1. Introduction: Ten Theses on Big Data and Computability

Wolfgang Pietsch, Jörg Wernecke

1.1 An experiment in interdisciplinarity

This volume is an experiment. It addresses various aspects of computability in a thematic and methodological breadth as it was once demanded by Francis Bacon at the beginning of modern Western science, but as it is rarely practiced today in a time of ever-increasing scientific specialization. At the beginning of the third millennium, the question of the computability of the physical, the psychological, and the social world arises once again—mainly due to advances in information and communication technology and the concurrent data deluge not only in the sciences but also in many other areas, e.g. in the economy and even our personal lives.

The volume is *interdisciplinary,* it tries to gather a large number of perspectives —from the formal sciences like mathematics, statistics, or computer science to the natural and social sciences to the humanities including philosophy. Furthermore, the volume assumes a *foundational* viewpoint, in that many of the issues that are discussed concern fundamental concepts and methods. Finally, it is *application-oriented* by examining a current issue in the empirical sciences, namely to what extent the data deluge has an impact on scientific concepts and methodology.

These three aspects also characterize the work of the Munich philosopher of science Klaus Mainzer, to whom we dedicate this volume at the occasion of his retirement. The themes of the book reflect quite well some of the key research concerns

W. Pietsch et al. (Hrsg.), *Berechenbarkeit der Welt?*,
DOI 10.1007/978-3-658-12153-2_2

of Klaus Mainzer, who has repeatedly demonstrated a keen sense for uncovering innovative questions at the interface between philosophy and science. For example, he was among the first philosophers of science to reflect critically on developments in complexity research or to recognize the epistemological significance of Big Data.

Mainzer never approaches such issues from a philosophical and epistemological viewpoint only, but in his multifaceted scientific career has always undertaken great efforts for an interdisciplinary exchange. This grounding in scientific practice is a crucial premise that abstract philosophical reflection can be relevant beyond the narrow confines of specialized debates. Relatedly, Mainzer has always been a great communicator not only with other disciplines, but he has in many lectures, interviews, and books also approached a wider public.

If one opens the science pages of major newspapers today, outsiders may get the impression that interdisciplinary research constitutes the norm rather than the exception. This, however, is not the case. The path to academic success is still almost exclusively organized in a disciplinary manner. Besides, the problems of interdisciplinarity can only be understood if acknowledging the existence of various scientific cultures that start from different knowledge bases and in particular have different epistemic interests and value systems.

Despite all these difficulties, there are many cross-sectoral issues that can be addressed only in an interdisciplinary manner, ultimately because relevant knowledge and understanding are spread among different academic fields. A novel view on computability in the age of Big Data is certainly an excellent example for such a cross-sectoral issue.

1.2 Ten theses on computability and data-intensive science

The question of the computability of nature is as old as science itself, as old as human efforts to predict the physical and social environment in order to survive. Nevertheless, the perspective has repeatedly changed over the centuries, often in parallel with significant developments regarding both scientific method and the availability of data. The inductivism of a Francis Bacon as well as the novel focus on experimentation and observation during the Renaissance fostered the breakthrough of modern empirical science liberating itself from the dogmas of scholasticism. The beginnings of modern statistics, to name another example, are closely tied with efforts to cope with the 'Big Data' of the 18th and 19th centuries, when

many administrative institutions and statistical offices were founded resulting in a flood of printed numbers (Hacking 1990, Ch. 4). There are many indications that today we are once again at a turning point where novel information technologies in combination with huge data sets raise the question of the computability of the physical and social world.

As often in the history of science, Big Data as a slogan and leitmotif could prevail because the concept is so ambiguous and malleable that it can be interpreted differently in various fields from computer science to economics. While in the sciences conceptual inaccuracies can occasionally be fruitful, philosophy calls for greater precision. Therefore we must first clarify the meaning that is intended with the terms Big Data and data-intensive science.

In the literature, Big Data is usually characterized with reference to the so-called 3 Vs, i.e. the amount of data (Volume), the data rate (Velocity) and, finally, the heterogeneous structure of the data (Variety) (see Laney 2001). This approach has sometimes been extended to include 5 or even 7 Vs. While such a definition seems adequate when focusing on the technical challenges of large datasets, it proves largely unsuitable for addressing epistemological and methodological issues in connection with data-intensive science. The main problem lies in the exclusive focus on the data and its structure while neglecting the algorithms and methods of data processing. Quite plausibly, these latter are much more important to establish novel epistemological aspects in data-intensive science.

Therefore, let us highlight in the following two methodological aspects of data-intensive science. First, there is the often-invoked assertion that Big Data is supposedly able to map all individuals of a population, what is often summarized as the "N = all" quality of Big Data (e.g. Mayer-Schönberger & Cukier 2013, p. 26-31). If one were to accept this characteristic as part of a definition of Big Data, then the aforementioned objection is mitigated since "N = all" has clear methodological implications. For example, conventional statistical approaches typically start with samples, which should be as representative as possible of a population with respect to a given research question. On the basis of such samples, probabilities can be determined for other individuals in the studied population. Such a methodologically sophisticated process of extrapolation apparently no longer happens in the case of a data-intensive science defined by means of "N = all". Instead, the novel approach seems much simpler resembling a process of looking up in a table of all possibilities what is the case for a particular individual.

