The Causal Nature of Modeling with Big Data[1]

Wolfgang Pietsch, pietsch@cvl-a.tum.de

Abstract

I argue for the causal character of modeling in data-intensive science, contrary to wide-spread claims that big data is only concerned with the search for correlations. After discussing the concept of data-intensive science and introducing two examples as illustration, several algorithms are examined. It is shown how they are able to identify causal relevance on the basis of eliminative induction and a related difference-making account of causation. I then situate data-intensive modeling within a broader framework of an epistemology of scientific knowledge. In particular, it is shown to lack a pronounced hierarchical, nested structure. The significance of the transition to such 'horizontal' modeling is underlined by the concurrent emergence of novel inductive methodology in statistics such as non-parametric statistics. Data-intensive modeling is well equipped to deal with various aspects of causal complexity arising especially in the higher-level and applied sciences.

[1] Forthcoming in *Philosophy & Technology* (DOI: 10.1007/s13347-015-0202-2)

1. Introduction

For some time, computer scientists have been speaking of data-intensive science as a 'fourth paradigm' in scientific research, in addition to—as they say—theory, experiment, and simulation. The classic statement is by Jim Gray, a Turing award winner and former employee of Microsoft Research. In one of his last talks before he went missing at sea in 2007, Gray declared: 'The world of science has changed, and there is no question about this. The new model is for the data to be captured by instruments or generated by simulations before being processed by software and for the resulting information or knowledge to be stored in computers. Scientists only get to look at their data fairly late in this pipeline. The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science as a new, fourth paradigm for scientific exploration.' (Gray 2007, xix) The talk was transcribed and resulted in a collected volume titled *The Fourth Paradigm* (Hey et al. 2009).

Of course, from a philosophical perspective, the terminology is somewhat misleading as data-intensive science is certainly not a paradigm in the sense coined by Thomas Kuhn, but rather concerns the emergence of novel scientific methodology that presumably does not replace, but complements existing approaches. Furthermore, it is at best an oversimplification to put data-intensive science and computer simulations on the same methodological level as theory and experiment. But such conceptual inaccuracies should not obscure the fact that the abundance of data in various fields of modern science has significant epistemological implications which merit the attention of philosophy of science and technology.

The current debate on big data is laden with philosophy-of-science concepts like explanation, modeling or causation. However, lack of conceptual clarity and rigor has led to considerable confusion regarding the real impact and methodological novelty—for example when debating controversial statements such as that big data allegedly involves a shift from causation to correlation (Mayer-Schönberger & Cukier 2013, Ch. 4) or that it implies 'the end of theory' (Anderson 2008). In a recent article, the editor of this journal highlighted the 'epistemological challenge' presented by big data (Floridi 2012). My essay is certainly not intended as a definite solution to this problem, but rather wants to give a broad outline of some of the relevant developments and thus serve as a starting point for further discussion. More specifically, I examine the nature of modeling in data-intensive science arguing for two interrelated theses. First, data-intensive science essentially aims at identifying causal structure. Second, I situate the causal modeling in data-intensive science within a general epistemology of scientific knowledge and argue on this basis that it lacks much of the hierarchical structure familiar from more conventional scientific modeling.

In Section 2, I briefly outline the types of problems addressed in data-intensive science comparing it with more traditional approaches in statistics. I also propose a definition for data-intensive science with the following two components: (i) first, it requires data covering all configurations of a phenomenon that are relevant with respect to a specific research

question; this is related to the popular claim that big data often implies data on all instances, in short 'N=all'.[2] (ii) Second, in the spirit of the quote by Jim Gray, data-intensive science requires an automation of much of the scientific process. Note that both features are meant to delineate an idealized scientific method and do not directly apply to all scientific fields in which the term big data is currently evoked.

In Section 3, two examples of data-intensive science are sketched. The first, machine translation, shows in an exemplary manner the shift to horizontal modeling that is at the focus of Section 5. The second, microtargeting, stems from the social sciences and illustrates the promise that data-intensive science provides a better grip on the causal structure of complex phenomena leading to more reliable short-term predictions.

In Section 4, I argue for the causal nature of data-intensive modeling, contrary to popular claims that data-intensive science supposedly is interested only in correlations. Employing a difference-making account of causation that is closely linked with eliminative induction in the tradition of Francis Bacon, John Herschel and John Stuart Mill, I show how some widely-used big data algorithms can under certain circumstances identify causal factors. The causal laws determined in data-intensive science often show various aspects of causal complexity that are familiar from methodological studies in the higher-level and applied sciences. For example, such laws specify a large number of conditions under which a phenomenon will occur or they exhibit complicated functional dependencies. The principal reason why data-intensive science turns out quite apt to deal with complexity lies in the mentioned automation of the entire scientific process. After all, the epistemic conditions and epistemic limits of modern information technology are substantially different from those of science as carried out by humans, e.g. in terms of storage capacity and computational abilities.

In Section 5, the modeling in data-intensive science is situated within a general framework of an epistemology of scientific knowledge. It is shown to lack a number of features that are typical for more conventional scientific modeling aiming at an efficient reduction of data and an adequate structuring of knowledge. Data-intensive models (i) have no pronounced hierarchical structure, which implies that (ii) they lack substantial explanatory power. (iii) They rely on few modeling assumptions and (iv) they are quite complex because little of the original data is actually discarded. Data-intensive modeling will be identified as *horizontal* modeling in contrast to the *hierarchical* modeling characteristic for more traditional scientific methodology.[3]

Further evidence for a qualitative change in the nature of modeling is provided by concurrent methodological shifts in statistics. For example, predictions in data-intensive science are often calculated on the basis of the original data and a suitable algorithm without formulating

[2] It is hard to pin down the origin of this phrase, but it is used in several analyses of big data (e.g. Mayer-Schönberger & Cukier 2013, 197 or Kitchin 2014, 1).

[3] As Peter Norvig, research director at Google, writes: 'In complex, messy domains, particularly game-theoretic domains involving unpredictable agents such as human beings, there are no general theories that can be expressed in simple equations like $F = m\,a$ or $E = m\,c^2$. But if you have a dense distribution of data points, it may be appropriate to employ non-parametric density approximation models such as nearest-neighbors or kernel methods rather than parametric models such as low-dimensional linear regression.' (2009) Many ideas elaborated in this essay take inspiration from scattered writings of Norvig.

parametric models, which attempt to summarize the data in terms of a relatively simple equation involving a few parameters, e.g. a linear function or a Gaussian distribution.

Section 6 concludes by summarizing how data-intensive modeling provides new ways to deal with causal complexity.

2. The nature of data-intensive science

2a Defining data-intensive science

The concept of data-intensive science should be distinguished from the notion of big data itself, i.e. from the mere fact that science—and other areas of human endeavor—are today dealing with increasingly massive amounts of data. Big data is usually defined with respect to the pure amount of information or to the technical challenges which it poses in terms of the so-called 'three Vs': Essentially, data is supposed to be big if it is huge in *v*olume, high in *v*elocity and diverse in *v*ariety (Laney 2001; for an extended definition on this basis see Kitchin 2014, 1-2). However, definitions based on the three V's are beset with a number of problems. Most importantly, volume, velocity, and variety are all relational concepts, as Luciano Floridi has pointed out (2012, 435). Thus, the big data of today could easily be the small data of tomorrow depending for example on progress in technological hardware. Just emphasizing the sheer amount of data thus fails to establish any interesting developments with respect to scientific methodology.

Let us therefore shift the focus from big data to data-intensive science, i.e. to the techniques with which large amounts of data are being processed. One should further distinguish methods of data-acquisition, data-storage, and data-analysis. Although the first two are clearly important for data-intensive practices, in this essay, I will concentrate on the third. With this in mind, let me propose two fundamental characteristics for data-intensive science that will later serve to establish novel methodological aspects in the analysis of large data sets:

(i) Data-intensive science requires *data representing all (or at least a substantial part) of those configurations of the examined phenomenon that are relevant with respect to a specific research question*. This first characteristic ensures that the causal structure of a phenomenon can be determined according to eliminative induction.

For complex phenomena, characteristic (i) implies (a) high-dimensional data, i.e. data sets involving many variables or parameters[4], as well as (b) a large number of observations covering a wide range of different configurations. The different variables are all potentially causally relevant to the phenomenon or are at least symptoms or proxies of other variables that might be causally relevant. These variables are more or less directly measureable or should at least be accessible to operationalization. A configuration consists in a specific combination of values for the different variables, and relevant are all those configurations that are required to carry out eliminative induction in a specific research context and to the desired level of precision (cp. Section 4).

[4] The notion of parameter is to be understood here in a non-technical manner and it is used interchangeably with the term variable.

According to characteristic (i), the notion of data-intensive science is relative to the complexity of the examined phenomenon. Sometimes, when considering a simple phenomenon like a system consisting of a light switch and a lamp, 'data-intensive' science may require only a few data points, e.g. in this case two observations: (switch=on, lamp=on); (switch=off, lamp=off). To avoid this unwanted consequence, one could restrict the practice of data-intensive science to the study of causally complex phenomena.

Note that big data can also be defined from this perspective: as data about complex phenomena that satisfies characteristic (i). This would render the notion independent of advances in information technology and would thus solve Floridi's problem of relativity. On the other hand, many huge data sets would not constitute big data anymore, no matter how large they are, if they do not allow for causal modeling based on eliminative induction.

In a way, characteristic (i) is supposed to render more precise the idea that with big data novel approaches to statistical sampling emerge. Several authors have claimed that in data-intensive science the whole population is examined instead of just representative samples, in short 'N=all' (e.g. Mayer-Schönberger & Cukier 2013, 197 or Kitchin 2014, 1). By contrast, characteristic (i) only requires evidence in terms of those instances that constitute different configurations, i.e. different combinations of values for the variables. For example, with respect to the microtargeting discussed in Section 3b, it is not necessary to have data on all voters, but rather data suffices on a smaller number of individuals that are representative of the possible variations that can occur in the population. This illustrates quite well an interesting shift with respect to sampling, from representativeness in terms of relative frequencies to representativeness in terms of possible variations of parameters.[5]

(ii) The second characteristic concerns the *automation of the entire scientific process*, from data capture to data processing to modeling. This allows sidestepping some of the limitations of the human cognitive apparatus but also leads to a loss in human understanding regarding the actual results of data-intensive science, as we will see in Section 5d. Certainly, automation is only part of the picture and technology takes over a large diversity of roles, e.g. by speeding up processes, by linking previously unrelated domains of experience, or by providing visualization tools. However, following Gray, automation of the entire scientific process is deemed crucial in that it brings to bear the epistemic conditions of the information technology such that science can overcome some limits of conventional modeling, especially when dealing with complex phenomena. If humans needed to interfere at a certain step this would create a bottleneck requiring again an overall reduction of the data and a strong simplification of the models.

Note once more that these features are meant to delineate an idealized scientific practice. Only rarely will there be data on all relevant configurations. Consequently, the resulting causal knowledge will only be approximate, compromising predictive reliability. Furthermore, many applications involving large data sets, for example in high-energy physics or in genetics, do not directly fit the notion of data-intensive science as proposed here—mainly since in these

[5] Note that frequency data, i.e. data how often certain configurations occur, will still be required if not all causally relevant variables are known. But a detailed discussion of this issue would lead too far.

domains theoretical considerations and data are deeply intertwined. But the discussion of idealized methodologies is familiar from philosophy of science. For example, purely inductive or purely hypothetico-deductive approaches do not exist in actual scientific practice nor do purely exploratory or purely theory-driven experiments, but working out the details of these methodologies has proven extremely helpful for understanding more complex scientific endeavors.

