives.

However, deriving knowledge from data and Data science methods must not lose touch with their fundamentals, including their technical nature. Promises of better decisions or hypothesis-free research are baseless proclamations if they cannot establish the connection to the concrete question and if their scope is not transparent.
Under these conditions, information and knowledge emerge; without these foundations, results are mere assertions. Such assertions can make sense in technical-deterministic perspectives, when data as technical processes are seen as something that has immediate influence on us. In a technical determinist perspective, it is also possible to take the position that data science leads to better decisions in a simple and direct way.
However, in my opinion, this thesis is not tenable for several reasons.
I have shown that decisions in data processes take place on many different levels. Data itself is the result of production processes, even before it is prepared in such a way that it can be stored in databases and analyzed with computer science methods. These transformations begin with the basic question and lead via the question of what should be regarded as data for the specific case to the question of the form in which results should be made accessible so that they can be analyzed.
Simplified perspectives on data science often only focus on the analysis process: Here, exciting models and algorithms are used to produce results. However, analyses can only operate within the framework that a preceding phase of data collection and preparation has delineated.
The emphasis on these two phases and their mutually dependent contraints limits many of the promises of data science as a new hyperproductive method.
One limitation concerns the promise of better decisions: The scope is limited by the decisions made in the first phase.
A second constraint concerns the expectation of hypothesis-free research: Again, a minimum set of hypotheses is necessary to collect and process data. Even if analysts do not formulate their own hypotheses, their work builds on the hypotheses of others.
A third limitation concerns the options for action: Even if data science itself would lead to better decisions with its result alone, decisions still have to be accepted and set in motion.
Neither data science (nor machine learning or artificial intelligence) trigger fatalistic deterministic processes that cannot be controlled. This is also indicated by the fact that technologies used, such as database systems, data modeling techniques, and querying languages, are changing.

Finally, a fourth limitation arises from tradeoffs that are necessary despite all the efficient automation and digitization. Data science can work with large amounts of data and formulate seemingly precise answers with very broad scopes. However, these answers are usually either very general in nature, they only become more specific with limited scope. Examples include sociodemographic analyses that can classify very well in general terms (men/women, old/young) but can only provide operationalizable answers to specific questions.
These limitations can be efficiently addressed by emphasizing the context and purpose of data science processes. I see these two features, which in simplified perspectives may just be the biggest limitations that computational methods seek to overcome, as prerequisites for giving concrete meaning and scope to the results of data science processes.
Context and purpose as central categories of technical inquiry also help circumvent the problem of imagining a neutral technology (or neutral data) that is biased by inappropriate questions. Bias does not falsify data after the fact; bias is inherent in the core of any question even before data collection and data preparation, and it can serve the positive function of providing for more concrete results. Unaddressed bias remains a problem (and ill-conceived questions in analysis can still additionally lead to misleading results).
Finally, context and purpose as specifying frameworks are useful guidelines for resisting data science overreach. In political or popular discourses, "data" is often heralded as a future problem solver; data will provide answers to all questions (an example of this I discussed in Chapter 3.4 with concept of promissory data). It is not a fault of data projects, it is not a consequence of poorly set up technology, not to be able to do that. Pressing issues such as the recent Corona pandemic have shown how quickly science as a method can become overwhelmed when concrete policy recommendations are required.
Science provides concrete answers to concrete questions; the question "What should we do?" is not a concrete scientific question. Data science can answer more questions faster - that is the real advantage of computational methods. However, meaningful information emerges from this only if these questions are related to each other for a specific purpose.

Around data science it is often discussed whether the core goal of data science would be explanations or predictions. Predictions are often considered to be the real kingpin, making explanations superfluous anyway. The prioritization of predictions is further amplified by the fact that many predictions are also formulated about the future developments and effects of data science and machine learning - instead of explaining the fundamentals and modes of action of these methods.

The focus on predictions, like the emphasis on better decisions, promotes a determinism that, in my opinion, cannot be justified. Above all, this determinism sometimes disguises itself as technical determinism - but the fact that technology and its environment are very closely interrelated can be observed very clearly, especially in the case of data science and machine learning. The determinism can therefore only be a social and political one.
This leads me to a final consequence: data science cannot deliver on many promises. Data science projects are most successful when they go beyond statistics and computer science to include pragmatic aspects. Data science and the follow-up topics machine learning and artificial intelligence are core topics that need to be analyzed with a pragmatic focus. Concrete technical knowledge is only a fraction of what is required to explain these processes. Data science is the basis of a series of new methods and tools that will bring new insights, perhaps also highlight new anomalies and help to overturn world views (Ihde 1990). For this, we need to see them in purpose and context.