Of course, it is unrealistic to capture in a data set all the individuals of a population, for example future cases can hardly be taken into account. It is also implausible that there is no reduction of complexity at all, for example by eliminating completely identical data points. Another undesirable consequence of "N = all" is

that for simple systems a small number of data already covers all possibilities, for example, the data points (switch on, light on), (switch off, light off) already contain the full variation of a system consisting of a switch and a light bulb. Obviously, it is not appropriate to speak of Big Data or data-intensive science in such cases. To work around this problem, one could explicitly restrict the definition to sufficiently complex phenomena, which then necessarily require large amounts of data.

A proposal in the spirit of "N = all", which takes into account the mentioned objections, is:

> *Data-intensive science refers to the systematic collection and analysis of data sets which represent a large part of all possible variations of a complex phenomenon*[1]*, such that the relevant causal structure can be inferred algorithmically without further theoretical background assumptions.*

Thus, data-intensive science is above all an inductivist research program. We will later return to some details of this definition, in particular the relative theory-independence and the focus on causality. Furthermore, substantial similarities between data-intensive science and exploratory experimentation will be pointed out.

The second methodological aspect concerns the automation of the entire scientific process (see for example Leonelli 2012). This has the immediate consequence that science takes place under epistemic premises that fundamentally differ from those of the human cognitive apparatus, particularly as regards storage capacity and processing speed of the data. Thus, scientific knowledge can be generated, which is no longer accessible to the human mind. In particular, a 'machine science' can examine phenomena, which from the perspective of a 'human science' are too complex to make reliable predictions or to establish effective interventions, because these phenomena involve too many variables or because the dependencies between the variables cannot be represented by simple functions. One might thus specify that the phenomena studied in data-intensive science commonly are so complex that automated methods yield more accurate predictions compared with traditional approaches of a 'human science'.

In the following, the topic of computability shall be considered from the perspective of data-intensive science in ten theses. Let us start with the big promise:

> *First thesis: Big Data leads to the increasing predictability of complex phenomena, especially to more reliable short-term predictions.*

1 In this context, one often speaks of variational evidence.

This property follows from the two above-mentioned characteristics that, on the one hand, reliable predictions are feasible if the data reflect a sufficient amount of different states of a complex phenomenon, and that, on the other hand, the change in epistemic conditions, particularly the improved ways of storing and processing data, allows for new approaches to analyze complexity.

Applications in data-intensive science often exhibit the following structure: A data set is given linking a large number of predictor variables with a smaller number of response variables. Using a part of the data set, called the training set, a model of the data is algorithmically developed, which is then validated with the aid of the remaining data, called the test set. For example, the results of an Internet search may be determined using the search history or products may be recommended in an online store based on user profiles and past purchases. The crucial novelty is that due to the automation of data collection and analysis many more variables as well as data points can be taken into account compared with conventional statistical approaches. In this twofold sense, we are thus dealing with Big Data.

Let us now take a brief look at the concept of complexity. First, one might distinguish complexity in the descriptions *(descriptive* complexity) from complexity in the phenomena *(phenomenological* complexity). The best-known example of the first type is the so-called Kolmogorov complexity, according to which complexity is perceived as "the minimal length of a programme by which the object x can be obtained following the programming method S." (Kolmogorov 1983, p. 33) According to Kolmogorov, there is a close relationship between complexity and randomness, as random objects exhibit maximum complexity. By contrast, a small complexity must be attributed to objects that can be derived from a few basic assumptions. Furthermore, such descriptive complexity depends on the specific representation, in particular on the programming method, the assumptions made and the allowed derivation steps.

Herein lies the main difference compared with phenomenological complexity, which is located in the phenomena themselves, and thus should ideally be objective and independent of the chosen representation. In the case of phenomenological complexity it is plausible to introduce a further distinction between *emergent* and *irreducible* complexity.[2] In the first case the complex phenomenon results from a relatively simple system of laws or instructions, i.e. from simple basic phenomena that can be accounted for by only a few variables and well defined dependencies between these variables. The complex dynamics then is emergent, it usually results from nonlinearities, which often imply a strong sensitivity in the system behavior

2 A closely related distinction between compositional and dynamic complexity is developed by Meinard Kuhlmann (2011; see also Mitchell 2008).

with respect to small variations in the initial conditions. Typical systems with emergent complexity are treated in the physics of non-equilibrium and in chaos theory, for example, the double pendulum, simple predator-prey systems, or the Rayleigh-Bénard convection.

In contrast, irreducible complexity is characterized precisely by the fact that such phenomena cannot be reduced to a few equations with a small number of variables and clearly defined dependencies. Rather, one is confronted with a complicated arrangement of a large number of very diverse components interacting according to complicated causal interdependencies. Here, dynamic complexity is not emergent in the sense that it can be derived from an underlying simple model. Rather, the system itself is irreducibly complex. Of course, all the characteristics of emergent complexity may still occur such as nonlinearities, feedback loops, or the lack of additivity. But crucially, the representation of the phenomenon cannot be reduced to a few simple assumptions. Relatedly, in the case of irreducible complexity one can no longer introduce a strictly hierarchical structure of various levels of coarse-grained description, as it exists for example in physics: quarks, elementary particles, atoms, molecules, etc. Various interactions between these levels are possible, ultimately leading to a de facto dissolution of the description levels.