There is considerable similarity between the two characteristics proposed in this section and the notion of data-intensive science developed by Sabina Leonelli, whose work on data-intensive science is the most elaborate to date in the philosophical literature focusing mainly on the role of big data in biology and more exactly on biomedical data bases as research tools, e.g. for classification (e.g. 2012a, 2012b, 2013). Leonelli is quite hesitant with respect to a universal characterization of data-intensive science given 'the wide range of activities and epistemic goals currently subsumed under this heading' (2012a, 1). Nevertheless, she identifies two central features: one concerns 'automated reasoning', the other 'induction from existing data […] as a crucial form of scientific inference' (ibd.). Leonelli criticizes the first on the grounds that machine science cannot replace human judgment and expertise. With respect to the second feature, she points to difficulties concerning the concept of induction invoking the notorious epistemological debates on this issue. As a consequence, Leonelli stresses the methodological and epistemic complexity of data-intensive science—a perspective that well correlates with her focus on big data in biology, which indeed does not constitute a pertinent example for the idealized practice that I sketch in the present article.

Still, some of Leonelli's worries can be addressed. By examining specific algorithms, this article identifies the type of induction used in data-intensive science as eliminative induction. Furthermore, Pietsch (forthcoming) distinguishes various aspects of theory-ladenness examining whether they occur or are absent in data-intensive science. This provides some indication where human expert knowledge is required and to what extent automation is feasible.[6]

2b The problem structure in data-intensive science

Typical problems in data-intensive science concern classification or regression of an output variable y with respect to a large number of input variables x, also called predictor variables or covariates. Generally, one wants to determine the nature of dependence of the output variable from the input variables on the basis of given data points linking certain values of the input variables with a value of the output variable. The main differences compared with conventional problems in statistics consist in the high-dimensionality of the input variable (sometimes also the output variable) and in the amount of data available about various configurations or states of the system. For example, an internet store wants to know how likely someone buys a certain product depending on surf history, various cookies and a user profile as well as based on data of other users who have either bought or failed to buy the

[6] Pietsch argues that data-intensive science involves external theory-ladenness concerning the framing of a research question, but mostly lacks internal theory-ladenness concerning the causal structure of the examined phenomenon.

product. A medical researcher examines which combinations of genetic and environmental factors are responsible for a certain disease. Or a political adviser is interested how likely a specific individual is going to vote for a certain candidate based on a profile combining for example voting history, political opinions, general demographics, and consumer data.

In a *classification* problem, the output variable can assume a number of discrete values. In a *regression* problem, the output variable is continuous. In order to establish an adequate and reliable model, extensive training and test data is needed. Usually, one starts with a data set comprising a number of instances that each gives a value for the output variable dependent on at least some values for the input variables (so-called supervised learning). The data set is then divided into a training set and a test set. The training data is used to build the model, the test data to validate and verify the model.[7] In this essay, we cannot delve into the technical details of the various algorithms employed in data-intensive science, such as support vector machines, forests or neural networks. I will now however introduce two simple algorithms, classification trees and naïve Bayes. Later on in Section 6, non-parametric modeling will be discussed as a further class of methods requiring large amounts of data.

Classification trees (Russell & Norvig 2010, Ch. 18.3.3) are used to determine whether a certain instance belongs to a particular group A depending on a large number of parameters $C_1, \ldots, C_N$, which can each take on a finite number of discrete values.[8] For example, A could classify an email as spam (A=1) or not (A=0), and the C_i could denote the number of times that certain keywords occur in the email. The tree is set up recursively with the help of training data, e.g. in the example data linking keywords to emails which are known to be spam or not. First, the parameter C_X is determined that contains the largest amount of information with respect to the classification of the training data, as formally measured in terms of Shannon entropy. In the mentioned example, this would be the single keyword that best classifies the emails in the training data, e.g. 'sex' or 'lottery'. If C_X classifies all instances in the training set correctly, the procedure is terminated. Otherwise, several subproblems remain, namely classifying when C_X is present a specific number of times or when it is absent. The procedure is repeated until either all instances are classified correctly or no potential classifiers are left. If the algorithm is successful and the problem is set up correctly, the resulting tree structure gives an expression of necessary and sufficient conditions for A, which can be interpreted as complex causal laws. In Section 4b, we will come back to the problem, under which circumstances such 'laws' really are predictive.

Another simple big-data algorithm is naïve-Bayes classification, which for example is also widely used in the identification of spam emails. The problem structure is the same as in the case of classification trees. A number of parameters $C_1, \ldots, C_N$, representing again certain words or sequences of words appearing in emails, is used to determine the probability that a specific instance is A or not, e.g. that an email is spam or not. Using Bayes' Theorem:

$$P(A|C_1, \ldots, C_N) = [P(A) / P(C_1, \ldots, C_N)] \, \Pi_{i=1,\ldots,N} P(C_i|A)$$

[7] An excellent introductory textbook from a computer-science point of view is Russell & Norvig (2009).
[8] In a pioneering book on machine learning and scientific method, Donald Gillies also used the example of classification trees to argue for the Baconian nature of these novel developments (1996). While Gillies does not discuss causation, the general thrust of his book points in a similar direction as the argument given in Section 4.

The 'naïve' part of the algorithm is that the parameters C_i are assumed to be independent given A, i.e. $P(C_1, \ldots, C_N|A) = \Pi_{i=1,\ldots,N} P(C_i|A)$, which of course may not be the case. As with classification trees, a training set is used to develop the model. It provides representative frequencies for joint occurrences of A and the different C_i and thereby the probabilities $P(C_i|A)$, P(A), and $P(C_1, \ldots, C_N)$. On this basis, new instances can be classified given certain values C_i. Usually, the value of A is chosen that has the highest probability. Again, test instances can be set aside to validate the model.

3. Examples of data-intensive modeling

3a Machine translation

Machine translation belongs to the standard repertoire of big-data success stories. It illustrates particularly well the shift from complex models with relatively scarce data to simple models with a lot of data that will be discussed in Section 5. Although somewhat of an oversimplification, two different approaches can be distinguished (Halevy et al. 2009). The *rule-based* approach models the complex hierarchy of grammatical rules of both languages and translates sentences by using a conventional dictionary. The *data-driven* or statistical approach largely neglects the grammatical structure and works instead with huge corpora of texts in combination with Bayesian inferential statistics. Usually, there will be monolingual corpora, e.g. in English and a foreign language, and bilingual corpora containing sample translations, all of them representative of current speech practice. The frequencies of words in these corpora and of sequences of words, so-called n-grams, can be used to calculate the most probable translation of a foreign word sequence f into English e using Bayes' rule[9]: argmax_e P(e) P(f|e), where the probabilities are evaluated in terms of relative frequencies in the corpora.

The data-driven approach has been strikingly successful. Apparently, probability distributions of words and word sequences yield reasonable results for many tasks such as spellchecking or translation, while grammatical knowledge is largely dispensable. Two quotes from practitioners well illustrate this remarkable situation. Peter Norvig, who for a long time headed Google's machine translation group, once stated that they have been able 'to build models for languages that nobody on the team speaks.'[10] Frederick Jelinek, a pioneering and by now legendary figure in the field, is often quoted with saying that 'every time I fire a linguist, the performance of the speech recognizer goes up.'[11]

Data-driven machine translation fits well the notion of data-intensive science developed in Section 2a. With respect to the first criterion, practitioners have often emphasized that the data-driven approach requires enormous corpora (e.g. Halevy et al. 2009, 9). Already in 2006, Google Translate relied on a trillion-word corpus assembled from the internet with frequency counts of all sequences up to five words (ibd.). Plausibly, such a corpus approximately

[9] Jelinek 2009, 492. Cp. also '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.
[10] Ibd. 43:45.
[11] http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/speechreco/team/, accessed 1.8.2013

represents for some areas of speech a considerable fraction of all possible phrases, i.e. of all relevant configurations. Data-driven machine translation is also a good example for the second criterion of Section 2a regarding the automation of the research process. After all, the corpora are collected online, stored in data bases and algorithmically analyzed without much human intervention and in particular without notable input in terms of theoretical knowledge from linguistics.

While machine translation constitutes a pertinent example for the shift from hierarchical to horizontal modeling as described in Section 5, it is less obvious that it fits the notion of causal modeling, since it obviously does not refer to the physical necessity of empirical causal laws. However, the logic of necessary and sufficient conditions inherent in eliminative induction as depicted in Section 4 works also for the 'conventional necessity' of translation rules. The context of a word, i.e. all other words at various distances, eventually determines a condition for a suitable translation, resulting in an almost infinite number of complex and highly context-specific translation rules. Note that simple and general rules may entirely fail to exist, i.e. more precisely, that general rules have so many and so diverse exceptions that these cannot be fully listed. Under such circumstances, the data-driven horizontal modeling could be the only option available.

Machine translation is a good example to illustrate the shifts in theory-ladenness and the role of expert knowledge occurring in the wake of data-intensive science. As the mentioned quotes by practitioners like Frederick Jelinek or Peter Norvig prove, for certain tasks in linguistics specific kinds of theory and thus of expert knowledge can be dispensed with. An interesting question for the sociology of science concerns the impact of these shifts on scientific practice in a wide array of fields, from medicine to political science.

3b Microtargeting

The second example comes from the social sciences regarding the use of data-intensive methods in American elections, in particular Barack Obama's 2008 and 2012 bids for presidential office (Issenberg 2012). Political campaigning is a typical big data problem as outlined in Section 2b. Voters are characterized in terms of hundreds or thousands of features ranging from demographic data like age, race or gender to political opinions gathered in surveys to consumer data provided for example by credit card companies. Campaign managers are then interested in causal relationships between these predictors and outcome variables like commitment to vote or allegiance to a certain candidate.[12] The approach has been aptly called *microtargeting*.

In the United States, abundant training data exists because citizens are often willing to volunteer information about their voting habits. The resulting models are developed algorithmically with little input of political expert knowledge. They are used to determine the probabilities that certain persons can be convinced to vote for a specific candidate and which means are most appropriate in terms of political message and medium, e.g. contact by mail,

[12] Of course, these variables often do not constitute direct causes, but rather symptoms or proxies of direct causes, as discussed in Section 4b.

telephone or a personal visit. While previously, political campaigns addressed larger groups of people characterized by just a few parameters such as middle-class Caucasian male, microtargeting focuses on individual voters characterized by hundreds or thousands of variables. This allows correcting many implicit assumptions about the alleged relevance of variables like race, gender or class, essentially redrawing the conceptual boundaries between groups on an empirical basis. Indeed, data-intensive science is especially suited for the categorization and classification of phenomena in view of a specific purpose.

Certainly, current data-intensive approaches are far from identifying necessary and sufficient factors such that the voting behavior of specific individuals can be reliably predicted and manipulated. But big data algorithms seem capable of increasing the probability for predictions by taking into account a larger number of plausible predictors than conventional approaches.[13] Such predictive success can only be due to a better grip on the causal relationships of a phenomenon, since after all, reliable predictions require that the algorithms are capable of better approximating necessary and sufficient conditions for a phenomenon. Many applications of big data in the social sciences have a structure that is analogous to the microtargeting in election campaigns. Individuals are characterized in terms of a large number of parameters with a specific aim in mind, e.g. to find an appropriate search result or to make someone click a certain link or buy a certain product. Data-intensive approaches in the social sciences are particularly suited for short-term predictions and manipulations.

Microtargeting well illustrates two developments that are characteristic for the emergence of data-intensive approaches in the social sciences. The first concerns personalization, i.e. the idea that one is not anymore interested in representative samples, but rather tries to model each individual of a population. In fact, the data-intensive modeling used by the Obama campaigns tried to account for every single voter—which broadly corresponds to the idea of 'N=all' evoked in Section 2a. Examples for this development in other areas of scientific research are the promise of personalized medicine that many health professionals see in big data, personalized web search or individualized online advertising.

The other development concerns the shift in the role of expert knowledge that was already mentioned, which directly correlates with the automation of science as described in Section 2a. The conflict between two expert cultures in recent election campaigns is well described in the chapter on 'Geeks vs. Gurus' of Sasha Issenberg's 'Victory Lab: The Secret Science of Winning Campaigns' (2012). The geeks rely on data and statistics with a profound distrust in anything they could not measure, while the gurus are 'the celebrated political wise men' with their little pet theories about which things happened when. It seems that data-intensive science has considerably shifted the balance from gurus to geeks in recent years.

