Aspects of Theory-Ladenness in Data-Intensive Science

Wolfgang Pietsch*†

Recent claims, mainly from computer scientists, concerning a largely automated and model-free data-intensive science have been criticized by several philosophers of science. The debate suffers from lack of detail regarding the actual methods used in data-intensive science and in which ways these presuppose theoretical assumptions. I examine two widely used algorithms, classificatory trees and nonparametric regression, and argue that they are theory laden in an external sense, regarding the framing of research questions, but not in an internal sense, concerning the causal structure of the examined phenomenon. With respect to the novelty of data-intensive science, I draw an analogy to exploratory experimentation.

1. Introduction. Over the past decade, computer scientists have claimed that a new scientific methodology has become possible through advances in information technology (e.g., Gray 2007). This approach is supposed to be data driven, strongly inductive, and relatively theory independent. The epistemology of such data-intensive science has recently emerged as a novel topic in philosophy of science. Generally, the reactions of philosophers have been rather critical, often referring to the more or less trivial observation that some kind of theory-ladenness always occurs in scientific research. But, as I will argue, this means throwing out the baby with the bathwater, since interesting shifts in the role of theory can indeed be observed when examining specific methods employed in data-intensive science.

In section 2, I suggest a definition for data-intensive science reflecting those features that are interesting from an epistemological perspective. Then, in section 3, I briefly introduce the debate on theory-ladenness in data-intensive

*To contact the author, please write to: Munich Center for Technology in Society, TU München, Arcisstrasse 21, 80333 Munich, Germany; e-mail: pietsch@cvl-a.tum.de.

†I am grateful to Mathias Frisch, Sabina Leonelli, and Sylvester Tremmel for very helpful insights and discussions.

Philosophy of Science, 82 (December 2015) pp. 905–916. 0031-8248/2015/8205-0015$10.00
Copyright 2015 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/683328 Published online by Cambridge University Press

science. To assess the various arguments, I discuss two algorithms that are widely used, namely, classificatory trees (sec. 4) and nonparametric regression (sec. 5). For both of these methods, I identify the specific ways in which theory has to be presupposed to identify causal connections and thus yield reliable predictions. I conclude in section 6 that these algorithms require an external theory-ladenness concerning the framing of research questions but little internal theory-ladenness concerning the causal structure of the examined phenomena. I also point out remarkable analogies to the analysis of theory-ladenness in exploratory experimentation.

2. Defining Data-Intensive Science. The problems usually addressed in data-intensive science bear close resemblance to standard problems in statistics. They concern classification or regression of an output variable y with respect to a large number of input parameters x, also called predictor variables or covariates, on the basis of large training sets. The main differences compared with conventional problems in statistics consist in the high dimensionality of the input variable and the amount of data available about various configurations or states of the system. For example, an online store wants to know how likely it is that someone will buy a certain product based on surf history, various cookies, and a user profile, as well as on data of other users who have either bought or failed to buy the product. A medical researcher examines which combinations of genetic and environmental factors are responsible for a certain disease. A political adviser is interested to determine how likely it is that a specific individual will 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 has a number of discrete possible values. In a regression problem, the output variable is continuous. In order to establish an adequate and reliable model, extensive training and test data are needed. Each instance in the training and test sets gives a value for the output variable dependent on at least some of the input parameters. The training data are used to build the model, the test data to validate and verify the model.[1]

In this article, we cannot delve into all the technical details of the various algorithms employed in data-intensive science, such as support vector machines, forests, and neural networks. Instead, we look at two simple but widely used algorithms, namely, classificatory trees and nonparametric regression, to examine how much and what kind of theory must be presupposed in order for these algorithms to yield meaningful results.

The term ‘data-intensive science’ is notoriously blurry, as has been emphasized, for example, by Sabina Leonelli: “a general characterisation of

1. An excellent introductory textbook is Russell and Norvig (2009).

https://doi.org/10.1086/683328 Published online by Cambridge University Press

data-driven methods is hard to achieve, given the wide range of activities and epistemic goals currently subsumed under this heading" (2012, 1). However, in order to say something substantial about the role of theory, we have to be more specific about the kinds of practices we want to include as data-intensive science even if such an exact definition does not fully correspond to common usage of the term.