It is particularly this latter form of complexity for which data-intensive science constitutes a great advantage. After all, in order to properly represent irreducibly complex dependencies, which cannot be fully grasped by the human cognitive apparatus, one necessarily requires a large amount of data. From what we know today, such complex phenomena occur in almost all application-oriented sciences, especially the social sciences, medicine, or biology. By contrast, phenomena in the theoretical sciences, especially in physics, are usually only complex in the emergent sense, which in part already results from the deliberate focus on a small number of supposedly paradigmatic phenomena (see discussion of the second thesis).

With respect to the subject of the collected volume, a related and just as many-faceted term should also be briefly analyzed, namely computability. In computer science, computability mainly refers to the solvability of mathematical problems. If for a particular field, e.g. arithmetic, general statements or axioms are given, can one determine for each proposition in this field, whether it is true or false? This form of computability addresses the question what can be concluded from a small number of general statements.

Especially for phenomena of irreducible complexity the mentioned conception of computability is not appropriate, precisely because by definition there are no general axioms or fundamental laws. Computability in the context of data-intensive science then is not so much about deriving valid statements from a limited number of universal propositions, but rather about predicting whether something is the

case in the world or not, based on a data set of individual events that is at least in principle arbitrarily extensible. One might again roughly classify these two viewpoints as descriptive computability on the one hand and phenomenological computability on the other. Obviously, in the case of phenomenological computability, the much discussed problem of undecidable propositions no longer occurs, for the simple reason that the interest shifts to propositions which are either realized in the world or not, i.e. which always have a determined truth value. Thus, different concepts of truth are employed in both cases.

Now, the fact that data-intensive science analyzes irreducibly complex phenomena is an important reason that changes in modeling occur (see also Norvig, this volume):

> *Second thesis: A change in modeling occurs from heavily theory-laden approaches with little data to simple models using a lot of data.*

Bearing in mind the distinction between emergent and irreducible complexity, this development is hardly surprising. When dealing with emergent complexity, theoretical background knowledge can fix the basic structure of the dependencies between the individual variables. For example, the relevant equations of the double pendulum can be derived from the fundamental laws of mechanics. Then, a few sufficiently accurate and well-chosen data points will suffice to determine the additional parameters of the model. For example, an ordinary double pendulum has four parameters, the two masses and the lengths of the two arms.

However, in the case of irreducible complexity, more or less by definition, the structure of the equations cannot be determined using theoretical background assumptions, but must instead be inferred from the data, if a sufficient amount is available. This leads to the mentioned inductivist and largely theory-free approach of data-intensive science.

The classic example for the change in modeling as depicted in the second thesis stems from linguistics (Halevy et al. 2009). From a much simplified perspective, two distinct approaches exist in the field of machine translation, one rule-based, the other data-driven. The first proceeds by modeling the different languages in terms of all their grammatical rules. A translation program must then recognize by means of these rules the grammatical structure of a sentence in the source language, then using an ordinary dictionary translate each word of the sentence into the target language, and ultimately implement the sentence into the grammatical structure of the target language. However, this historically prior approach quickly reached its limits. The main reason for its failure presumably lies in the (irreducible) complexity of language. There are just too many grammatical rules and again too many

exceptions to these rules as that it would be possible to manually implement all of them into a computer program.

Somewhat surprisingly, a data-driven approach turned out much more successful, which completely refrains from an explicit modeling of the grammatical structure. Instead, it relies on large text corpora, derived for example from the Google Books project or from the Internet. The predetermined model structure is now far less complex than in the rule-based approach. Instead of modeling a large number of grammatical rules, a simple Bayesian algorithm is used that determines the most probable translation based on the relative frequencies of word sequences in the different corpora.[3] Machine translation still exhibits serious weaknesses, but the data-driven approach makes it possible today to at least grasp the basic ideas of a text in a foreign language. Thus, in the example one can observe just as alleged in the second thesis that simple model structures with a lot of data are sometimes much more successful than complex modeling approaches with little data.

This change in modeling can be further illustrated by means of a distinction between phenomenological and theoretical science that can be found with a number of authors, e.g. Pierre Duhem (1954) or Nancy Cartwright (1983). Pertinent examples for a phenomenological approach are the engineering sciences, while physics, on the other hand, is a classic theoretical science.

Differences exist on almost all epistemological levels. For example, while the laws in phenomenological sciences are causal and contextual, theoretical sciences mainly aim at abstract and universal relationships (see also thesis six below). Thus, the engineering sciences, to use the aforementioned example, are interested in diverse and highly specialized causal knowledge to solve technical problems, while physics mainly wants to discover and explore a small number of fundamental laws of nature[4], from Newton's axioms to the fundamental laws of quantum mechanics. Relatedly, phenomenological sciences investigate a much greater breadth of phenomena compared with theoretical sciences. The engineering sciences for example create an enormous variety of artifacts that are to fulfill their respective purposes under an extremely diverse range of conditions. By contrast, physics is interested mostly in a very limited range of exemplary and paradigmatic phenomena such as

3 In essence, the model structure is as follows: $Pr(e|f) = \text{argmax}_e\ Pr(e)\ Pr(f|e)$, where Pr(e) denotes the relative frequency of a word sequence e in the target language and Pr(f|e) the relative frequency of the word sequence f in the source language, given that the word sequence e can be found at the corresponding position in the target language (cf. 'The Unreasonable Effectiveness of Data', talk given by Peter Norvig at UBC, 23.9.2010, http://www.youtube.com/watch?v=yvDCzhbjYWs at 38:00).