4. The causal nature of data-intensive modeling

In a much-cited and influential article, journalist Chris Anderson, at the time editor in chief of the technology and lifestyle magazine Wired, wrote some controversial remarks how big data

[13] This has been widely reported in the press, e.g. http://www.businessweek.com/articles/2013-05-31/obamas-data-team-totally-schooled-gallup (accessed 5.8.2014)

affects science: ‘Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.’ (2008) The ideas condensed in this brief statement have been readily picked up by mainstream media, but also in academia. Occasional critical reactions have mostly focused on Anderson’s thesis concerning ‘the end of theory’.[14] By contrast, I will now explain why the wide-spread claims about the significance of correlation as opposed to causation are mistaken. Instead, the modeling in data-intensive science is very much of causal nature.

Nancy Cartwright once highlighted as central feature of causation that causal knowledge can ground effective strategies (1983, Ch. 1). A mere correlation cannot tell how to successfully intervene in the world, e.g. the amount of rain cannot be reduced by banning the use of umbrellas, even though there exists a robust correlation between rain and the use of umbrellas. By contrast, headaches can be cured by taking acetylsalicylic acid because, according to current medical knowledge, there is a causal mechanism connecting the two phenomena. Thus, if big data is about making predictions regarding interventions, e.g. about making people vote for a specific candidate or click on a certain ad, then it must aim for causal knowledge and cannot be satisfied only with correlations.

Note that correlations which allow for reliable predictions, but do not enable effective interventions, nevertheless require a causal justification. Consider as an example the correlation between barometer reading and weather, which obviously can be used for prediction, but not for manipulation. Tinkering with the barometer needle will never let the sun come out. Of course, the underlying causal structure consists in a common cause in terms of the air pressure that determines both weather and barometer reading. It seems plausible to assume that it should always be like that. If we can reliably predict certain phenomena on the basis of some other phenomenon, i.e. if the success-rate is higher than expected due to chance, then either one phenomenon is a difference-maker for the other or there must be some common cause that is a difference-maker for both phenomena. All this holds of course independent of the fact whether we are aware of the causal connection or not.

Notwithstanding this simple argument, the phrase ‘correlation supersedes causation’ is ubiquitous in the debate on big data.[15] The confusion essentially results from a conflation of causation with mechanistic explanation. The viewpoint defended in the present article is the following: the notion of causation grounds the important distinction between correlations that allow for successful predictions or manipulations and those that do not. Now, a wide-spread but ultimately mistaken belief holds that causation cannot be read off the data itself, but needs a deeper justification in terms of a mechanistic explanation, how the causes bring about the effects in terms of more fundamental laws. However, recent technical work on causality (e.g. Pearl 2000; Spirtes et al. 2000) as well as conceptual analysis (e.g. Woodward 2003) has

[14] It is quite revealing that Anderson misquotes Google research director Peter Norvig with the statement: ‘All models are wrong, and increasingly you can succeed without them.’ (2008) In a reply on his web page, Norvig clarifies: ‘That’s a silly statement, I didn’t say it, and I disagree with it.’ (2009) Certainly, there will always be modeling assumptions in any scientific endeavor. Norvig’s actual point had concerned changes in the nature of modeling resulting from big data (cp. Section 5).

[15] Compare for example the recent compilation on http://www.forbes.com/sites/gilpress/2013/04/19/big-data-news-roundup-correlation-vs-causation/ accessed 15.6.2013

shown that causal knowledge can be derived without a deeper understanding of any underlying mechanism. In the following, I will argue for the same point on the basis of a difference-making account of causation and will show how this account plays a role in some of the classic big-data algorithms.

4a Difference-making: An appropriate account of causality for data-intensive science

The framing of big-data problems as a mapping of input variables to an outcome variable fits well with eliminative induction[16]—a scientific method whose history reaches back at least to the methodological writings of medieval thinkers like Robert Grosseteste and William of Ockham. The most elaborate frameworks are Francis Bacon's *method of exclusion* (1620/1994, Bk. 2), which arguably was considered the methodological foundation for modern science until the end of the 19th century, William Herschel's methodology as laid out in his *Preliminary Discourse* (1851), and John Stuart Mill's *methods of elimination* (1886, Bk. III, Ch. VIII). Writers in the tradition of eliminative induction have repeatedly stressed the inadequacy of the primitive notion of regularity for causal and inductive reasoning (e.g. Bacon 1620/1994, 21; Mill 1886, 204). In the modern debate, Federica Russo, in particular, has emphasized the importance of variation for causal reasoning as opposed to essentially Humean regularity conceptions (2009; Illari & Russo 2014, Ch. 15-16). In the 20th century, eliminative induction has received little attention presumably due to prevailing anti-inductivist and anti-causalist commitments.[17] In the following, I can only highlight a few features that are crucial for the discussion of data-intensive science. For a more comprehensive overview of the method, compare for example Pietsch (2014).

In eliminative induction, a phenomenon A is examined under the systematic variation of potentially relevant conditions $C_1, \ldots, C_N$ with the aim of establishing *causal relevance* or *irrelevance* of these conditions, relative to a certain context B determined by further conditions. Obviously, the framing corresponds exactly to that of big-data problems as outlined in Section 2b.

The best known and arguably most effective method of eliminative induction is the so-called *method of difference* that establishes causal relevance of a condition C_X by comparing two instances which differ only in C_X and agree in all other circumstances C. If in one instance, both C_X and A are present and in the other both C_X and A are absent, then C_X is causally relevant to A.[18] There is a twin method to the method of difference, that one might call the *strict method of agreement*, which establishes causal irrelevance, if the change in C_X has no influence on A. Eliminative induction can deal with functional dependencies and an extension of the approach to statistical relationships is straightforward.

A method for establishing causal irrelevance is usually not included in accounts of eliminative induction, such as that of Mill or of Mackie. Some authors contemplate but reject inferences

[16] not to be confused with a looser use of the same term in the sense of eliminating hypotheses until only the correct one remains

[17] There are notable exceptions, e.g. Mackie (1980) or Baumgartner & Graßhoff (2004).

[18] A number of further problems arise here, e.g. concerning time direction. Details under which additional premises these inferences are actually valid can be found in Pietsch (2014).

to causal irrelevance. For example, Baumgartner and Graßhoff argue that inferring the irrelevance of a condition in some context overlooks the possibility that it may be causally relevant with respect to other conditions that are currently not instantiated (2004, p. 212). For example, a burning match can be causally relevant to a barn-fire, even though in a concrete instance it makes no difference due to the absence of inflammable material. An account that includes the notion of causal irrelevance and circumvents the described problem by rendering the notion of causal relevance and irrelevance background-dependent is given in Pietsch (2014).

A further conceptual difficulty of the strict method of agreement is that in scientific practice causal irrelevance can be determined only up to a certain degree of measurement accuracy. A change in C may have minute consequences for A that are not detectable in a specific situation. While considerable problems can arise, for example when several small contributions add up to measureable changes, the dependence on measurement accuracy constitutes no reason to reject inferences to causal irrelevance. They are ubiquitous in scientific practice and causal irrelevance can play an important conceptual role in various contexts, for example regarding the evaluation of counterfactuals, as described below.

The methodology of eliminative induction suggests a corresponding account of causation which will be called *difference-making account*. While I essentially agree that one should be careful to separate definition and methodology of causation, the dividing line is nevertheless blurry. Conceptual questions about how to define causation cannot be addressed without at the same time considering epistemological issues how causal relationships can be identified in the world and vice versa. In a way then, eliminative induction and the difference-making account should be considered as two sides of the same coin.

The difference-making account is closely related to the counterfactual approach. In fact, it adopts a counterfactual definition of causation, along the lines of Hume's famous statement: "we may define a cause to be *an object followed by another, and* […] *where, if the first object had not been, the second never had existed*" (Hume 1777, Sec. 7, §60). In the difference-making account, the notions of causal relevance and causal irrelevance are fundamental, thus: *in a context B, in which a condition C and a phenomenon A occur, C is causally relevant (irrelevant) to A, iff the following counterfactual holds: if C had not occurred, A would also not have occurred (if C had not occurred, A would still have occurred).* Note that there is an issue regarding the direction of causation, which for lack of space we cannot address here. That every causal statement is relative to an instance, in which the conditions and the phenomenon actually occur, is meant to reflect the contextuality of causal statements as well as the primacy of singular causal statements over causal regularities. The latter only follow from a specific causal relation if the context and the relevant conditions can be replicated.

According to the perspective of the difference-making account, causal (ir-)relevance is a three-place relation: a condition C is (ir-)relevant to a phenomenon A with respect to a certain background B of further conditions that are held constant if potentially causally relevant to A and that are allowed to vary if causally irrelevant. More exactly, a causal context is determined by a background B of potentially causally relevant conditions that are held constant, while a number of other potentially relevant conditions C_1, …, C_N are allowed to

vary. Certainly, conditions in the background that are relevant to A in virtue of being relevant to one of the Cs, i.e. that lie on causal chains leading through the Cs to A, may vary as well. A good example for such a causal background or context is an experimental set-up under laboratory conditions. The restriction to a context B is required because there is no guarantee that in a different context B*, the causal relation between C and A will continue to hold. Since the relevant conditions in the background can only rarely be made explicit, if at all, causal laws as established by the difference-making account have a distinct *ceteris-paribus* character.

A crucial difference with respect to conventional counterfactual accounts of causation like that of David Lewis concerns the way the counterfactual conditional is evaluated. Lewis, for example, refers to the similarity between the actual and possible worlds, essentially: *'If C were the case, A would be the case' is true, if some C-world where A holds is closer to the actual world than is any C-world where A does not hold.* (Lewis 1973, 560) Here, a C-world is just a possible world in which C holds. For Lewis, the challenge is to find a proper construal of the notion of possible worlds and of the similarity between them.

By contrast, the difference-making approach evaluates causal counterfactuals in terms of causal irrelevance: *'If C were not the case, A would not be the case' is true with respect to an instance in which both C and A occur in a context B, if there exists at least one instance in which neither C nor A occurs in the same context B and if the context guarantees homogeneity.*[19] The context guarantees homogeneity, if only conditions that are causally irrelevant to A can change with the following exceptions: (i) C itself, (ii) conditions that are causally relevant to A in virtue of being causally relevant to C, and (iii) conditions that are causally relevant to A in virtue of C being causally relevant to them. Intuitively speaking, the last two items denote conditions that lie on a causal chain leading through C to A.

The causal counterfactual that occurs in the definition of causal irrelevance is evaluated in a completely analogous manner. Note that this approach covers only a small amount of all conceivable counterfactual statements—a restriction that is intended to be conveyed by using the term 'causal counterfactuals'. For example, propositions like 'if Hilary Clinton had been a man, she would have become president' cannot be evaluated in the above manner since an instance that differs only in terms of irrelevant circumstances cannot plausibly be found in the actual world nor be idealized from other phenomena in the actual world. Furthermore, there is a threat of circularity, since the concepts of a constant context and of homogeneity themselves require the notions of causal irrelevance and of causal relevance. However, this circularity is not vicious, but essentially corresponds to a consistency requirement for all propositions stating causal relevance and irrelevance.

This construal of the truth conditions for counterfactuals is of course directly inspired by the method of difference and the strict method of agreement. One might be tempted to consider it a variant of Lewis's approach that implements a specific measure of similarity. However, this

[19] One should also consider defining the truth-value for counterfactuals if there are no situations, in which C does not occur and that differ only in terms of irrelevant circumstances. For example, it may be the case that C belongs to a complex of conditions that occur only together. In such a case, nothing can be said about the causal relevance of C alone, only about the relevance in conjunction.

is clearly not the case for the following reasons. (i) First of all, the difference-making approach does not compare different possible worlds, but different phenomena or events in the actual world. (ii) Furthermore, the 'similarity measure' of the difference-making approach is not continuous, but a two-valued function: either the instances differ only in terms of irrelevant circumstances or not. (iii) Moreover, there is no universal similarity measure at all. Rather, similarity always depends on C and A, because these appear in the definition of a constant context. For all these reasons, the difference-making approach to counterfactuals cannot be considered a special case of Lewis's account.