In the computer science literature, various definitions have been proposed for the closely related concept of 'big data'. Most of these refer to the pure amount of information or to the technical challenges that such big data pose in terms of the so-called three Vs—volume, velocity, and variety of data (Laney 2001). However, from a philosophy of science perspective, these definitions do not provide much insight. After all, larger amounts of data do not automatically imply interesting methodological developments.

Leonelli, partly following Gray (2007, xix), identifies two characteristic features for data-intensive methodology: "one is the intuition that induction from existing data is being vindicated as a crucial form of scientific inference, which can guide and inform experimental research; and the other is the central role of machines, and thus of automated reasoning, in extracting meaningful patterns from data" (2012, 1). She adds that these features are themselves quite controversial and criticizes that they are difficult to apply in research contexts.

In defining data-intensive science, I largely follow Leonelli, while attempting to be more precise about the type of induction. I argue that eliminative induction in the tradition of Mill's methods plays the crucial role.[2] The first part of my definition (referred to as premise 1) thus focuses on the evidence that is necessary to carry out eliminative induction: data-intensive science requires data representing all configurations of the examined phenomenon that are relevant with respect to a specific research question. For complex phenomena, this implies high-dimensional data, that is, data sets involving many parameters, as well as a large number of observations or instances covering a wide range of combinations of these parameters. We will see later that this premise underpins the characteristic data-driven and inductive nature of data-intensive science.

The second feature (premise 2) concerns the automation of the entire scientific process, from data capture to processing to modeling (cf. Gray 2007, xix). This allows sidestepping some of the limitations of the human cognitive apparatus but also leads to a loss in human understanding regarding the results of data-intensive science. Again, being more precise about the type of induction allows us to determine under which circumstances automation is really possible.

2. This is not to be confused with a related but looser use of the same term in the sense of eliminating hypotheses until only the correct one remains.

https://doi.org/10.1086/683328 Published online by Cambridge University Press

3. Theory-Free Science? Proponents of data-intensive science claim that important changes are happening with respect to the role of theory. An extreme but highly influential version of such a statement is by Chris Anderson, former editor in chief of *WIRED*, who notoriously proclaimed "the end of theory" altogether (2008). More nuanced positions can be found, for example, in the writings of Google research director Peter Norvig: "Having more data, and more ways to process it, means that we can develop different kinds of theories and models" (2009). Simpler models with a lot of data supposedly trump more elaborate models with less data (Halevy, Norvig, and Pereira 2009, 9).

A number of philosophers have objected to claims of a theory-free science—generally by pointing out various kinds of theory-ladenness. For example, Werner Callebaut writes, "We know from Kuhn, Feyerabend, and . . . Popper that observations (facts, data) are theory-laden. Popper . . . rejected the 'bucket theory of knowledge' in favor of the 'searchlight theory,' according to which observation 'is a process in which we play an intensely active part.' Our perceptions are always preceded by interests, questions, or expectations—in short, by something 'speculative'" (2012, 74). Leonelli concurs in her work on big data biology:

> Using data for the purposes of discovery can happen in a variety of ways, and involves a complex ensemble of skills and methodological components. Inferential reasoning from data is tightly interrelated with specific theoretical commitments about the nature of the biological phenomena under investigation, as well as with experimental practices through which data are produced, tested and modelled. For instance, extracting biologically meaningful inferences from high-throughput genomic data may involve reliance on theories about gene expression and regulation, models of the biological processes being regulated and familiarity with the instruments and organisms from which data were obtained. In this context, 'inductive' clearly does not mean 'hypothesis-free'; nor can automated reasoning be seen as a substitute to human judgment based on specific expertise and laboratory experience. (2012, 2)