4 These laws are no longer in a strong sense causal, a discussion of this point, however, would lead us too far astray.

the pendulum, the inclined plane or, in more recent times, particle accelerators. Apparently, an underlying assumption is that such paradigmatic phenomena demonstrate the relationships in nature in a pristine manner and thereby help uncover unifying principles of nature. The expectation is that the world can be understood in terms of these unifying principles. Thus, to a strong degree, theoretical sciences aim at explanation and unification, while phenomenological sciences mainly want to provide reliable predictions and justify successful interventions in the world (see thesis eight below). In other words, the epistemic aims differ substantially.

With respect to this distinction, data-intensive science chiefly remains on the phenomenological level, as can clearly be seen in the example of machine translation. After all, the data-driven approach in machine translation dispenses with the explicit modeling of grammatical rules, which could otherwise be used for explanations, and instead focuses on the 'prediction' of useful translations. In the employed text corpora, linguistic practice is recorded in its full breadth instead of focusing on examples that might illustrate particularly well specific grammatical rules. Finally, the aim of the data-driven approach is not to establish a small number of universally valid grammatical rules, but to find rules that in very specific contexts provide the correct translation.

To summarize, in the example of machine translation an inductive approach proved much more successful than a hypothetico-deductive approach based on theoretical background knowledge and complex modeling assumptions. This is insofar remarkable as inductivist approaches were mostly frowned upon in the 20th century—not least in statistics. To this day, many statisticians consider any statistical approach as necessarily involving sophisticated modeling, thereby excluding from the outset many inductive methods, for example from machine learning. This leads us to the third thesis:

> *Third thesis: Conventional statistics is only partly equipped to deal with the data deluge, novel inductive methodology is necessary.*

This statement is well illustrated by a much-discussed essay of Leo Breiman "Statistical Modeling: The Two Cultures" (2001; cp. also Norvig, this volume), in which he distinguishes a culture of data modeling from a culture of algorithmic modeling. The first denotes the conventional approach in statistics, in which informed by background theory a parametric stochastic model is postulated usually depicting an explicit functional relationship between predictor and response variables. Subsequently, it is examined how well this model matches the data, e.g. whether the deviations from the postulated function satisfy a normal distribution. Linear regression methods are a typical example of such a data modeling approach.

By contrast, according to the algorithmic modeling approach, which was developed mainly outside of academic statistics for example by computer scientists or physicists, an algorithm is used to calculate the response variables from the predictor variables. There is no model anymore specifying a clearly defined functional relationship, the parameters of which are to be determined. Rather, the relationship between predictor and response variables remains largely a black box and the quality of the algorithm is measured just by how precise the predictions turn out. Examples of this second type of modeling are decision trees or neural networks. According to Breiman, in 2001 about 98% of all professional statisticians use data modeling, while only 2% employ algorithmic models. These figures have surely changed since then.

A closely related, somewhat older distinction in statistics is between parametric and non-parametric modeling.[5] Parametric models are fixed by a limited number of free parameters, while the structure of non-parametric models is undetermined in the sense that there is no restriction in the number of possible parameters, as is the case for neural networks or decision trees. Obviously, parametric models are closely related with Breiman's data models, both of which rely to a considerable extent on theoretical background knowledge, while non-parametric models are quite similar to algorithmic models, both proceeding in a strongly exploratory and inductive manner.

Several authors have pointed out that non-parametric modeling only became scientifically interesting and promising with the advent of the first computers. In fact, this type of modeling is almost always data-intensive, because the model structure itself is indeterminate to a considerable extent and has to be inferred from the data. Consequently, there are also different requirements regarding the data. For the determination of a few parameters in parametric modeling usually a relatively small number of precise data points suffice, while in non-parametric approaches the quantity of data becomes more important than the precision of the individual data points. In summary, then, if irreducibly complex phenomena are to be modeled on the basis of large data sets, one almost always deals with algorithmic and non-parametric modeling, as is the case for example in data-driven machine translation.

Apparently, such non-parametric models no longer can be represented in terms of coupled systems of equations, the mathematical paradigm that has dominated science for centuries. Particularly in physics, partial differential equations reflected

5 A query to the Google Books Ngram Viewer shows that terms such as "parametric statistics" and "parametric modeling" have been used only since the 50s and 60s of the last century, exactly since the time when the first computers entered scientific research.

the dynamics of the system variables, for example, Hamilton's equations in classical mechanics or the Schrödinger equation in quantum mechanics. Obviously, such coupled systems of equations are always parametric models. It follows:

Fourth thesis: New types of formal representation are required for modeling irreducibly complex phenomena.

In particular, one increasingly finds discrete, fully scalable modeling structures, especially networks (e.g. decision trees, neural networks, or Bayesian networks). These models are often so extensive that they can no longer be manually displayed. Thus, the medium for representation changes as well. The sheet of paper, the book or the human brain are increasingly being replaced by the computer, which alone can develop these models.

Central to the two previous theses is the distinction between theory-driven and exploratory approaches in statistics. Remarkably, there exists an analogous distinction with respect to experimentation, which leads us to the next thesis:

Fifth thesis: Strong analogies exist between exploratory experimentation and data-intensive science.

