proceedings

MDPI

Proceeding Paper

Epistemological Considerations about Big Data and Prediction in Ecology †

Léo Trocmé--Nadal

Centre Gilles Gaston Granger, UMR 7304, Université d'Aix-Marseille-Site Schuman Maison de la Recherche, 29, Avenue Robert Schuman, CEDEX 1, 13621 Aix-en-Provence, France; leo.trocmenadal@outlook.fr

† Presented at Philosophy and Computing Conference, IS4SI Summit 2021, online, 12–19 September 2021.

Abstract: The aim of this paper is to analyze the philosophical implications of the techno-scientific promises and discourses that generally surround big data, but also to nuance their content with respect to the concrete uses of these technologies in the natural sciences. To achieve this, I will rely on a case study focusing on the hopes, as well as the effective practical and theoretical issues, that accompany the development of big data and numerical models within the ecological sciences.

Keywords: big data; ecology; prediction; techno-scientific promises; theory; inductivism; scientific practices; pluralism; knowledge; action

check for updates

Citation: Trocmé--Nadal, L. Epistemological Considerations about Big Data and Prediction in Ecology. *Proceedings* **2022**, *81*, 86. https://doi.org/10.3390/proceedings2022081086

Academic Editor: Peter Boltuc

Published: 29 March 2022

1. Introduction

The concept of big data was initially coined in the fields of digital industry and commerce at the turn of the 2000s [1,2]. It commonly refers to a set of massive accumulation technologies and powerful tools for the visualization and automated analysis of digital data. Due to these characteristics, big data are regularly called upon to revolutionize the practices and theoretical frameworks of scientific research. They are even called on to impose themselves as a new unique model for the production of scientific knowledge led by data sciences and data mining technologies [3]. The aim of this short paper is to analyze the philosophical implications of the techno-scientific promises that structure these types of discourse (as they are formulated, for instance, in: [4,5]), but also to nuance their content through a case study of the actual practical and theoretical issues that accompany the development of big data and numerical models in ecology.

2. Ecological Big Data Promises and Discourses

The discourses which commonly surround big data belong to what sociologists of science have called "technoscientific promises" [6]. The idea of techno-scientific promise refers to a set of discourses and coordination regimes of actors oriented towards the promotion of a technical innovation. The aim of this promotion is to justify substantial economic and human investments, given the supposed capacity of this innovation to respond to major societal problems [6]. In ecology, this economy of techno-scientific promises translates into the hope of developing a global and exhaustive knowledge of the biosphere and global biodiversity [1,7] and a more predictive science able to support sustainable natural resource management strategies and to respond to the challenges of ecological crises [8]. For the rest of this paper, I will call these techno-scientific discourses surrounding big data "the big data discourses" (BDD). Here, I will discuss one of the epistemological components that underlie these BDD: namely, the idea that big data technologies would lead to the advent of a single purely inductive and empirical scientific model, within which the work of theoretical reflection, modelling, or hypothesis formulation would become useless. The automated statistical processing of data, made possible by machine learning, for example, would make it possible to extract patterns and correlations inaccessible to the human brain and sufficiently robust to accurately predict the evolution of the systems under study—or so it is claimed. Thus, big data analytics and the predictions they produce could effectively guide action, without having to identify the causes of the observed dynamics.

To discuss this epistemological dimension of BDD, I will focus on some issues related to the status of predictions in ecology.

3. Big Data for Better Prediction, but without the Need for Better Understanding?

One of the main features of BDD is the focus on the capacity of artificial intelligence and machine learning tools to provide strong and reliable prediction about the phenomenon under study, thanks to the available computing power and the alleged completeness associated with the amount of data handled [8]. With this confidence in technological tools comes a corelative decrease in the attention paid to the theoretical dimension of scientific work. It was even said that, in the context of big data technologies, the "scientific method" (it must be understood as a hypothetico-deductive method) could become "obsolete" [4]. In ecology, the operationalization of these kinds of techno-scientific claims and promises is illustrated by several examples. It leads, for example, to the collection of big databases aiming at quantifying global biodiversity, as exemplified by the GBIF program [7]. Additionally, this kind of initiative comes with the underlying idea that this big amount of global biodiversity data could allow us to predict, and therefore to manage, the global responses of ecosystems to global change. Examples of the use of these big databases in ecology was the so-called "species distribution models" (SDM), which aim at forecasting the evolution of species distribution at a global scale, according to the evolution of climatic conditions linked to global changes [9,10]. These BDD about prediction in ecology also find an illustration in the growing works on the forecasting of "catastrophic shifts" [11] through the search, thanks to the complex systems mathematical and numerical modeling tools, of possible "early warning signals" [12]. Finally, one can also mention here the debates about "non-parametric models", which aims at predicting the evolution of an ecological dynamic (e.g., a fish population variation caused by fisheries) with machine learning tools that parametrize the predictive models automatically from the data. One common feature of these expressions of BDD in ecology is their focus on temporal prediction in order to improve action, with comparatively little attention paid to improving knowledge about the phenomenon under study [7,8]. As a result, these three examples of epistemological big data standard narratives have generated big debates. Big biodiversity databases, such as GBIF, generated epistemological debates about the quality and usability of the data considered for improving ecological knowledge goals, but also about the ontological, political, and axiological reliability of the predictions produced from this data, in the case of SDM, for instance (see, e.g., [7,9,10]). Furthermore, early warning signals generated epistemological debates about the specificity of the patterns extracted from the data on endangered ecosystems in relation to ecosystems that are not threatened by collapse because of "the frequency of false positives" [8]. The application of this type of model, generated from controlled conditions or relatively simple systems, to real and/or more complex systems, is also questioned. More generally, authors discuss the explanatory power of these early warning signals. Indeed, they are only based on correlations, although generated from big amounts of data, and so "they say nothing about underlying mechanisms" [8] that could explain the observed dynamics. Finally, predictions produced by non-parametric models have also generated debates, but I will analyze this example in more detail below because I think it is one of the most exemplary cases of BDD expression in ecological sciences.