In the end, it comes down to a matter of taste, if one prefers to classify the difference-making account as a variant of the counterfactual approach or if one considers the differences crucial enough to merit a proper name. I tend towards the latter. To sum up, there are basically three main differences: the construal of counterfactuals, the inclusion of a notion of causal irrelevance, and the introduction of background dependence. The term 'difference-making' aptly reflects the fact that counterfactuals are evaluated not with respect to possible worlds but with respect to actual situations that either really occur or are idealized from the actual world.

These are the basic building blocks of the difference-making account. All other notions can be construed from the elementary definitions of causal relevance and irrelevance, they all have what one might call a specific causal signature. Consider as an example the notion of a *causal factor* C for a phenomenon A with respect to a background B. The basic idea is that C is not sufficient for A with respect to background B but requires the presence of additional conditions X. For example, a short circuit is not enough to start a barn fire but requires also some inflammable material. In terms of causal relevance and causal irrelevance, a causal factor can be identified in the following way: There exists an X such that C is causally relevant to A with respect to B∧X and irrelevant to ¬A with respect to B∧¬X; X is relevant with respect to B∧C and causally irrelevant to ¬A with respect to B∧¬C.[20]

Similarly, the notion of an *alternative cause* C for A with respect to a background B, i.e. both C and some X are by themselves sufficient for the phenomenon A, can be explicated as follows: There exists an X such that C is causally relevant to A with respect to a background B∧¬X, but causally irrelevant to A with respect to a background B∧X; equally, X is causally relevant to A with respect to a background B∧¬C, and causally irrelevant to A with respect to a background B∧C. As an example, C could be a short circuit and X lightning as possible causes for a fire A.

Based on causal factors and alternative causes, the difference-making account can identify causal laws in terms of necessary and sufficient conditions for a phenomenon, relative to a certain background or context. More specifically, a cause established by this method can be formulated as an *INUS*-condition (Mackie 1965), i.e. an *I*nsufficient, but *N*on-redundant part of an *U*nnecessary but *S*ufficient condition, with the further requirement that these *INUS*-conditions must in general be seen relative to a context. Extensive information about potentially relevant conditions in as many different configurations as possible is necessary to

[20] Note that there are some technical difficulties in defining irrelevance to ¬A, but intuitively the meaning should be clear.

approximate reliable causal knowledge of complex phenomena by means of eliminative induction and the difference-making account. Exactly this kind of information is provided by big data.

We have already compared the difference-making approach with the counterfactual account. Let me now give a brief overview how it relates to other popular approaches to causation. With respect to naïve regularity theories an important distinction is that the difference-making account does not explicate causation in terms of constant conjunction, but rather in terms of the variation of circumstances, for example by taking into account negative instances. Strictly speaking, the mere repetition of an event under the exact same circumstances does not furnish any evidence at all about causal relationships according to the difference-making account.

A further crucial difference concerns the conditions that need to be fulfilled for reliable inferences. As emphasized in Pietsch (2014, Sec. 3f), there is a distinct problem of induction for the difference-making approach in comparison with naïve regularity theories. While the latter rely on an essentially indefensible principle of the uniformity of nature, the difference-making account presupposes a set of premises that seem much more reasonable and realistic, including: (i) determinism; (ii) constancy of background conditions; (iii) adequate causal language. These are certainly challenging to establish, but they can be approximated by relying on evidence in terms of varying conditions.[21]

Note that the difference-making account bears considerable resemblance to more sophisticated regularity theories like those of Mill (1886) or of Mackie (1980). With both, it shares the focus on eliminative induction and just like Mackie's account, it emphasizes the importance of background dependence and of counterfactual reasoning.

The difference-making account is also related to the interventionist approach, in particular there exist considerable similarities between the way it evaluates counterfactuals and how the notion of intervention is construed for example in Jim Woodward's approach (2003). Moreover, the difference-making account shares with the interventionist approach the ambition to evaluate counterfactuals with respect to actual events rather than possible worlds. However, the term intervention conveys the misleading impression that for the identification of causal relationships observational data cannot suffice, but that experimentation or manipulation is somehow required. And indeed, according to Woodward, interventions have some peculiar ontological properties, e.g. they require a distinct intervention variable and they are able to disrupt other causal relationships. This puts a big question mark behind the appropriateness of the interventionist approach for data-intensive science which is more often than not dealing with data sets of observational nature. By contrast, eliminative induction and the difference-making account work equally well for experimental and observational data.[22]

In summary, an account of causation that is appropriate for data-intensive science should fulfill a number of criteria of adequacy. (i) The account should fit the variational nature of

[21] Some preliminary ideas can be found in Pietsch (2014, Sec. 3f), while a lot of the details still have to be worked out.
[22] In a similar vein, Reutlinger (2012) criticizes the notion of intervention in Woodward's approach and argues for eliminating it.

evidence used in data-intensive science, i.e. that one is usually dealing with many instances in terms of different combinations of the high-dimensional predictor variable and the response variable. (ii) The account should not require a strong notion of intervention, since the data in data-intensive science is often of observational nature. (iii) Finally, it should not rely on elaborate theoretical background assumptions, since data-intensive science is usually taken to be relatively theory-free. The first requirement poses problems for the regularity account. The second rules out any interventionist or manipulationist accounts that deserve the name. And the third requirement excludes mechanistic or causal-process accounts, since data-intensive science often aims to establish predictive inferences without any knowledge of underlying processes. By contrast, the difference-making account fares quite well with respect to all these criteria.

4b Difference-making in big-data algorithms

We are now in a position to establish that the big-data algorithms introduced in Section 2b rely on a logic of eliminative induction. Several aspects are important. First, data-intensive methods generally employ evidence of variational nature as required by eliminative induction, i.e. they rely on a large number of different combinations of the predictor and response variables. Also, the perspective of eliminative induction is useful when examining the premises under which big-data algorithms lead to reliable results. Finally, several data-intensive methods explicitly rely on difference-making to single out crucial parameters.

In the case of classification trees, the algorithm determines those parameters that contain the most information, i.e. make the largest difference with respect to classification, measured in terms of the Shannon entropy. More exactly, the algorithm proceeds as follows for a classification problem with respect to an outcome variable A that can take on the values a_1, …, a_n. In the training set these values shall appear with relative frequencies $p(a_1)$, …, $p(a_n)$. The Shannon entropy is then calculated as $H(A) = -\sum_i p(a_i) \log_2 p(a_i)$. Notably, it has the property that it is maximal, when all outcomes are equally likely, and minimal in the case of perfect classification, i.e. if the probability of a single a_x equals one, and the probabilities for all other a's are zero.

One can then introduce a classification taking into account a parameter C_X with possible values x_1, …, x_k. Again, the $p(x_j)$ and $p(a_i \mid x_j)$ can be determined on the basis of relative frequencies in the training set. Accordingly, a conditional Shannon entropy can be calculated: $H(A|C_X) = \sum_j p(x_j) H(A|C_X = x_j) = -\sum_j p(x_j) \sum_i p(a_i|x_j) \log_2 p(a_i|x_j)$. Given the conditional entropy, the so-called information gain is defined as $H(A) - H(A|C_X)$. It is always positive or zero and quantifies a possible improvement of the classification $A|C_X$ with respect to just A. The information gain will be the largest, namely H(A), for a perfect classification, i.e. a classification, where all $p(a_i \mid x_j)$ are either 1 or 0. On this basis, the C_X can be determined that exhibits the largest information gain of all C resulting in a number of subtrees corresponding to the different possible values of the parameter C_X. The whole procedure is then repeated for every subtree and so on.

In the case of a perfect classification, the algorithm yields an expression of necessary and sufficient conditions for each value of A: e.g. iff $(C_1 = y_1 \wedge C_2 = y_2) \vee (C_1 = y_3 \wedge C_3 = y_4)$, then $A = a_1$. If the classification is not perfect, for some values a_i, the conditional probabilities given certain C will be $\neq 1$ and $\neq 0$.

Thus, if a full set of relevant conditions is among the parameters C and if there is enough data in terms of variation of parameters to exclude accidental correlations, then the algorithm will identify the actual INUS conditions—possibly with some redundancies, but redundant variables can easily be identified and eliminated after the algorithm terminates. In the example above, $C_1 = y_1$, $C_2 = y_2$, $C_1 = y_3$, and $C_3 = y_4$ are all INUS-conditions for $A = a_1$. Certainly, these will not always be made explicit by an algorithm, especially if they turn out highly complex involving a large number of variables. But whenever the algorithm yields fairly reliable predictions, the reason must be that it managed to at least approximate some necessary and sufficient conditions for the phenomenon.

There are various ways, in which the evidence may not be ideal. For example, we may lack information in terms of specific configurations or a causally relevant parameter may be missing from the variables C. Furthermore, it may be the case that some parameters are only symptoms or proxies of the actual causally relevant variables, which may be unknown. A symptom here is understood as a condition that is causally related to the actual cause, but need not always co-occur with it. It could be a consequence of the real cause that requires for its instantiation further conditions. Now, if the probability of co-occurrence is large then the classification tree algorithm will yield decent predictions on the basis of proxies. Broadly speaking, the algorithm can be understood as a heuristic generalization of eliminative induction for situations that are less ideal than textbook examples.

With respect to the conceptual analysis of causation in the previous section, the conditions identified by the classification tree algorithm are causes according to the difference-making account in situations of ideal evidence and if the additional premises mentioned towards the end of Section 4a are fulfilled.[23] By contrast, these conditions are in no obvious way causes in terms of Lewis's counterfactual account as the algorithm does not evaluate possible worlds or the similarity between instances, nor are they causes in terms of interventions since there are no obvious intervening variables as required according to Woodward's definition of intervention.

Furthermore, the notion of regularity is not really helpful to understand the logic of classification trees. Rather, it must again be emphasized that the algorithm relies chiefly on evidence in terms of parameter variation while identical instances are often redundant. For

[23] More exactly, the conditions, under which classification trees function successfully, are identified in Pietsch (forthcoming, Sect. 4): "(a) one has to know all parameters C that are potentially relevant for the phenomenon A in a given context determined by the background B; (b) one has to assume that for all collected instances and observations the relevant background conditions remain the same, i.e. a stable context B; (c) one has to have good reasons to expect that the parameters C are formulated in stable causal categories that are adequate for a specific research question; (d) there must be a sufficient number of instances to cover all potentially relevant configurations of the phenomenon. If such theoretical knowledge can be established, then there is enough data to avoid accidental correlations and to map the causal structure of the phenomenon without further internal theoretical assumptions about the phenomenon."

example, if there are three parameters C_1, C_2, and C_3 with C_2 being the actual cause of A, the classification tree algorithm will be at a loss if the data comes as a mere regularity of many instances in which both C_2 and A are present, while not much is known about C_1 and C_3. Rather, to identify the correct cause, it suffices to know one instance for all possible combinations of C_1, C_2, C_3. In summary, the analysis of the classification tree algorithm in terms of causality is best carried out using the difference-making account.

In a similar manner, the naïve-Bayes algorithm can be interpreted in terms of eliminative induction although the connection is a little less obvious. In the interest of space, I will refrain from a detailed analysis. Let me just summarize that for a Bayesian approach without the naïve independence assumption $P(C_1, \ldots, C_N|A) = \Pi_{i=1,\ldots,N} P(C_i|A)$, it can be proven that a correct classification results in the presence of sufficient conditions as well as in the absence of necessary conditions. Other than the classification trees, the Bayesian approach will not explicitly identify causal factors, but if they are present in terms of INUS conditions, a correct classification of novel instances will result.