First, let us delineate the important distinction between exploratory and theory-driven experimentation, which, somewhat astonishingly, has been explicitly introduced into the philosophy of science literature only a few years ago independently by Richard Burian and Friedrich Steinle in 1997. This late moment is surprising especially because the distinction concerns the most basic categorization of experimental research that is conceivable.

Nevertheless, both viewpoints on the role of experiments can already be found in the works of earlier authors, while the juxtaposition is lacking. Karl Popper, for example, characterizes experiments exclusively as theory-driven: "the theoretician puts certain definite questions to the experimenter, and the latter, by his experiments, tries to elicit a decisive answer to these questions, and to no others." (Popper 2002, p. 89) A pertinent example of such a theory-driven experiment is Arthur Eddington's expedition in 1919 to the volcanic island Principe off the West African coast in order to verify during a solar eclipse the deflection of light rays in a gravitational field as predicted by Einstein's general theory of relativity.

An apt description of exploratory experiments, by contrast, can be found in Ernst Mach's "Knowledge and Error": "What we can learn from an experiment resides wholly and solely in the dependence or independence of the elements or conditions of a phenomenon. By arbitrarily varying a certain group of elements or

a single one, other elements will vary too or perhaps remain unchanged. The basic method of experiment is the method of variation." (Mach 1976, p. 149) Such variable variation constitutes the classical procedure in most laboratory experiments. For example, when Wilhelm Röntgen in the late 19th century discovered a new type of radiation, for him, in contrast to Eddington, an elaborate theory was not available. In order to explore the phenomenon, he changed systematically all the variables that he considered potentially relevant, examining their respective impact on the radiation.

Thus, a crucial feature to distinguish these two types of experimentation concerns the theoretical integration. Theory-driven experiments compare the predictions of an elaborate and detailed theoretical framework with experience, while exploratory experiments aim at developing such a theory in the first place or at least discovering the most important phenomenological laws. In other words, in the first case there is a considerable theory-ladenness in experimental practice, in contrast to the second case, where a theoretical understanding is largely unavailable. Relatedly, in exploratory experimentation, the central concepts still have to be developed together with the causal dependencies while the scientific terminology is determined by the theoretical framework in theory-driven experimentation.

There are some striking similarities between exploratory experimentation and data-intensive science. Most importantly, a substantial theory-independence characterizes the two scientific practices.[6] Also, both share the logic of variable variation, that the impact of changing conditions or circumstances on the phenomenon is examined. Finally, the ultimate aim of this procedure is in both cases to determine the causal structure of the examined phenomena.

There are some differences as well. From an epistemological perspective, it is perhaps most remarkable that data-intensive science usually does not start from experimentally obtained data, but rather works with observational data. Another crucial issue concerns the complexity of the phenomena that are being studied. Those examined in data-intensive science usually are so complicated and context-dependent that they do not fit into a controlled laboratory environment.

While it is generally accepted that exploratory experimentation aims at causal knowledge, in the case of data-intensive science the alleged central role for causality stands in contradiction with the relevant literature. For example, Chris Anderson wrote in a much-quoted article in the technology magazine WIRED that in

6 Referring to the relative theory-independence of exploratory experimentation, Ian Hacking could claim that experiments have a life of their own (Hacking 1983). Similarly, one could say that data-intensive science also has a life of its own, in that it is largely independent if more theory-laden scientific practices.

the wake of Big Data causality is replaced by correlation (2008). Similarly, Viktor Mayer-Schönberger and Kenneth Cukier claim that Big Data implies "a move away from the age-old search for causality" (2013, p. 14. See also Ch. 4). The opposite is the case:

> *Sixth thesis: Causality is the central concept to understand why data-intensive approaches can be scientifically relevant, in particular establish reliable predictions or allow for effective interventions.*

The argument for this thesis is quite simple and can be illustrated with the following quotation by the British philosopher of science Nancy Cartwright: "causal laws cannot be done away with, for they are needed to ground the distinction between effective strategies and ineffective ones." (Cartwright 1979, p. 420) A mere correlation, even if it is empirically well documented, e.g. between the stork population and the human birth rate in some regions, cannot be used to influence phenomena. Even the settlement of a large number of storks will hardly cause people to have more children. In comparison, causal relationships can always be used to systematically interact with the world. This holds even if the relationships are only of statistical nature, then the corresponding strategies are of course effective only on average. When people are encouraged to smoke less, the lung cancer rate will very probably decrease in the population.

Now if data-intensive science is about manipulating phenomena then correlations are obviously not sufficient, but there must be a causal connection—for example, if Facebook wants its members to spend as much time as possible on the social network or if Google wants its users to click on certain advertising links.

One might object that sometimes data-intensive science is not about influencing or manipulating phenomena but only concerned with reliable predictions and for this correlations should be sufficient. However, two different types of correlations must be further distinguished, first those that can be attributed to a common cause, and then those which have arisen purely by chance. Correlations can establish reliable predictions only in the former case. Given a barometer reading one can relatively well predict the weather, just because air pressure is a common cause of both variables. However, water levels in Venice cannot help to predict bread prices in the UK even if the corresponding correlation is empirically well established (Sober 2001). Such random correlations occur with high probability whenever a very large number of different variables is considered. It is clear then that in accordance with the sixth thesis reliable predictions must just as well be causally justified, only that there need not be a direct causal link, but an indirect causal relationship via a common cause is enough.