Certainly, the idea of an entirely theory- or model-free science is absurd. Hence, Callebaut and Leonelli rightly point out various kinds of theoretical assumptions that enter scientific analyses. But this kind of argument turns out to be too general and in the end fails to do justice to the remarkable shift toward a strongly inductive approach. Instead, the interesting question is in which ways data-intensive science is indeed theory laden, and, more importantly, in which sense it can be theory free. To provide an answer, I now take a detailed look at two algorithms that are widely employed, namely, classificatory trees and nonparametric regression. I link these methods to eliminative induction and then determine the kind of theoretical knowledge that has to be presupposed.

https://doi.org/10.1086/683328 Published online by Cambridge University Press

4. First Case Study: Classificatory Trees. Classificatory trees (e.g., Russell and Norvig 2009, sec. 18.3.3) are used to determine whether a certain instance belongs to a particular group A depending on a number of parameters C1, . . . , CN and thus they perfectly match the scheme of data-intensive problems as described in section 2. With the help of training data, the tree is set up recursively. First, the parameter CX is determined that contains the largest amount of information with respect to the classification of the training data, as formally measured in terms of the Shannon entropy. If CX classifies all instances correctly, the procedure is terminated. Otherwise, two subproblems remain, namely, classifying when CX is present and when it is absent. This step is repeated until either all instances are classified correctly or no potential classifiers are left. If the algorithm is successful, the resulting tree structure gives a Boolean expression of necessary and sufficient conditions for A, which can be interpreted as a complex scientific law: for example, if (C3C2 ∨ C4¬C2)C1 ∨ C6C5¬C1, then A.

The framing of classificatory trees in particular and of problems in data-intensive science in general in terms of a mapping of boundary conditions to an outcome variable fits well with eliminative induction as exemplified in John Stuart Mill's methods of elimination (1886, bk. 3, chap. 8) and with a predecessor in Francis Bacon's method of exclusion (1620/1994, bk. 2). Whereas until the end of the nineteenth century Bacon's approach was widely considered the methodological foundation for modern science, eliminative induction has not been very popular since. Hence, there exist comparably few modern accounts, including von Wright (1951), Mackie (1980, appendix), Skyrms (2000), Baumgartner and Grasshoff (2004), and Pietsch (2014).[3]

In eliminative induction, a phenomenon A is examined under the systematic variation of potentially relevant boundary conditions C1, . . . , CN with the aim of establishing causal relevance or irrelevance of these conditions, relative to a certain context or background B consisting of further boundary conditions. The best-known and arguably most effective method is the so-called method of difference that establishes causal relevance of a boundary condition CX by comparing two instances that differ only in CX and agree in all other circumstances C. If in one instance both CX and A are present and in the other both CX and A are absent, then CX is causally relevant to A. 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 CX has no influence on A. Eliminative induction can deal with functional dependencies, and an extension of the approach to statistical relationships is straightforward.[4]

3. In the following, I largely rely on the last account.

4. For further discussion, see Pietsch (2014).

https://doi.org/10.1086/683328 Published online by Cambridge University Press

Thus, causal (ir)relevance is a three-place relation: a boundary condition C is (ir)relevant to a phenomenon A with respect to a certain background B of further conditions that remain constant if causally relevant or are allowed to vary if causally irrelevant. The restriction to a context B is necessary because there is no guarantee that in a different context B* the causal relation between C and A will continue to hold. Causal laws established by eliminative induction thus have a distinctive contextual or ceteris paribus character. Extensive information about all potentially relevant boundary conditions in as many different situations as possible is necessary to establish reliable causal knowledge by means of eliminative induction. Exactly this kind of information is provided in data-intensive science.

Eliminative induction corresponds to a difference-making account of causality, which is closely related to the counterfactual approach. However, the truth-value of counterfactuals is determined via the method of difference or the direct method of agreement, and thus by comparison with actual situations that differ from the counterfactual statement only in terms of irrelevant circumstances, and not by means of a possible-world semantics as in traditional counterfactual approaches like that of David Lewis.