Complications arise again in situations of less-than-ideal evidence, for example if not all causally relevant factors are known or if only proxies of the actual factors are available. But in those cases it will generally help to have abundant data in terms of parameter variation. The independence assumption of the naïve-Bayes approach strongly simplifies calculations but is not always compatible with an INUS perspective. For example, if $(C_1 \wedge C_2) \vee C_3$ is the complete cause for A, the independence assumption will not hold, since e.g. in general $P(C_1 \wedge C_2 \wedge \neg C_3|A) \neq P(C_1|A)\, P(C_2|A)\, P(\neg C_3|A)$.

Again, eliminative induction provides a useful perspective also for the methodological analysis of the naïve-Bayes algorithm. Obviously, the algorithm relies on evidence in terms of parameter variation as generally required by eliminative induction. The quality of the results depends on the same premises that were pointed out in Section 4a, in particular constancy of the background, appropriate causal language, and sufficient number of instances.

Clearly, difference-makers C among the predictor variables will decisively influence the classification as they correlate substantially with the phenomenon A, i.e. mostly $P(C|A) >> P(C|\neg A)$—even if C is not sufficient, but only a necessary condition for A. By contrast, if C is causally unrelated to A, then we can plausibly assume that A and C are independent $P(C|A) \approx P(C|\neg A) \approx P(C)$. Thus, there is a good chance that the outcome of the naïve-Bayes algorithm is determined by the largest difference-makers or proxies of them and thus a correct classification will result. These few remarks are intended to underline that an analysis of the naïve-Bayes algorithm in terms of a difference-making approach can lead to a better understanding under which circumstances the algorithm will actually be successful.

In contrast to classification trees, the naïve-Bayes approach does not explicitly identify the most relevant parameters. Therefore, it is often combined with a separate procedure of feature selection. For this purpose a large number of different approaches exist and many of them at least implicitly rely on difference-making (e.g. Guyon & Elisseeff 2003). One important example is variable ranking, in which the predictor variables are ranked according to a specific measure and then a threshold is introduced up to which value of the measure the

variables are taken into account. A typical measure is ‘mutual information’ which corresponds to the information gain discussed before, i.e. at least in this case the ranking again relies on the amount of difference-making by the various parameters.[24]

A further conceptual difficulty is that the difference-making account as presented in the previous section constitutes prima facie a deterministic approach to causation. This is particularly problematic since the evidence in data-intensive science only rarely includes all relevant parameters such that phenomena are fully determined by their conditions. One might conclude then that algorithms like naïve Bayes require a probabilistic concept of causation. While it is impossible to fully address this complex issue in the present article, I believe that the difference-making account can be generalized to indeterministic contexts. The basic idea is to reformulate the examined phenomenon on a coarse-grained level of description such that it becomes deterministic. As a simple example, quantum mechanics is deterministic when considering the evolution of the wave-function according to the Schrödinger equation instead of examining single collapse events. Note that this requires a further premise ensuring a stable context: variables which are not fully determined by conditions should be identically and independently distributed as is for example the case for position and momentum in quantum mechanics.

As a third example of a big-data algorithm, the non-parametric regression to be discussed in Section 5c can also be interpreted in terms of eliminative induction. The causal nature of the resulting curve can be justified in terms of the method of concomitant variations which in turn derives its inferential power from the method of difference (Pietsch 2014, Sec. 3d). If the required premises are fulfilled, in particular stable context, stable categories, and sufficient number of instances, then non-parametric regression just implements the method of difference to establish complex functional relationships (Pietsch forthcoming).

In summary, while the algorithms of data-intensive science are less rigorous than pure eliminative induction, at least some of them can be considered as heuristic variants for less-than-ideal situations of evidence. They are optimized for computational feasibility and stand a good chance to yield useful approximations of some causal story behind a phenomenon. It largely remains a challenge for contemporary statistics to work out the mathematical details, under which premises these methods are successful and to what extent. In general, including more parameters C will increase the probability that the actual cause of A might be among them, while admittedly also increasing the probability for accidental correlations, i.e. that conditions produce the correct classification merely by chance. However, more data in terms of instances of different configurations can reduce the probability for such accidental correlations. Thus, more data in terms of parameters and instances will generally increase the probability that correct causal relations are identified by big-data algorithms.

[24] Recently, deep learning techniques have become an immensely popular and successful approach to feature extraction.

4c Big-data laws

Building on work by Nancy Cartwright (1999) among others, the philosopher of biology Sandra Mitchell has recently outlined the enormous challenge posed by complexity in many higher-level sciences (2008). Questioning wide-spread reductionist assumptions, she argues that causal relationships in many higher-level and applied sciences show a number of remarkable features, including: (i) they are complicated, in particular there are usually many contributing factors instead of one dominating cause. Also, there may well be a number of different possible causes for the same phenomenon. Consequently, causal relationships are often strongly context-dependent; (ii) the causal dependencies are frequently non-linear, mostly they do not follow any simple function at all; (iii) causal interaction takes place between different levels of ontology; (iv) the composition of causes does not follow simple addition laws. Mitchell identifies a number of paradigmatic examples mainly from her own field of expertise, the biomedical sciences, including depression and various kinds of cancer.

Accordingly, Mitchell calls for a novel epistemology that is able to deal with these and other aspects of causal complexity (2008, p. 22-23). In general, the naïve regularity accounts that are still commonplace in the sciences will fail, since due to strong contextuality some laws may be instantiated only a small number of times or even only once (cp. again Russo's argument for a variational epistemology, 2009). A certain type of depression or cancer may for example be very specific to a particular individual.

By contrast, the algorithms discussed in the previous Section 4b are well-equipped to address the mentioned aspects of complexity. Obviously, both classification trees and naïve Bayes are able to identify or approximate causal relationships that depend on a large number of parameters, that react very sensitively to all kinds of changes in the parameters, and that may sometimes be instantiated only a small number of times. Relatedly, the composition of causes may be arbitrarily complex. All these complications are not particularly problematic for eliminative induction and a difference-making account of causation as long as the problem is properly framed in terms of constancy of background and an adequate causal language and as long as a sufficient number of instances are known covering at least a considerable fraction of the relevant configurations of the examined system.

Furthermore, eliminative induction and the discussed big-data algorithms are neutral with respect to any supposed levels of ontology. There is no reason, why necessary and sufficient conditions in terms of difference-making could not link different levels of ontology. In fact, the difference-making account grounds a pluralistic and perspectival view on causation, where it is easily possible that different causes are identified depending on the formulation of a problem. Finally, note that both algorithms discussed so far rely on the presence or absence of certain conditions and therefore cannot immediately account for non-linear relationships or complex functional dependencies. However, we will see in Section 5c that data-intensive algorithms based on the logic of eliminative induction are also able to deal with non-linearity and arbitrarily complex functional dependence.[25]

[25] Compare the preliminary discussion about functional dependence in Pietsch (2014, 3d).

The more complex and contextual the causal relationships are, the less plausible it becomes that these relationships will adhere to the old reductionist ideal that the laws of science can be integrated into a *hierarchy of increasing universalit*y. This constitutes one of the major differences between the causal complexity discussed here and the complexity traditionally invoked in physics, in particular in chaos theory. The latter deduces complex phenomena from underlying simple equations, like the logistic equation modeling population growth or the Lorenz equations in meteorology. However, the reduction of complex phenomena to simple laws is only plausible if the relevant ontology can be reduced to a small number of different types of entities, and if the causal structure can be reduced to a small number of laws, which in turn requires that the composition of causes adheres to simple rules. There is no reason apart from metaphysical prejudice why we should expect this to be the case for all phenomena.

Before the advent of big data, the causal structure of complex phenomena was extremely difficult to analyze as it was almost impossible to efficiently collect and handle high-dimensional data. Mostly, scientists worked with dubious simplifications, e.g. that all but a few main influences on a phenomenon could be neglected and that these main influences adhered to simple functional relationships. But these assumptions, which are for example implicit in the structural equation modeling that is ubiquitous in the social sciences, were chiefly motivated not by empirical considerations but merely by the need to make the analysis fit the available scientific toolbox. Certainly, generalized laws can always be formulated, but for the price that such laws exhibit a large number of exceptions which renders them fairly useless beyond basic heuristics. Data-intensive science seems much more suitable than these conventional approaches as a methodology for the causally complex sciences.

4d Data threshold

As Halevy et al. point out, there exists for many phenomena a relatively sudden change when data-driven approaches become effective (2009)—a transition point that could be called a *data threshold*. They provide a plausible explanation for its existence: ‘For many tasks, once we have a billion or so examples, we essentially have a closed set that represents (or at least approximates) what we need, without generative rules.’ (2009, 9) This aspect further corroborates the methodological analysis of the previous sections. The data threshold constitutes the point, at which the data covers a considerable fraction of ‘all configurations that are relevant with respect to a specific research question’ or to a predictive task as required in Section 2a. The deeper justification for this premise was given in Section 4a, namely that it enables to carry out eliminative induction in terms of the method of difference and the strict method of agreement. As the example of machine translation showed, the required number of instances may be huge when dealing with complex phenomena. Beyond the data threshold, the equally large number of resulting laws need not be integrated into a hierarchical structure to make predictions. No abstract or general laws are necessary, which leads us to the notion of horizontal modeling to be described in the next section.

5. Horizontal modeling

5a The role of causal modeling in science

In the following, I will first provide a brief sketch how the causal modeling in data-intensive science fits into a general epistemology of scientific knowledge. I will then discuss some characteristics of this type of modeling and finally present an example from statistics. In the past, several authors have sketched a hierarchy in our knowledge about the world that broadly consists in an observational and a theoretical level. For example, many logical positivists endorsed the distinction. More recently, proponents of the Stanford School, in particular Nancy Cartwright (1983), have argued that the real distinction is not between theory and observation but between a phenomenological and an abstract level in science—a viewpoint that can be traced back at least to Pierre Duhem.[26] While the difficulties of separating the supposedly directly observable from the non-observable are notorious, a broad distinction between the phenomenological and the abstract is more plausible. This does not mean that every statement can be uniquely classified in these terms, but rather that the general distinction is useful to account for scientific practice. It basically concerns the boundary between that part of our knowledge that can be used for efficient interventions and reliable predictions and other parts that primarily serve an adequate structuring of the knowledge, mainly in the interest of what Ernst Mach once called intellectual economy.

Cartwright developed her account broadly in the context of the new experimentalism of the Stanford School. At the phenomenological level, one mostly deals with experimental laws and is interested in the phenomena in their full complexity. These laws are of causal nature, since their main function is to ground reliable predictions and interventions. Eliminative induction based on parameter variation constitutes the primary scientific practice to establish such experimental laws. Note that the notion of a phenomenological or causal level should not be misunderstood in terms of some ontologically fundamental level. Rather, depending on context, the causal level can be macroscopic or microscopic, or it may even include both micro- and macroconcepts.[27] Consequently, it is somewhat of a simplification to speak of a single phenomenological level. Several layers can be included as long as the relationships between these layers are causal and can serve for manipulation and prediction.[28]

By contrast, the theoretical level deals with abstract laws that are universal and mainly serve conceptual and explanatory purposes. One is not interested anymore in the full complexity of

[26] "A physical theory […] is a system of mathematical propositions, deduced from a small number of principles, which aim to represent as simply, as completely, and as exactly as possible a set of experimental laws. […] These principles may be called 'hypotheses' in the etymological sense of the word for they are truly the grounds on which the theory will be built; but they do not claim in any manner to state real relations among the real properties of bodies. These hypotheses may then be formulated in an arbitrary way. […] The various consequences […] drawn from the hypotheses may be translated into as many judgments bearing on the physical properties of the bodies. […] These judgments are compared with the experimental laws which the theory is intended to represent." (Duhem 1954, 19-20)

[27] Thus the causal level can comprise different levels of ontology (cp. Section 4c). One should keep the distinction between these different notions of level in mind.