Of course, with these few remarks the problem of causality is not sufficiently dealt with. After all it is one of the most controversial concepts in scientific methodology, to which one can hardly do justice in a brief introduction. One finds in history and still in current science both extremes: prominent voices that completely reject causation as an incoherent concept, but also the opposite view that scientific knowledge must always be of causal nature. This conflict can obviously only be resolved if it is further specified what exactly one means by causality—leading us to the next thesis.

Seventh thesis: The conceptual core of causality in data-intensive science consists in difference-making that a change in circumstances produces a change in the examined phenomena.

In fact, most of the familiar conceptual analyses of causality seem inappropriate for the context of data-intensive science. For example, the currently popular interventionist approaches (e.g. Woodward 2003) rely on a strong notion of intervention which is difficult to reconcile with the fact that in data-intensive science causal relationships are often established based on observational data, i.e. data collected in the absence of explicit interventions.[7]

Mechanistic approaches or the related process theories (e.g. Salmon 1984), another large group of interpretations, presuppose that the connection between cause and effect can be reconstructed in terms of fundamental laws, for example by tracing the electric current moving from a light switch to the bulb according to Maxwell's equations. But that is just not possible in the case of data-intensive science. After all, the mentioned theory-independence of data-intensive science implies that deeper reasons for the connection between cause and effect are commonly unknown.

Finally, with respect to a third large class of interpretations, so-called regularity theories of causality, these are difficult to reconcile with the variational evidence commonly employed in data-intensive science. As pointed out at the beginning of the section, data-intensive science is not so much concerned with the observation of regularities, in which a condition B is always followed by phenomenon A, but above all with the observation of variation, how a phenomenon is influenced by the change of circumstances or conditions.

7 Even Woodward's relatively weak construal of interventions assumes for example that interventions eliminate alternative causal paths between the predictor and the response variables. Thus, an ontologically hard distinction remains between experimental and observational data that is totally unsuitable for the context of data-intensive science.

One should discuss here in much greater detail the ramified philosophical debate on causation, which for lack of space is not possible. Instead, I want to briefly argue that the notion of difference-making is crucial to understanding the role of causality in data-intensive science. After all, it is generally examined which conditions make a difference for a phenomenon, just as required in John Stuart Mill's famous formulation of the method of the difference: "If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance save one in common, that one occurring only in the former; the circumstance in which alone the two instances differ is the effect, or the cause, or an indispensable part of the cause, of the phenomenon." (Mill 1886, p. 256)

Both the method of difference as well as related counterfactual approaches to causality exhibit a variety of problems, which we cannot discuss here due to lack of space.[8] We can nevertheless conclude that the method of difference meets the criteria that were discussed above. First of all, it does not presuppose a strong notion of intervention. At least in principle, then, causal relationships should be determinable on the basis of observational data. Also, a deeper theoretical understanding of the relationship between cause and effect is not required for the method of difference. Ideally, causal relations can be identified without knowledge of the underlying causal processes if as many conditions as possible can be held fix. Finally, the method of difference does not rely on evidence in terms of regularities but rather in terms of variation. After all, it examines the influence of a change in circumstances.

In the literature, the idea that correlation replaces causation is often connected with another characteristic of data-intensive science that it supposedly cannot explain the phenomena. Without causation no explanations, thus the common logic. However, this conclusion holds only under certain construals of causality. For example, in the case of mechanistic causality it is plausible that the underlying mechanism may also explain the corresponding causal relationship. By contrast, if causality is understood in terms of difference-making, as suggested here, then causal relationships are generally not in a strong sense explanatory. This leads to the eighth thesis:

Eighth thesis: Data-intensive science can explain by specifying causes, but not by referring to unifying principles.

8 See for example the manuscript on causation as difference making by one of the authors: http://philsci-archive.pitt.edu/11913/.

This thesis has a lot to do with the aforementioned distinction between phenomenological and theoretical science. In this context, Nancy Cartwright has introduced two types of explanation that one might term causal explanation on the one hand and unifying explanation on the other: "there are two quite different kinds of things we do when we explain a phenomenon in physics. First, we describe its causes. Second, we fit the phenomenon into a theoretical frame." (Cartwright 1983, p. 16) Obviously, the second type of explanation remains reserved largely for theoretical sciences, while the first type can also be found in phenomenological sciences.

As stated before, data-intensive science is confined primarily to the phenomenological level, it determines causal dependencies but refrains from searching for unifying principles. Again the main reason is that data-intensive science usually deals with irreducibly complex phenomena for which such principles do not exist. Obviously then, explanations that refer to unifying principles are impossible in data-intensive science, even though such unifying explanation is regarded by many as a hallmark of science. This applies particularly to physics. A satisfactory explanation of the tides, for example, should ultimately refer to general laws of mechanics and gravitation.

In phenomenological sciences, only a much weaker form of explanation can be realized that refers to causes in the sense of difference-making. Thus, one can explain the occurrence of the tides by referring to the positions of the moon and sun. Such explanations can establish why certain predictions are successful, but they almost entirely lack a unifying nature. In particular, they can hardly be used for guidance to understand related phenomena. A possible link between a stone falling to earth and the tides could hardly be established in this manner.