Obviously, classificatory trees rely on eliminative induction. Thus, to assess their quality, one has to look at the premises required for eliminative methods to yield the correct causes. Partial analyses of this problem are given, for example, in Keynes (1921, chap. 22), von Wright (1951, chap. 5), Baumgartner and Grasshoff (2004, sec. 9.2.4), and Pietsch (2014, sec. 3f). I will again follow the exposition in the last reference. There are at least three main assumptions: (i) determinism, that is, that the phenomenon A is fully determined by boundary conditions C and background B; (ii) constancy of the background, that is, that no relevant parameters in the background change when two instances are compared via the method of difference or the strict method of agreement; and finally (iii) an adequate vocabulary, that is, that the parameters C reflect suitable causal categories for the given context B. Applied to classificatory trees, we can say, for example, that if there is a single sufficient condition CX among the parameters C and there are sufficient data in terms of instances of the system in various configurations to avoid spurious correlations, then the classificatory tree algorithm will return CX as the cause. Certainly, these assumptions i–iii are quite strong. There are supposedly weaker constraints for causal relations of statistical nature, but this issue goes beyond the scope of the present paper.

We can now identify the elements of theory that have to be presupposed. In particular, (*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, that is, a stable context B;

https://doi.org/10.1086/683328 Published online by Cambridge University Press

(*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; and (*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 are enough data to avoid spurious correlations and to map the causal structure of the phenomenon without further internal theoretical assumptions about the phenomenon.

This motivates and explains the definition of data-intensive science given in section 2. In particular, premise 1 turns out to be the fundamental condition allowing for a strongly inductive approach based on parameter variation. This viewpoint is further corroborated by the fact that in many cases data-driven approaches become effective rather suddenly—a transition point that could be called a data threshold (Halevy et al. 2009). Halevy et al. give 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). At this threshold, the data represent a large fraction of the relevant configurations of the considered phenomenon.

Of course, in scientific practice full theoretical knowledge (*a–d*) is rarely available. However, in general, including more potentially relevant parameters C will increase the probability that the actual cause of A might be among them, while admittedly also increasing the probability for spurious correlations, that is, that boundary conditions accidentally produce the right classification. However, more data in terms of instances of different configurations can reduce the probability for such spurious correlations. Thus, more data in terms of parameters and instances will generally increase the chance that correct causal relations are identified by data-intensive algorithms.

5. Second Case Study: Nonparametric Regression. A recent paradigm shift in statistics closely mirrors the change from a hypothesis-directed to a more inductive, data-driven approach. It has been described as a transition from parametric to nonparametric modeling (e.g., Wasserman 2006; Russell and Norvig 2009, sec. 18.8), from data to algorithmic models (Breiman 2001), or from model-based to model-free approaches.[5] Since the shift concerns methodology and not theoretical or empirical content, it differs in important ways from scientific revolutions. Nevertheless, the statistics community has experienced over the past two decades some of the social ramifications and 'culture clashes' that are typical for scientific paradigm shifts, as documented, for example, in Breiman (2001) or in Norvig's dis-

5. One could argue that these refer to different, but closely related, paradigm shifts. Owing to lack of space, a detailed discussion must be left to another occasion.

https://doi.org/10.1086/683328 Published online by Cambridge University Press

pute with Noam Chomsky on data-driven machine translation (Norvig 2011).

The paradigm shift has the following basic features: (1) Parametric methods usually presuppose considerable modeling assumptions. 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, nonparametric modeling presupposes few modeling assumptions, that is, allows for a wide range of functional dependencies or of distribution functions. (2) In nonparametric modeling, predictions are calculated on the basis of 'all' the data. There is no detour over a parametric model that summarizes the data in terms of a few parameters. (3) While this renders nonparametric 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 nonparametric modeling is a relatively recent development in scientific method strongly dependent on advances in information technology.

Let me give a simple example as an illustration: the comparison between parametric and nonparametric regression. 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 nonparametric regression, the data are not summarized in terms of a small number of parameters a and b, but rather all data are kept and used for predictions (e.g., Russell and Norvig 2009, sec. 18.8.4). A simple nonparametric 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.