[28] From the perspective of the difference-making account nothing precludes the possibility that macrovariables cause microvariables or vice versa as long as the various causal relations are consistent with each other. A detailed defense of this point would go beyond the scope of the present article.

the world but often in exemplars, i.e. paradigmatic phenomena. Also, these general laws are not of a pronounced causal nature anymore for the following reasons. First, the function of the theoretical level is not so much on the side of prediction and manipulation, but rather it serves an adequate structuring of knowledge. Also, the theoretical laws are usually not established experimentally by parameter variation, but rather are developed from the phenomenological level in a process of abstraction mostly according to pragmatic criteria like simplicity. It is of course again a simplification to speak of a single theoretical level. Especially in physics there exists a hierarchy of increasingly general theoretical laws, at the top of which stand the fundamental axioms that are often taken to be the core of physical theories.

The two-level approach allows for a compromise between realist and antirealist intuitions.[29] It explains the continuous and increasing empirical success of science by referring to the causal level that remains relatively stable even during major scientific upheavals. Rather, what changes during scientific revolution is not the causal content but the abstract framework, i.e. scientific revolutions concern mainly the theoretical level. The two-level framework can thus account for both the predictive stability and the explanatory instability of science in the course of its history. This already suggests a relative independence of the phenomenological from the theoretical level.

That one can develop the causal level to some degree without making reference to a theoretical level, is of course a central tenet of the new experimentalism—more or less expressed by Ian Hacking's famous slogan that "experimentation has a life of its own" (Hacking 1983, xiii).[30] Accordingly, there should be scientific practices that remain largely on the phenomenological and causal level. An example in this regard concerns exploratory as opposed to theory-driven experimentation (Burian 1997; Steinle 1997, 2005; cf. also Pietsch 2014). The former consists basically in systematic parameter variation and the link to causation can again be established by means of eliminative induction. There is little doubt that exploratory experimentation is a crucial practice in particular in the early stages of the scientific examination of experimentally accessible phenomena. Interesting case studies have been developed in particular by Steinle (2005).

Now, data-intensive science is another scientific practice that is mostly restricted to the phenomenological level. As described in Pietsch (forthcoming, Sec. 6), there are a lot of similarities between this type of research and exploratory experimentation, in particular the restriction to a causal level that is accessed via a methodology of eliminative induction based on parameter variation. There are also some key differences, maybe most importantly that exploratory experimentation is—of course—an experimental practice, while the data in data-intensive science is usually of observational nature. But from the perspective of the difference-making account of causation as sketched in Section 4, this does not constitute a methodological obstacle since this approach works equally well for observational and experimental data, in contrast to accounts that put more emphasis on the ontological

[29] In analogy to John Worrall's "The best of both worlds"-argument for structural realism (1989).
[30] More exactly, Ian Hacking does not explicitly identify experimental knowledge as causal and theoretical knowledge as 'less' causal. This element is introduced by Cartwright, who is generally counted as a proponent of the new experimentalism as well. But Hacking cites Cartwright's approach approvingly and points out the close similarity of their respective antitheoretical stances (1983, Ch. 0).

importance of interventions or manipulations for causation. A related difference is that data-intensive science often deals with phenomena that are much more complex than those accessible to exploratory experimentation.

As a consequence, the two practices, very broadly speaking, dominate different domains. While exploratory experimentation is especially useful in sciences with a relatively simple fundamental ontology and a pronounced hierarchy of theoretical levels, such as physics, data-intensive science is often employed in complex sciences that do not allow for significant levels of generalized theoretical laws. Of course, this is not to deny that there are examples of exploratory experimentation in the complex sciences or that data-intensive science has applications in physics. To the contrary, there are important data-intensive practices for example in particle physics or astrophysics, even though some of these may not fall fully under the definition given in Section 2a, mainly due to the considerable theory-ladenness of these practices.

An important question concerns the scope of this epistemological picture. When Cartwright developed her views, she was chiefly inspired by physics which was her main research focus at the time. Duhem, of course, also wanted to provide an epistemology for physics. By contrast, most success stories of data-intensive science come from other areas, for example biology or the social sciences. Thus, it is an important question whether the two-level framework sketched above applies to those areas as well. My position on this issue is that all sciences have a more or less well developed phenomenological or causal level, as long as part of their epistemological enterprise is to derive predictions based on a practice of difference-making that can range from simple experiments in a laboratory to more sophisticated practices like randomized control trials or quasi-experiments.

It is sometimes argued that the causes in complex sciences like the social sciences are of a different nature compared with physics. While in the latter, one deals chiefly with causal regularities, in the former these may often fail to exist. However, eliminative induction building on counterfactuals can identify causal structure without relying on causal regularities, while these can subsequently be derived under the premise that the relevant conditions reoccur. Thus, an absence of strict regularities does not imply the absence of causal structure according to the difference-making account of causation.

The differences between the sciences are more pronounced with respect to the theoretical level. For some fields, the development of an elaborate theoretical structure is an important aim, the prime example being again physics, where one finds a complex hierarchy of theoretical concepts of increasing generality that is built on top of the phenomenological relations. Certainly, physics draws its impressive explanatory power from the unificatory virtues of this hierarchical superstructure. In other sciences, past attempts to build a similar theoretical structure have largely failed. A case in point are the social sciences, in which 19th century dreams of a social physics based on a small number of axioms never materialized. Social knowledge apparently resembles rather a patchwork of more or less isolated areas of interrelated causal laws.

Of course, the main reason, why there are these differences with respect to the theoretical level has to do with the complexity of the respective phenomena. Physics has the advantage that its ontology can be reduced to a small number of basic entities and the history of physics suggests that it is also possible to formulate a small number of fundamental laws for those entities that can then be aggregated in straight-forward ways to account for composite physical phenomena of restricted complexity. Of course, there are no a priori reasons why this should be the case for other fields as well. In sciences that do not fulfill the mentioned requirements concerning ontology and laws, it is implausible that a pronounced hierarchical structure can be developed.

The epistemological framework presented in this section motivates the distinction between horizontal and hierarchical modeling that will now be elaborated. Again, horizontal modeling remains on the causal and phenomenological level while hierarchical modeling is primarily concerned with building a theoretical superstructure.[31]

5b Characteristics of horizontal modeling

If data-intensive science is largely constrained to the phenomenological level, then this suggests a fairly theory-independent scientific practice. This may be the true core of the widely exaggerated claims concerning an alleged "end of theory" (Anderson 2008). Again, the modeling of data-intensive science is particularly suited for the causal analysis of complex phenomena in areas that lack a pronounced theoretical level—when large amounts of data have to be taken into account, with which human memory and computational capacities cannot deal anymore. The success of data-intensive science in dealing with causal complexity is possible due to the automation of the entire scientific process from data collection to data processing and model building to making novel predictions—which was identified as a central feature in Section 2a. This automation allows that the epistemic conditions for data-intensive science can differ substantially from those under which the human cognitive apparatus models phenomena[32], in particular in terms of perceptive faculties, storage capacity, and computational power.

Most importantly, while humans have to be very efficient in determining which data to keep and which to forget or not even perceive in the first place, computers can often store and handle all the data they are collecting. Consequently, conventional scientific modeling is geared at an efficient data reduction and an adequate structuring of knowledge resulting in a hierarchy of laws of increasing generality. By contrast, data-intensive modeling has a different nature due to the ability to quickly access and handle enormous amounts of data. The hierarchical structuring becomes largely dispensable for prediction and manipulation, hence the term *horizontal modeling* for the big-data approach. Distinctive features of such modeling are the following, which in principle are all implied by the characterization of horizontal

[31] This terminology does not correspond to the way statisticians speak of hierarchical modeling in terms of individual and aggregate variables, e.g. individuals, firms, markets (e.g. Russo 2009, 315). Again, the causal level can easily include variables from all of these 'ontological' levels.
[32] A similar argument is given by Humphreys 2004 in the first chapter on 'epistemic enhancers'.

modeling as causal modeling for complex phenomena. The characteristics are well illustrated by the example of data-driven machine translation[33] from Section 3a:

i) Predictions in horizontal modeling are made *rather directly from the data without taking recourse to substantial modeling assumptions about the causal structure of a phenomenon* (Pietsch forthcoming), often just by looking up instances in the data that are sufficiently similar to the instance that is to be predicted. Thus, the causal relationships can be very complex and highly context-specific, often involving a large number of parameters. Consequently, the number of laws will usually dwarf that in conventional scientific modeling, as well illustrated by the example of machine translation, which relies on many thousands of empirical relationships.

ii) Since the data already represents (a significant fraction of) all relevant configurations of the phenomenon, there is little need to introduce abstract levels of description. Data-intensive models thus largely *lack the hierarchical, nested structure* that is characteristic of most conventional science. Again, this is well illustrated by the example of statistical machine translation which apparently functions largely without modeling the grammatical structure of a language.

iii) Relatedly, *the explanatory power of horizontal models is much smaller* than that of hierarchical models. After all, models become more explanatory according to many accounts of scientific explanation the more pronounced the hierarchical structure is with each new level of laws or rules constituting a new level of explanation. Consequently, the horizontal models provide little understanding[34], e.g. the understanding of a language is poor without knowledge of the grammatical structure. This aspect lies behind wide-spread claims that data-intensive science can account only for the fact that something is happening but not for why it is happening. I will briefly elaborate on this in Section 5d.

iv) *Idealizations and simplifications play only a minor role* in horizontal modeling compared with the hierarchical approach, since these are usually introduced to link different levels of generality. Certainly, there are modeling assumptions also in the horizontal approach for example concerning the choice which data to collect and how to analyze it. But in the horizontal approach there is no need to formulate general rules that hold only approximately and have considerable exceptions.

Note further that data is handled in specific ways in data-intensive science. Mostly, few restrictions are imposed on the extent and kind of data that is gathered and analyzed (‘messy data’). Also, there are often only a few very general modeling assumptions guiding the formulation and analysis of the data. Finally, few data is discarded in the modeling process compared to conventional data modeling. These features are linked with some remarkable developments in statistics to which we will turn now.

[33] One should stress again that translation rules are of course not causal relationships. As we had discussed in Section 3a, eliminative induction works just as well for the ‘conventional necessity’ of rules as for the ‘empirical necessity’ of laws.

[34] The term ‘understanding’ is used from now on in the sense of the theoretical explanation described in Section 5d.

5c Science without equations: Novel paradigms in statistics

Much of statistics in the 20th century was model-driven relying to considerable extent on the existence of some levels of generalized theoretical laws. This holds in particular for the hypothesis-testing of classical statistics, which has been the dominant paradigm for much of the 20th century. By contrast, the horizontal modeling that was described in the previous sections is strongly inductive and largely data-driven. From this follows a challenge for contemporary statistics to develop methods that are suited for the novel data-rich contexts lacking substantial theoretical background knowledge. This need for novel methodology is increasingly recognized in the statistics community. For example, in a 2010-piece in *Amstat News*, the magazine of the American Statistical Association, Mark van der Laan and Sherri Rose argue under the heading "Statistics ready for a revolution" that the "next generation of statisticians must build tools for massive data sets".[35] By now, major research initiatives on the topic have formed in various countries.