The advent of data-intensive science thus seems to have the effect that various disciplines such as medicine or the social sciences, which were often seen as theoretical sciences in analogy with the fundamental natural sciences, are now increasingly perceived as phenomenological sciences. In particular for irreducibly complex phenomena in these fields, scientific explanations in the conventional sense are not feasible, causal manipulation and prediction however remain possible. And since in the future, only the computer and less the human scientist can "understand" these complex phenomena, the following holds:

> *Ninth thesis: The increasing automation of science changes fundamentally the role of scientific experts.*

We must accept that due to the limitations of the human cognitive apparatus many scientific processes that take place in the computer are no longer traceable for us humans. A good example is the proof of the four-color theorem, in which only a

computer could check all the remaining variants. In some areas the resulting changes for the role of human experts are already under way, when, as mentioned, grammatical knowledge is no longer relevant for writing translation programs or when reading recommendations of online book shops are no longer written by cultural journalists, but rather are algorithmically generated, which at least from a financial perspective turns out extremely successful.

These changes and the consequences for the labor market exemplify the social impact of data-intensive science. In fact, the ethical dimension of the current developments reaches far beyond the ubiquitous debate on privacy. The latter is indisputably important, but the improved capabilities for predicting complex phenomena, for example in the social sciences or in medicine, as well as the use of data-driven predictions for developing new products such as automated driving lead to a large variety of mostly unanswered ethical questions.

It should however be noted:

Tenth thesis: The ethics and the epistemology of data-intensive science can hardly be separated.

Most ethical problems related to data-intensive science can only be addressed if the epistemological framework is already understood. In other words, the epistemology determines the scope of what is possible, within the confines of which one may deliberate what is ethically desired.

1.3 Overview of the volume

In line with the topics that were just discussed, the volume is divided into four sections: "Big Data and the Sciences", "Computability", "Complexity and Information" and "Ethical and Political Perspectives".

The first part begins with an essay by Google research director Peter Norvig that was originally published as a blog post on his website. For one of the editors (Wolfgang Pietsch), this essay played a crucial role in understanding the epistemological significance of data-intensive science. The subsequent contribution by the Copenhagen philosophers of technology Gernot Rieder and Judith Simon also examines several epistemological issues related to Big Data, for example, theory-independence or the supposed objectivity. They question the thesis of a new empiricism and highlight some political and social consequences of the use of big data.

The third article by the Cottbus philosopher of technology Klaus Kornwachs also takes on a critical perspective with respect to some standard epistemological claims in connection with big data, arguing for the virtues and advantages of traditional scientific modeling.

The following articles focus on selected epistemological aspects of data-intensive science. The Texan mathematician and computational biologist Edward Dougherty argues that even in the age of big data a pressing problem in many areas remains the lack of data when it comes to modeling complex stochastic and non-linear phenomena, such as cell regulation. The statisticians Anne-Laure Boulesteix, Roman Hornung, and Willi Sauerbrei then discuss in their article how much room for interpretation remains in the analysis of large data sets. In particular, they compare different measures how so-called "fishing for significance" can be avoided. In the last contribution of the first part, the Munich mathematicians Nadine Gissibl, Claudia Klüppelberg, and Johanna Mager investigate the significance of large data sets for the analysis of extreme risks, such as plane crashes.

The second part of the volume considers aspects of computability from various disciplinary perspectives. It begins with an essay, in which the Russian philosopher Andrei Rodin discusses whether spatial aspects should also be taken into account in models of computation besides the temporal ordering. Next is an article by the biophysicist Leo van Hemmen on the feasibility of mathematical modeling in complex biological systems, especially in neurobiology. The physicist Konrad Kleinknecht then asks to what extent climate modeling can yield reliable predictions in view of abundant non-linearities. The science TV-host, physicist, and natural philosopher Harald Lesch considers the relationship between physics and metaphysics concentrating in particular on the limits of physical knowledge, for example that the ultimate beginning of the universe is not computable.

The focus then shifts from the natural sciences towards the social sciences. Chadwick Wang, a Chinese researcher in science and technology studies, portrays a controversial debate among various factions of Chinese scientists, whether society is computable or not. The Stuttgart physicist and pioneer of sociophysics Wolfgang Weidlich provides a brief overview of his approach and discusses some important methodological problems. Finally, the Augsburg philosopher of science Theodor Leiber outlines a SWOT analysis of new methodological approaches in computational social science and Big Data.

The Salzburg psychologist Günter Schiepek who like Wolfgang Weidlich belongs to the broader circle of the Stuttgart school of synergetics, examines in his essay issues of complexity, computability, and the use of large data sets in psychotherapy. The Paderborn philosopher Ruth Hagengruber then discusses differences

between human and machine knowledge raising the question to what extent creativity can ever be attributed to machines. The second part concludes with a historical perspective by the Munich philosopher of science Tobias Jung, who pursues the question how important aspects of computability are in the work of Immanuel Kant. Similar to Harald Lesch, Jung focuses primarily on the limits of human cognition and knowledge.