Let us briefly reflect how these regression methods illustrate the differences between parametric and nonparametric modeling (features 1–3). While in parametric regression linear dependency is presupposed as a modeling assumption, the nonparametric method can adapt to arbitrary dependencies. In parametric regression, the nature of the functional relationship has to be independently justified by the theoretical context, which prevents an automation of the modeling process. Certainly, nonparametric regression also makes modeling assumptions, for example, a suitable kernel function must be chosen that avoids both over- and underfitting. However, since pre-

https://doi.org/10.1086/683328 Published online by Cambridge University Press

dictions often turn out relatively stable with respect to different choices of kernel functions, an automation of nonparametric modeling remains feasible.

While nonparametric 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 nonparametric case, a regression problem must be solved 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 nonparametric model consists in all the data and an algorithmic procedure for making predictions.

The main difference in terms of theoretical assumptions is that in parametric regression the type of functional dependency is presupposed, in contrast to nonparametric regression. The latter again relies on eliminative induction. Essentially, it constitutes a case of Mill's method of concomitant variations, which derives its inferential power from the method of difference, as argued, for example, in Skyrms (2000, sec. 5.9) or Pietsch (2014, sec. 3d). Therefore, the conditions for identifying a causal relationship are largely the same as those discussed in the previous section—determinism, constancy of the background, and correct causal language—resulting in the same premises in terms of theoretical assumptions *a–d*. In particular, when mapping a functional dependency, all causally relevant conditions in the background must remain constant. And there must be sufficient data points such that the functional dependence can be traced in adequate detail.

6. Conclusion: Data-Intensive Science and Exploratory Experimentation. We are finally in a position to evaluate the claims concerning a theory-free science. In both case studies, certain elements of theory had to be presupposed in order to yield reliable results in terms of causal structure that in turn can ensure successful prediction and manipulation. In particular, among the considered parameters must be those that are causally relevant for a phenomenon in a considered context and not too many that are causally irrelevant to avoid spurious correlations. Also, the parameters should reflect adequate causal categories. Finally, the collected instances or observations should cover all configurations that are relevant with respect to a given research question.

Because these aspects all concern the framing of the problem, one could speak of external theory-ladenness. By contrast, there is another kind of theory-ladenness that is largely absent from data-intensive science. For example, in classificatory trees no hypotheses are made about causal connections that link the various parameters. Equally, in nonparametric regression, no assumptions are presupposed about the functional dependencies between different quantities. Thus, the essential difference in comparison with a hypothesis-driven approach is that not much is presupposed about the

https://doi.org/10.1086/683328 Published online by Cambridge University Press

internal causal structure of the phenomenon. Rather, this structure is mapped from the data by parameter variation.

How novel is all this? On closer scrutiny, data-intensive science much resembles the practice of exploratory as distinguished from hypothesis-directed experimentation (Burian 1997; Steinle 1997; Waters 2007; see also Vincenti 1993, 291). Exploratory experimentation essentially consists in the very same parameter variation of eliminative induction, where the experimenter tries to map the system of interest in all those states that she considers relevant. It is this common methodological core that links exploratory experimentation and data-intensive science and speaks against the claim, for example, by Krohs (2012) that the latter constitutes a novel experimental approach focusing on data gathering.

Not surprisingly, the debate concerning theory-ladenness in exploratory experimentation parallels the discussion in the present article. For example, Steinle (2005, 285) suggests a distinction between different kinds of theory-ladenness. According to his view, exploratory experimentation presupposes theoretical knowledge in terms of classification systems or empirical rules, but not in terms of theories that postulate empirically inaccessible abstract entities. Steinle refers to Duhem, Hacking, and Cartwright as having drawn similar distinctions between an experimental/phenomenological and a theoretical level in scientific theories. Indeed, the distinction between exploratory and hypothesis-driven experimentation fits well with Hacking's (1983) claim that experiments have a life of their own and Cartwright's (1983) position of entity realism, which postulates a causal level in science that is mostly phenomenological and largely independent of the theoretical level.