A number of decisions have to be made in the process. Technical or mathematical-statistical factors play a subordinate role, in contrast to usual and frequent assumptions about data science. They do not create neutrality or more independent decisions. If there are compelling arguments, they arise from constraints that have an impact on what can be considered a datum.
Such constraints are particularly evident in Hoeyer (Ch. 3.2); the concept of promissory data also makes clear that and how data create their own universe with rules of meaning and validity and why this is necessary.
This world- and condition-building consists of a series of choices.
What this ultimately means is that what appears to another as a bias-free goal, to another as abusive falsification, is instead a different data project with a different universe created under different conditions that it can meet for specific purposes. For others, however, it cannot.
Bias, better decisions, and statistics are of little use when discussing meaning and authority in data processes. I see the greatest potential to provide meaningful answers around data science and machine learning in pragmatic approaches that focus on interaction in networks.

References

Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired*. https://www.wired.com/2008/06/pb-theory/

Becker, J., Schütte, R., & Rosemann, M. (1995). Grundsätze ordnungsmäßiger Modellierung. *Business and Information Systems Engineering*, *37*(5), 435–445. https://doi.org/10.1007/978-3-663-10233-5_2

Blank, M. (2019). *Open Data Maturity Report 2019* (European Commission, Ed.). 10.2830/073835

Christiano, P., & Leike, J. (2023). *Deep reinforcement learning from human preferences*. https://doi.org/10.48550/arXiv.1706.03741

Coeckelbergh, M. (2020). *Introduction to Philosophy of Technology*. Oxford University Press.

Douglas, H. (2000). Inductive Risk and Values in Science. *Philosophy of Science*, *67*(4), 559–579.

Ess, C. (2009). *Digital Media Ethics*. Polity Press.

Floridi, L. (2011). *The Philosophy of Information*. Oxford University Press.

Floridi, L. (2019). *The Logic of Information. A Theory of Philosophy as Conceptual Design*. Oxford University Press.

Giere, R. (2010). An agent-based conception of models and scientific representation. *Synthese*, *172*, 269–281. https://doi.org/10.1007/s11229-009-9506-z

Gitelman, L. (Ed.). (2013). *Raw Data Is An Oxymoron*. MIT Press.

Glasersfeld, E. von. (1997). *Radikaler Konstruktivismus*. Suhrkamp.

Gold, M., & Klein, L. (Eds.). (2016). *Debates in the Digital Humanities*. University of Minnesota Press. https://dhdebates.gc.cuny.edu/read/untitled/section/dfaeaa1b-e682-4b0b-bbd6-3955db0db3be

Gray, J. (2014). *Towards a Genealogy of Open Data*. Conference Paper, General Conference of the European Consortium for Political Research in Glasgow. https://ssrn.com/abstract=2605828

Haig, B. D. (2020). Big Data Science: A Philosophy of Science Perspective. In S. E. Woo, L. Tay, & R. Proctor (Eds.), *Big data in psychological research*. American Psychological Association. https://doi.org/10.1037/0000193-002

Hao Tong, M., & Gunawi, R. L. (n.d.). *Experiences in Managing the Performance and Reliability of a Large-Scale Genomics Cloud Platform Proceedings of the 2021 USENIX Annual Technical Conference 2021*. https://www.usenix.org/conference/atc21/presentation/tong

Haslhofer, B., & Klas, W. (2010). A Survey of Techniques for Achieving Metadata Interoperability. *ACM Computing Surveys*, *42*(2). https://doi.org/10.1145/1667062.1667064

Hoeyer, K. (2019). Data as promise: Reconfiguring Danish public health through personalized medicine. *Social Studies of Science*, *Vol 49*(Nr 4), 531–555.

Ihde, D. (1990). *Technology and the Lifeworld*. Indiana University Press.

Irizarry, R. (2019). *Introduction to Data Science*. Chapman and Hall.