The third part is devoted to the great antagonist of computability, the notion of complexity. First, the Moscow philosopher and complexity-theorist Helena Knyazeva gives an overview of basic principles and aspects of complex systems emphasizing the interdisciplinary nature of any complexity science. The Darmstadt physicist and philosopher Jan Schmidt presents largely ignored aspects in the epistemological work of Pierre Duhem that make him appear as one of the pioneers of modern complexity research. The philosopher of technology Alfred Nordmann, also from Darmstadt, compares two very different approaches by Stuart Kauffman and Brian Goodwin on how biology changes on its way to becoming a science of complexity. Subsequently, the Oxford computer scientist and philosopher Hector Zenil examines the extent to which the theory of algorithmic information can help to analyze complex systems. Finally, the physicist and philosopher Holger Lyre discusses whether the concept of information should play a fundamental role in the natural sciences at all.

In the fourth and last part, the view widens towards ethical and political perspectives. Julian Nida-Rümelin, the Munich philosopher and former German Minister of State for Culture and the Media analyzes whether we need to rethink the concept of responsibility in light of autonomous robots. In the following contribution, the Konstanz philosopher Jürgen Mittelstrass examines the concept of emergence, which of course is closely intertwined with complexity. He shows in particular how emergence can contribute to the creation of novelty at the interface of various disciplines. The business ethicist Christoph Lütge outlines the research agenda of the new field of experimental ethics and raises the question to what extent ethical problems can be empirically examined and calculated. The philosopher and computer scientist Sabine Thürmel then analyzes the interplay of autonomy and control in Big Data based systems. A contribution by the physicist and energy researcher Thomas Hamacher follows that is also methodologically innovative by taking recourse to literature studies. It asks, whether the analysis of a contemporary novel can foster insights regarding energy and sustainability. The volume concludes with a study by the Japanese social philosopher Naoshi Yamawaki how an unpredictable and uncalculable disaster like the nuclear accident of Fukushima may require rethinking the role of ethics for the sciences.

References

Anderson, Chris. 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *WIRED Magazine* 16/07. http://www.wired.com/science/discoveries/magazine/16-07/pb_theory.

Breiman, Leo. 2001. Statistical Modeling: The Two Cultures. *Statistical Science* 16 (3): 199-231.

Burian, Richard. 1997. Exploratory Experimentation and the Role of Histochemical Techniques in the Work of Jean Brachet, 1938-1952. *History and Philosophy of the Life Sciences* 19: 27-45.

Cartwright , Nancy. 1979. Causal Laws and Effective Strategies. Noûs 13 (4): 419-437.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie.* Oxford: Oxford University Press.

Duhem, Pierre. 1954. *The Aim and Structure of Physical Theory.* Princeton, PA: Princeton University Press.

Hacking, Ian. 1983. *Representing and Intervening.* Cambridge: Cambridge University Press.

Hacking, Ian. 1990. *The Taming of Chance.* Cambridge: Cambridge University Press.

Halevy, Alon, Peter Norvig und Fernando Pereira. 2009. The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems* 24(2): 8-12.

Kolmogorov, Andrey N. 1983. Combinatorial foundations of information theory and the calculus of probabilities. *Russian Mathematical Surveys* 38 (4): 29-40.

Kuhlmann, Meinard. 2011. Mechanisms in Dynamically Complex Systems. In *Causality in the Sciences*, hrsg. P. McKay Illari, F. Russo, and J. Williamson, 880-906. Oxford: Oxford University Press.

Laney, Doug. 2001. 3D Data Management: Controlling Data Volume, Velocity, and Variety. Research Report. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Leonelli, Sabina. 2012. Introduction: Making sense of data-driven research in the biological and biomedical sciences. *Studies in History and Philosophy of Biological and Biomedical Sciences* 43 (1): 1-3.

Mach, Ernst. 1976. *Knowledge and Error: Sketches on the Psychology of Enquiry.* Dordrecht: D. Reidel.

Mayer-Schönberger, Viktor & Kenneth Cukier. 2013. *Big Data.* London: John Murray.

Mill, John S. 1886. *System of Logic*. London: Longmans, Green & Co.

Mitchell, Sandra. 2008. *Komplexitäten. Warum wir erst anfangen, die Welt zu verstehen.* Frankfurt a.M.: Suhrkamp.

Pietsch, Wolfgang. 2015. Aspects of Theory-Ladenness in Data-Intensive Science. *Philosophy of Science* 82 (5): 905-916.

Pietsch, Wolfgang. 2016. The Causal Nature of Modeling with Big Data. *Philosophy & Technology* 29 (2): 137-171.

Popper, Karl. 2002. *The Logic of Scientific Discovery.* London: Routledge Classics.

Salmon, Wesley. 1984. *Scientific Explanation and the Causal Structure of the World.* Princeton: Princeton University Press.

Sober, Elliott. 2001. Venetian Sea Levels, British Bread Prices, and the Principle of the Common Cause. *British Journal for the Philosophy of Science* 52: 331-346.

Steinle, Friedrich. 1997. Entering New Fields: Exploratory Uses of Experimentation. *Philosophy of Science* 64: S65-S74.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation.* Oxford: Oxford University Press.

http://www.springer.com/978-3-658-12152-5

Berechenbarkeit der Welt?
Philosophie und Wissenschaft im Zeitalter von Big Data
Pietsch, W.; Wernecke, J.; Ott, M. (Hrsg.)
2017, XII, 562 S. 24 Abb., Softcover
ISBN: 978-3-658-12152-5