Building on Burian and Steinle's work, Kenneth Waters emphasizes a subtle difference between "theory-directed" and "theory-informed." While in exploratory experimentation background theories are used "to set up experiments, generate data, and draw conclusions," such experiments "are not 'directed' by the aim to test, develop, or otherwise articulate an existing theory or hypothesis" (Waters 2007, 280). Laura Franklin makes a similar point that exploratory experiments are theory laden in terms of background knowledge, but not in terms of local theories (2005, 891).

These remarks closely resemble the previous discussion regarding external and internal theory-ladenness. The distinction between a phenomenological and a theoretical level is also helpful for the methodological analysis of data-intensive science, which supposedly concerns the phenomenological level regarding local, causal structure of phenomena but does not rise to the theoretical level.

An important difference between exploratory experimentation and data-intensive science is that in the former, data are usually of experimental nature, while the latter often deals with observational data. But this is largely

https://doi.org/10.1086/683328 Published online by Cambridge University Press

irrelevant from the perspective of a difference-making account of causation according to which experimental intervention has only pragmatic advantages over observational data. Another difference concerns the complexity of the phenomena. While mapping the causal structure by parameter variation is as old as science itself, carrying it out on the computer can address phenomena that were previously largely inaccessible to causal analysis. This new handle, which data-intensive science provides, for mapping the causal structure of highly complex phenomena will make all the difference to scientific practice.

REFERENCES

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

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

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

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.

Franklin, Laura R. 2005. "Exploratory Experiments." *Philosophy of Science* 72:888–99.

Gray, Jim. 2007. "Jim Gray on eScience: A Transformed Scientific Method." In *The Fourth Paradigm: Data-Intensive Scientific Discovery*, ed. Tony Hey, Stewart Tansley, and Kristin Tolle, xvi–xxxi. Redmond, WA: Microsoft Research.

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

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

Keynes, John M. 1921. *A Treatise on Probability*. London: Macmillan.

Krohs, Ulrich. 2012. "Convenience Experimentation." *Studies in History and Philosophy of Biological and Biomedical Sciences* 43 (1): 52–57.

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

Leonelli, Sabina. 2012. "Making Sense of Data-Driven Research in the Biological and Biomedical Sciences." *Studies in History and Philosophy of Biological and Biomedical Sciences* 43:1–3.

Mackie, John L. 1980. *The Cement of the Universe*. Oxford: Oxford University Press.

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

Norvig, Peter. 2009. "All We Want Are the Facts, Ma'am." http://norvig.com/fact-check.html.

———. 2011. "On Chomsky and the Two Cultures of Statistical Learning." http://norvig.com/chomsky.html.

Pietsch, Wolfgang. 2014. "The Nature of Causal Evidence Based on Eliminative Induction." *Topoi* 33 (2): 421–35.

Russell, Stuart, and Peter Norvig. 2009. *Artificial Intelligence*. Upper Saddle River, NJ: Pearson.

Skyrms, Brian. 2000. *Choice and Chance*. Belmont, CA: Wadsworth.

Steinle, Friedrich. 1997. "Entering New Fields: Exploratory Uses of Experimentation." *Philosophy of Science* 64 (Proceedings): S65–S74.

———. 2005. *Explorative Experimente*. Stuttgart: Steiner.

https://doi.org/10.1086/683328 Published online by Cambridge University Press

Vincenti, Walter. 1993. *What Engineers Know and How They Know It*. Baltimore: Johns Hopkins University Press.
von Wright, Georg H. 1951. *A Treatise on Induction and Probability.* New York: Routledge.
Wasserman, Larry. 2006. *All of Nonparametric Statistics*. New York: Springer.
Waters, C. Kenneth. 2007. "The Nature and Context of Exploratory Experimentation." *History and Philosophy of the Life Sciences* 29 (3): 275–84.

https://doi.org/10.1086/683328 Published online by Cambridge University Press