Kelleher, J. D., & Tierney, B. (2018). *Data Science*. MIT Press.

Kitchin, R. (2014). *The Data Revolution*. Sage.

Knuuttila, T. (2011). Modelling and representing: An artefactual approach to model-based representation. *Studies in History and Philosophy of Science*, *42*, 262–271. https://doi.org/10.1016/j.shpsa.2010.11.034

Knuuttila, T. (2021). Epistemic artifacts and the modal dimension of modeling. *European Journal for Philosophy of Science*, *11*(65). https://doi.org/doi.org/10.1007/s13194-021-00374-5

Leonelli, S. (2020a). Learning from Data Journeys. In S. Leonelli & N. Tempini (Eds.), *Data Journeys in the Sciences*. Springer Open.

Leonelli, S. (2020). *Scientific Research and Big Data* (E. Zalta, Ed.). Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/sum2020/entries/science-big-data/

Leonelli, S., & Ankeny, R. A. (2020). *Model Organisms*. Cambridge University Press.

Leonelli, S., & Boumans, M. (2020). From Dirty Data to Tidy Facts. In S. Leonelli & N. Tempini (Eds.), *Data Journeys in the Sciences*. Springer Open.

Leonelli, S., & Tempini, N. (Eds.). (2020). *Data Journeys in the Sciences*. Springer Open. https://doi.org/10.1007/978-3-030-37177-7

Longino, H. (2020). Data in Transit. In S. Leonelli & N. Tempini (Eds.), *Data Journey in the Sciences*. Springer Open.

Lyon, A. (2016). Data. In P. Humphreys (Ed.), *The Oxford Handbook of Philosophy of Science*. Oxford University Press.

Marenko, B. (2019). Algorithm Magic. Gilbert Simonson and Techno-Animism. In S. Natale & D. Pasulka (Eds.), *Believing in Bits: Digital Media and the Supernatural*. Oxford Academic. https://doi.org/10.1093/oso/9780190949983.001.0001

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

Mayer-Schönberger, V., & Ramge, T. (2017). *Das Digital*. Econ.

Morgan, M. (1999). Learning from Models. In M. Morgan & M. Morrison (Eds.), *Models as Mediators. Perspectives on Natural and Social Science* (pp. 347–388). Cambridge University Press.

Morgan, M., Leonelli, S., & Tempini, N. (2020). The Datum in Context: Measuring Frameworks, Data Series and the Journeys of Individual Datums. In *Data Journeys in the Sciences*. Springer Open.

O'Neill, C. (2014). *Doing Data Science*. O'Reilly.

O'Neill, C. (2017). *Weapons of Math Destruction*. Penguin.

Pietsch, W. (2021). *Big Data*. Cambridge University Press.

Rosenberg, D. (2013). Data before the Fact. In L. Gitelman (Ed.), *Raw Data is an Oxymoron*. MIT Press.

Schiebinger, L. (1989). *The Mind Has No Sex?* Harvard University Press.

Schreibman, S., Siemens, R., & Unsworth, J. (Eds.). (2004). *A Companion to Digital Humanities*. Blackwell Publishing.

Shapin, S., & Schaffer, S. (1985). Seeing and Believing: The Experimental Production of Pneumatic Facts. In *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life* (pp. 22–79). Princeton University Press.

Suárez, M. (2003). Scientific representation: against similarity and isomorphism. *International Studies in the Philosophy of Science*, *17*(3), 225–244. https://doi.org/10.1080/0269859032000169442

Suárez, M. (2004). An Inferential Conception of Scientific Representation. *Philosophy of Science, 71*, 767–779. http://www.jstor.org/stable/10.1086/421415 .

Vermaas, P., Kroes, P., van de Poel, I., Franssen, M., & Houkes, W. (2011). *A Philosophy of Technology. From Technical Artefacts to Sociotechnical Systems*. Morgan & Claypool Publishers.

Vismann, C. (2000). *Akten. Medientechnik und Recht*. Fischer.

Weisberg, M. (2007a). Three Kinds of Idealization. *The Journal of Philosophy*, *104*(12), 639–659. http://jstor.org/stable/20620065

Weisberg, M. (2007b). Who is a Modeler. *British Journal for the Philosophy of Science, 58*, 207–233.

All digital references last accessed last accessed 2023/08/21.