In a classic paper entitled "Statistical Modeling: The Two Cultures", the renowned statistician Leo Breiman describes exactly this shift from statistics as a model-driven enterprise, a practice that he calls *data modeling*, to statistics as an inductive, data-driven enterprise, to which he refers as *algorithmic modeling* (2001). In the former, statisticians build a model from the data, where typical models involve predictor variables, random noise and model parameters, which are all mapped by a given function on the response variables. The model structure has to be motivated from a more comprehensive theoretical context and the models are either accepted or rejected largely on the basis of goodness-of-fit tests. By contrast, the second type of modeling does not assume a simplifying model to exist, but rather works directly with the entire data to make predictions. The data mechanism is treated as a black box, as either unknown or unknowable. Thus, there are no models to be evaluated on a yes-or-no basis, algorithmic modeling is evaluated purely in terms of predictive accuracy. (Breiman 2001, 199)

With respect to the discussion in the previous section, the data modeling culture assumes that there exists a meaningful theoretical level, from which the data can be understood, and thus engages in hierarchical modeling, while the algorithmic modeling culture is very much concerned with developing statistical tools for horizontal modeling. Breiman estimates that at the time of writing only two percent of the statistics community was concerned with algorithmic modeling with major input coming from other fields in particular of course from machine learning. He lists classification trees and neural nets as primary examples for the algorithmic approach, and regression analyses with a specified functional form such as logistic or linear regression as examples for data modeling. (Breiman 2001, 199)

It would be wrong to think of the emergence of an inductive, data-driven paradigm in statistics as a sudden phenomenon. Breiman describes how he developed his view on algorithmic modeling during his time working as a consultant outside of academia in the 1960s and 70s. A period of major progress in algorithmic modeling then begins around the mid-1980s. One does not have to be a historian to observe that these theoretical developments

[35] http://magazine.amstat.org/blog/2010/09/01/statrevolution/ accessed 31.1.2015.

parallel enormous advances in information technology around the same time period.[36] Many core ideas of algorithmic modeling have in fact been known for many centuries, e.g. the concept of classification trees or simple non-parametric regression methods like 'connect the dots'. However, they have gained significant scientific importance only due to the emergence of powerful information technology for acquisition, storage, and processing of large data sets. This in turn has led to further theoretical analysis regarding the reliability and limits of the various algorithmic approaches as well as to the development of novel statistical tools. In recent times, from the mid-90s onwards, it was in particular the increasing interconnectedness of information technology, in particular the internet, that has led to the emergence of very large high-dimensional data sets for example in the social domain. In a way, the recent hype concerning big data is the culmination of a development that has started much earlier.

Since these shifts concern methodology and not theoretical or empirical content, they differ in important ways from scientific revolutions. For instance, the emergence of novel methodologies does not imply the abandonment of older ones. Nevertheless, the statistics community currently experiences some of the social ramifications and 'culture clashes' that are typical for scientific paradigm shifts as documented for example in Breiman (2001) or in Peter Norvig's dispute with Noam Chomsky on data-driven machine translation (Norvig 2011).

There are by now several well established fields of algorithmic modeling, for example neural nets or classification trees, while other important developments may still lie ahead of us. I will in the following focus on a specific example that very well illustrates the distinction between data and algorithmic modeling, namely the shift from *parametric* to *non-parametric* modeling.[37] The latter is by now an established approach in statistics, but its ever-increasing significance for applications is mostly owed to advances in information technologies. Goeran Kauermann explicitly makes the link: 'Statistics and econometrics have been dominated by linear and parametric models over decades. A major reason for this were the numerical possibilities which simply forbid to fit highly structured models with functional and dynamic components.' (2006, 137)

The distinction between non-parametric and parametric refers to the number and nature of assumptions in the respective models. Unlike parametric models, non-parametric models cannot be characterized by a bounded set of model parameters.[38] Non-parametric models thus allow for a wide range of functional dependencies between input and output variables, while in parametric modeling the functional form is usually predetermined by a finite set of parameters, e.g. the mean μ and the standard deviation σ in case of a Gaussian distribution.

Let me illustrate the changes by means of two examples, first the comparison between parametric and non-parametric regression and second between parametric and non-parametric

[36] For a graphic illustration of this claim compare the terms 'computer' and 'non-parametric' on Google's Ngram Viewer https://books.google.com/ngrams.

[37] Hastie & Tibshirani (1990) is a milestone; a useful overview can be found in Kauermann 2005; from a philosophical perspective, Sprenger 2011 discusses an interesting example of non-parametric modeling, bootstrap resampling, and argues for its epistemic significance.

[38] Here, parameters are to be understood not in terms of variables but of constant values determining the properties of a specific model: e.g. in the model y=ax+b below, a and b are model parameters.

density estimation. Afterwards, I will discuss the findings in light of the discussion regarding causal modeling and in particular the distinction of horizontal vs. hierarchical modeling.

In a parametric univariate linear regression problem, one has reasonable grounds to suspect that a number of given data points $(x_i;y_i)$ can be summarized in terms of a linear dependency: $y = ax + b$. Thus, two parameters need to be determined, offset b and slope a, which are usually chosen such that the sum of the squared deviations $\sum_{i=1}^{n}(y_i - (ax_i + b))^2$ is minimized.

In non-parametric regression, the data is not summarized in terms of a small number of parameters a and b, but rather all data is kept and used for predictions (Russell & Norvig 2009, Ch. 18.8.4). A simple non-parametric procedure is *connect-the-dots*. Somewhat more sophisticated is locally weighted regression, in which a regression problem has to be solved for every query point x_q. The y_q-value is determined as $y_q = a_q x_q + b_q$ with the two parameters fixed by minimizing $\sum_{i=1}^{n} K(d(x_q, x_i))(y_i - (a_q x_i + b_q))^2$. Here, K denotes a so-called kernel function that specifies the weight of the different x_i depending on the distance to the query point x_q in terms of a distance function $d()$. Of course, an x_i should be given more weight the closer it is to the query point.

The generalization to higher dimensions is straight-forward though for next-neighbor methods an important issue arises that has been termed the 'curse of dimensionality' (Bellman 1961). With an increasing number of dimensions, i.e. of predictor variables, the average distance between neighboring points rapidly becomes very large of order $(1/N)^{1/n}$, where N is the total number of points and n the number of dimensions. Consequently, the data points will almost always be sparsely distributed in many dimensions.[39]

Let us briefly reflect how these regression methods illustrate differences between parametric and non-parametric modeling. While in the example of parametric regression, linear dependency is presupposed as a modeling assumption, the non-parametric method can adapt to arbitrary dependencies. In parametric regression, the nature of the functional relationship has to be independently justified by reference to a theoretical context, which prevents an automation of the modeling process. Certainly, non-parametric regression also makes modeling assumptions, e.g. a suitable kernel function must be chosen that avoids both over- and underfitting. However, within reasonable bounds the kernel function can be chosen by comparing the predictions of the algorithm with the values in a test set. Since often, predictions turn out relatively stable with respect to different choices of kernel functions, an automation of non-parametric modeling remains feasible.

While non-parametric regression is more flexible than parametric regression, it is also much more data-intensive and requires more calculation power. Notably, in the parametric case, a regression problem must be solved only once. Then all predictions can be calculated from the resulting parametric model. In the non-parametric case, a regression problem must be solved

[39] Note that this curse of dimensionality does not automatically apply to all big-data algorithms. To the contrary, it occasionally turns out helpful to artificially increase the dimensionality of the variable space in methods like decision trees or support vector machines (Breiman 2001, 208-209). Also, if additivity is assumed between the different influences, the curse loses its spell (Kauermann 2006, 144).

for every query point. In principle, each prediction takes recourse to all the data. While the parametric model consists in a relatively simple mathematical equation, the non-parametric model consists in all the data and an algorithmic procedure for making predictions. Note that the predictive reliability of non-parametric regression can again be evaluated in terms of eliminative induction and a difference-making account of causation. In fact, non-parametric regression just implements Mill's method of concomitant variation, which in turn essentially relies on the method of difference. In principle, the same premises already given in footnote 22 need to be fulfilled (Pietsch forthcoming, Sec. 5).

Consider density estimation as a second example (Russell & Norvig 2009, Ch. 20.2.6). The parametric approach makes an explicit assumption about the nature of the distribution function, for example a Gaussian distribution $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. This distribution is determined by two parameters, the mean μ and the standard deviation σ, which are chosen such that a best fit with the data is achieved.

A simple non-parametric approach is the histogram method, where the parameter space is partitioned into cells of equal volume ΔV and the number k_i of all N data points is counted for each cell i. The density is given by $f(x) = k_i / N \Delta V$, where k_i is the number of data points in the same cell as the query point x. A closely related often more effective non-parametric method is k-nearest-neighbors, where the same formula is used but k is now fixed and one determines the minimal volume ΔV surrounding the query point x such that k points are included. The parameter k should be chosen in a way to avoid overfitting, but still be sufficiently sensitive. A suitable k can be fixed by comparing the values in a test set with the predictions by the trained model allowing for straight-forward automation of the non-parametric approach.

Again, in the parametric case the data is summarized in terms of a model characterized by a few parameters μ and σ resulting in a simple formula, while the non-parametric method makes no assumptions about the nature of the distribution function and is thus much more flexible. On the other hand, the non-parametric method is very data-intensive since it uses the original data points to make predictions. The difference between the two types of models is striking: While parametric models usually are more or less simple equations, the non-parametric models consist in the original data plus an algorithm to derive predictions from the data. Since there is no bounded set of parameters in the latter, the non-parametric models cannot be framed in terms of general equations at all.

From such examples a list of several features can be drawn that distinguish parametric from non-parametric modeling. These are intricately connected with the characteristics of horizontal modeling of causal structure as described in Section 5b: i) Parametric methods usually presuppose considerable modeling assumptions that must be based on background theory. In particular, they summarize the data in terms of a 'small' number of model parameters specifying for example a Gaussian distribution or linear dependence, hence the name. By contrast, non-parametric modeling presupposes *few and weaker modeling assumptions*, e.g. allows for a wide range of functional dependencies or of distribution functions. Certainly, some parameters also occur in non-parametric modeling, e.g. the number

of neighbors taken into account in the k-nearest neighbors algorithm. ii) In non-parametric modeling, predictions are calculated *on the basis of 'all' data*. There is no detour over a parametric model that summarizes the data in terms of a few parameters.[40] iii) While this renders non-parametric modeling quite *flexible* with the ability to quickly react to unexpected data, it also becomes extremely *data- and calculation-intensive*. This aspect accounts for the fact that non-parametric modeling is a relatively recent phenomenon in scientific method.

Non-parametric models allow for novel ways to deal with complexity: iv) A crucial shift occurs from *equation modeling* to *algorithmic modeling*. Conventional parametric modeling in terms of equations, describing for example functional dependencies or distribution functions, already presupposes that the picture has been reduced to a (usually small) number of parameters as well as to (often relatively simple) functional relationships. By contrast, non-parametric modeling does not have such restrictions. It relies less on sophisticated mathematics and more on a brute-force execution of a large number of steps, when for example an algorithm searches a large data-base for similar cases. Since the set of parameters is by definition unbounded, non-parametric models cannot be expressed in terms of general equations. Consequently, the question arises how to represent them on a fundamental level. A basic suggestion in this regard would be that the algorithm together with the data take over the role of the equations. Thus, *non-parametric models consist of the original data and an algorithmic procedure to derive predictions from the data*.

v) The complexity of non-parametric models prevents a deeper understanding of the phenomena. Thus, there is a *shift in epistemic values* regarding the aims of modeling. Non-parametric modeling is geared almost exclusively at prediction and manipulation and rarely at understanding in terms of general laws or rules (cp. characteristic iii in Section 5b). By contrast, parametric modeling usually emphasizes understanding. While parametric modeling often correlates with a *realist* and *reductionist* viewpoint, non-parametric modeling has *instrumentalist* and *pluralist* connotations. The instrumentalist attitude is for example apparent in the wide-spread use of ensemble methods that combine different models even if these start from mutually contradictory assumptions. Presumably, this shift in epistemic values is at the root of the mentioned divide between Breiman's different 'cultures' of statistical modeling.

In summary, the increasing availability of data leads to the emergence of novel paradigms in statistics that are well suited for the data-driven, strongly inductive approach of data-intensive science aiming to analyze phenomena with a complex causal structure.

5d Outlook: Big data's alleged lack of explanatory power

There is one further issue that merits a detailed treatment in a separate article, but that I want to briefly address because it is intimately connected with the shift from hierarchical to horizontal modeling. It concerns the question in which ways data-intensive science can be explanatory or not. The idea that with big data and with the alleged shift from causation to correlation science ceases to be explanatory plays a considerable role in popular as well as

[40] Both i) and ii) are of course closely related to the first characteristic stated in Section 5b.

academic accounts. A representative example is the following quote: 'The correlations [found in big-data science] may not tell us precisely *why* something is happening, but they alert us *that* it is happening.' (Mayer-Schönberger & Cukier 2013, 21) In combination with the claim that correlations often are enough and that the notion of causation loses its significance in data-intensive science, this would imply a scientific practice that ceases to be explanatory.

Again, the conceptual details are rather sophisticated. To understand how data-intensive modeling can be causal but fail to exhibit certain explanatory virtues of conventional science, various notions of explanation have to be carefully distinguished as discussed in the extensive philosophy-of-science literature on this issue.[41] As with many methodological theses on big data, the original claims by data evangelists are rather exaggerated, while the rejoinders by those defending the traditional ways of science often fail to do justice to the interesting epistemic shifts that are happening. Consider again the example of machine translation from Section 3a. Obviously, the statistical approach lacks a lot of the explanatory virtues of the rule-based approach. For example, in the latter one can explain the position of words in a sentence by referring to grammatical rules for the sentence structure, or explain the ending of verbs by means of conjugation rules, the ending of substantives by means of declination rules etc.

In a purely statistical approach, all these explanatory virtues cease to exist—mainly because there is no hierarchical structure of increasingly general rules. Essentially, the only type of explanation that remains in the data-driven approach is explanation by similarity. A certain translation can be explained by referring to sufficiently similar instances of successful translation. Certainly, such an explanation does not provide as much intellectual satisfaction as explanations referring to grammatical rules, but it is explanatory in the rudimentary sense that it points to relevant evidence which justifies the translation—thereby answering the question why it was used in a certain context.

A good starting point for discussing the alleged lack of explanatory virtue in data-intensive science is the following distinction[42]: (i) to explain by giving an argument that derives what is to be explained from a number of general laws or rules thereby relating a phenomenon to other phenomena and achieving unification. Such explanations can be formulated for example in the rule-based modeling of languages; (ii) to explain by citing the causal factors that can account for a certain event, where these factors are difference-makers and can be identified by eliminative induction. With respect to the example of data-driven machine translation, one might explain a successful translation of a certain word by pointing to relevant word sequences in the vicinity of the word.[43] In the first category of explanation, general laws are explanatory, explanations have the structure of arguments, and they mostly aim at unification. In the second category, causal factors are explanatory and explanations consist in lists of factors. Since data-intensive science is about causal modeling in terms of eliminative

[41] An excellent introduction is Psillos (2002).
[42] A similar distinction is drawn in Gijsbers (2013). His terminology is quite useful for the present analysis, with some minor disagreements between our perspectives, the discussion of which would lead too far astray and has to be postponed to a more in-depth treatment of explanation in data-intensive science.
[43] Note that some overlap can exist between both kinds of explanation, in particular if the causal laws are sufficiently general.

induction but fails to produce a hierarchical structure, data-intensive models mostly yield explanation in the second sense but largely fail to be explanatory in the first sense.

The distinction fits well with the epistemological framework sketched in Section 4a. And in fact, Nancy Cartwright introduces a similar dichotomy: "there are two quite different things we do when we explain a phenomenon in physics. First, we describe its causes. Second, we fit the phenomenon in a theoretical frame" (1983, 16). The former stays at the causal level and thus fits well with horizontal modeling, the latter refers to theoretical structure and thus depends on hierarchical modeling.

Note that complex phenomena in the high-level sciences, e.g. the social sciences or medicine, may not be accessible to explanation (i) at all, if laws of significant generality fail to exist. Consequently, human understanding of these phenomena in terms of theoretical explanations may always be considerably impaired while reliable predictions and manipulations may still be possible using data-driven approaches. This may in the future force us to reconsider the role of human experts in the complex sciences that have been traditionally conceived to guide the research process by providing understanding. Data-intensive science often functions without much of a theoretical level, as some of the pertinent examples show: machine translation without knowledge of grammar, advertising without classical advertising knowledge, campaigning without in-depth political-science knowledge.

6. Conclusion: The new science of complexity

The horizontal modeling based on algorithmic procedures and novel statistical approaches will in the coming years greatly extend the causal knowledge in the complex sciences. Opportunities lie for example in medicine and epidemiology when dealing with complex diseases like allergies, asthma, and cancer or in ecology when trying to understand complex processes like the recent worldwide decline in bee populations. Presumably, more effective ways of management will become possible through big data in both economics and politics. However, there are also considerable dangers concerning potential abuse especially in the social sciences, where most of the large data sets are currently collected.

The knowledge established by data-intensive methods will consist in a large number of causal relationships that generally involve numerous predictor variables and that are highly context-specific. The complexity of these laws and the lack of a hierarchy into which they could be integrated prevent a deeper understanding, while allowing for predictions and interventions. Almost certainly, we will experience the rise of entire sciences that cannot leave the computers and do not fit into conventional textbooks.

Acknowledgments

I am grateful to Mathias Frisch, Sabina Leonelli, and Sylvester Tremmel for helpful discussions as well as to audiences in Enschede, Delft, Bielefeld, and at the 7th Workshop on the Philosophy of Information in London. The article would not be what it is without the enormous help and useful advice of four anonymous referees. I am particularly indebted to

one of them, who has put an incredible effort into improving the manuscript over several rounds of revisions.

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

Bacon, Francis. 1620/1994. *Novum Organum*. Chicago, Il: Open Court.

Baumgartner, Michael & Gerd Graßhoff. 2004. *Kausalität und kausales Schließen.* Norderstedt: Books on Demand.

Bellman, Richard E. 1961. *Adaptive Control Processes: A Guided Tour.* Princeton: Princeton University Press.

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.

Callebaut, Werner. 2012. "Scientific perspectivism: A philosopher of science's response to the challenge of big data biology." *Studies in History and Philosophy of Biological and Biomedical Science* 43(1):69-80.

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

Cartwright, Nancy. 1999. *The Dappled World.* Cambridge: Cambridge University Press.

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

Floridi, Luciano. 2012. "Big Data and Their Epistemological Challenge." *Philosophy & Technology* 25:435-437.

Frické, Martin H. 2014. "Big Data and its Epistemology." *Journal of the Association for Information Science and Technology*. DOI: 10.1002/asi.23212.

Gijsbers, Victor A. 2013. "Understanding, Explanation, and Unification." *Studies in History and Philosophy of Science* 44(3):516-522.

Gillies, Donald. 1996. *Artificial Intelligence and Scientific Method.* Oxford: Oxford University Press.

Gray, Jim. 2007. "Jim Gray on eScience: A Transformed Scientific Method." In Tony Hey, Stewart Tansley & Kristin Tolle (eds.). *The Fourth Paradigm. Data-Intensive Scientific Discovery.* Redmond, WA: Microsoft Research. http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_jim_gray_transcript.pdf

Guyon, Isabelle &André Elisseeff. 2003. "An Introduction to Variable and Feature Selection." *Journal of Machine Learning Research* 3(2003):1157-1182.

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

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

http://www.csee.wvu.edu/~gidoretto/courses/2011-fall-cp/reading/TheUnreasonable%20EffectivenessofData_IEEE_IS2009.pdf
Hastie T. & R. Tibshirani, 1990. Generalized Additive Models. London: Chapman and Hall.
Herschel, John F. W. 1851. *Preliminary Discourse on the Study of Natural Philosophy*. London: Longman, Brown, Green, and Longmans.
Hey, Tony, Stewart Tansley & Kristin Tolle. 2009. *The Fourth Paradigm. Data-Intensive Scientific Discovery.* Redmond, WA: Microsoft Research.
Hume, David. 1777. *An Enquiry Concerning Human Understanding.* Oxford: Clarendon Press.
Humphreys, Paul. 2004. *Extending Ourselves. Computational Science, Empiricism, and Scientific Method.* Oxford: Oxford University Press.
Illari, Phyllis & Federica Russo. 2014. Causality. Philosophical Theory meets Scientific Practice. Oxford: Oxford University Press.
Issenberg, Sasha. 2012. *The Victory Lab: The Secret Science of Winning Campaigns.* New York, NY: Crown.
Jelinek, Frederick. 2009. "The Dawn of Statistical ASR and MT." *Computational Linguistics* 35/4:483-494.
Kauermann, Goeran. 2006. "Nonparametric Models and their Estimation." In Olaf Hübler & Joachim Frohn (eds.). *Modern Econometric Analysis* (pp. 137-152). Springer: Berlin.
Kitchin, Rob. 2014. "Big Data, New Epistemologies and Paradigm Shifts." *Big Data & Society* 1:1-12.
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 (ed.). 2012a. Data-driven Research in the Biological and Biomedical Sciences. *Studies in History and Philosophy of Biological and Biomedical Sciences* 43(1).
Leonelli, Sabina. 2012b. "Classificatory Theory in Data-Intensive Science: The Case of Open Biomedical Ontologies." *International Studies in the Philosophy of Science* 26(1):47-65.
Leonelli, Sabina. 2013. "Integrating Data to Acquire New Knowledge: Three Modes of Integration in Plant Science." *Studies in the History and Philosophy of the Biological and Biomedical Sciences: Part C.*
Lewis, David. 1973. "Causation." *Journal of Philosophy* 70:556-67.
Mackie, John L. 1965. "Causes and Conditions." *American Philosophical Quarterly* 12:245-265.
Mackie, John L. 1980. *The Cement of the Universe.* Oxford: Clarendon Press.
Mayer-Schönberger, Viktor & Kenneth Cukier. 2013. *Big Data.* London: John Murray.
Mayo, Deborah. 1996. *Error and the Growth of Experimental Knowledge.* Chicago: University of Chicago Press.
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.

Norvig, Peter. 2011. “On Chomsky and the Two Cultures of Statistical Learning.” http://norvig.com/chomsky.html
Norvig, Peter. 2009. “All we want are the facts, ma’am.” http://norvig.com/fact-check.html
Pearl, Judea. 2000. *Causality. Models, Reasoning, and Inference.* Cambridge: Cambridge University Press.
Pietsch, Wolfgang. 2014. “The Nature of Causal Evidence Based on Eliminative Induction.” In P. Illari & F. Russo (eds.), *Topoi*. Doi:10.1007/s11245-013-9190-y
Pietsch, Wolfgang. Forthcoming. “Aspects of Theory-Ladenness in Data-Intensive Science”, *Philosophy of Science*. Preprint: http://philsci-archive.pitt.edu/10777/1/pietsch_data-intensive-science_psa.pdf
Psillos, Stathis. 2002. *Causation and Explanation.* Durham: Acumen.
Reutlinger, Alexander. 2012. “Getting Rid of Interventions.” *Studies in the History and Philosophy of Science Part C* 43 (4): 787-795.
Russell, Stuart & Peter Norvig. 2009. *Artificial Intelligence.* Upper Saddle River, NJ: Pearson.
Russo, Federica. 2009. *Causality and Causal Modelling in the Social Sciences. Measuring Variations.* Springer.
Schmidt, Michael & Hod Lipson. 2009. “Distilling Free-Form Natural Laws from Experimental Data.” *Science* 324(5923):81-85.
Siegel, Eric. 2013. *Predictive Analytics.* Hoboken, NJ: John Wiley & Sons.
Spirtes, Peter, Clark Glymour & Richard Scheines. 2000. *Causation, Prediction and Search.* Cambridge, MA: M.I.T. Press.
Sprenger, Jan. 2011. “Science without (parametric) models: the case of bootstrap resampling.” *Synthese* 180(1):65-76.
Steinle, Friedrich. 1997. “Entering New Fields: Exploratory Uses of Experimentation.” *Philosophy of Science* 64:S65-S74.
Steinle, Friedrich. 2005. *Explorative Experimente.* Stuttgart: Franz Steiner Verlag.
Suppes, Patrick. 1962. “Models of Data”, in Ernest Nagel, Patrick Suppes and Alfred Tarski (eds.), *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress*. Stanford: Stanford University Press, 252–261.
Winsberg, Eric. 2010. *Science in the Age of Computer Simulation.* Chicago, Il: University of Chicago Press.
Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.
Worrall, John. 1989. “Structural realism: The best of both worlds?” *Dialectica* 43: 99–124.