As mentioned above, on the issue of modelling and data analysis, central topics for the BDD, the development of tools for the automated analysis of big ecological data have raised hopes and debates. These debates particularly focused on the place of theory in ecology according to the possibility to develop more predictive models of ecosystem dynamics. An exemplary illustration of these debates is provided by the literature on the use of non-parametric models to produce predictions in ecology. The promoters of this type of approach defend the idea that the use of relatively simple non-parametric models, whose parameters are determined automatically by algorithms "learning" from the data, offer better results for predicting the behavior of a data set than large simulation models with manual parametrization [13]. Therefore, in the mind of their advocates, this kind of scientific approach is built on the explicit idea that according to the temporal constraints that weigh on certain management decisions (e.g., fisheries management), it could be more relevant to rely on the computing power of the machine for guiding action than on improving ecological scientific understanding [13]. This approach thus reflects, in the field of ecology, one central idea of BDD mentioned above: that, with a lot of data and sufficient computing power, it would be possible to produce useful information for action without the need to understand and analyze the processes involved theoretically. Against this argument, it is argued, first, that the quantitative assessment of the reliability of these models is difficult due to the lack of available mathematical tools [14]. As in the case of EWS mentioned above, their qualitative evaluation of non-parametric models is not easy either, since this type of strategy does not allow for the identification of possible mechanisms, which would in turn allow for the assessment of the reliability and relevance of the modelled predictions. The hopes for improving the predictive power of the models generated by big data have also been analyzed more specifically from the perspective of the philosophy of science. Several authors have shown the importance of theoretical reflection to determine the expectations and confidence that can be placed in this type of tool, (1) whether it is used to improve knowledge or (2) to assist in decision making [8,15,16]. The necessity of this theoretical reflection endeavor is analyzed at two levels by the authors quoted above. Firstly, it is necessary to clarify what is meant by prediction. It is an ambiguous concept that is used alternatively in ecology, on the one hand, to refer to a method of corroborating hypotheses (according to a Popperian falsificationist's framework) or, on the other hand, to refer to the anticipation of possible scenarios of the evolution of the target system. Among other things, the authors highlight that the importance of this clarification to temporal prediction, in the case of decision-making support, is possible or not, according to the properties of the system under study [8,16]. Secondly, the improvement of an ecological theoretical framework and reasoning is necessary to clarify the limits of the numerical tools, the methodologies used, and what is known about the properties of ecosystems in order to specify what can be expected from big data and predictive models. Among the limits imposed by ecological processes on prediction, Maris et al. note, in particular, their contingent and evolutionary nature, their stochasticity, and also their complexity, which generates emergent phenomena that are difficult to predict. If these limitations are not considered from a theoretical point of view, decision makers and their scientific advisors take the risk of applying a predictive model to an ecological system which is fundamentally unpredictable, and thus of being misguided entirely on the actions required for the preservation of this system [8].

4. Are We Replaying the Classical 1950's Debate Surrounding the Demarcation Criterion and the Place of Theory in the Knowledge Production Process?

The previous section has shown that the development of big data in ecology is far from making theory useless, and that it is also far from leading to a purely inductive way of performing research. The authors quoted above show, on the contrary, that the improvement of an ecological theoretical framework and hypothesis-testing methods are even more necessary to take proper advantage of the amount of data accumulated and of the tools available to analyze it [7,8,17]. Thus, the debates generated by BDD, especially about the possibility of a purely inductive science, seem to renew some of the features of philosophy of science debates of the middle of the twentieth century, although in a significantly different way. Despite significant historical, contextual, technological, scientific, and theoretical (etc.) differences, it is interesting to note that authors who highlight the importance of theory, even more so in the context of big data in ecology, against the idea of a possible purely inductive science refer to a Popperian framework to frame their analysis [8,15,16,18,19]. Indeed, Popperian falsificationism was initially formulated against logical empiricists' inductivism [20]; thus, it seems interesting to consider this controversy to think about the current forms of inductivism raised by the development of big data.

However, the Popperian framing of the scientific method has been widely criticized since the end of the 20th century. Many authors put forward the fact that there is not a universal scientific method (e.g., hypothetico-deductive in the case of Popper) or demarcation criterion between science and pseudo-science [21], and that the objectivity criterion and evidence value could change across scientific cultures or styles [22,23]. Interestingly, philosophical works developing in the wake of the pluralist movement in the philosophy of science have also led to a questioning of the representational conception of data, which has run through the philosophy of science since Hempel and Popper, and which makes data stable, objective, and realistic representations of the phenomena under study [17]. Even more in the context of data-intensive science, it was shown that data features change over the time and during their circulation across several situations [1,17].

Therefore, in the case under study, the Popperian framework could be helpfully convocated, especially to express more precisely what one means by prediction (anticipation but also corroboration), but one must not erase all of the work conducted about the pluralist and perspectivist nature of science [22–24], especially in ecology and the life sciences [25–27], which have typically exemplified the diversity of scientific methods and objectivity criteria against the unique and evolutive view of science conveyed by Popper. Moreover, in the case of big data "neo-inductivism", what is at stake is not, as in Popper's time inductivism debates, discussing the right place of theory and experience in knowledge production process, but the legitimacy and usefulness of scientific theoretical reasoning itself.

5. Conclusions

Finally, these analyses of the topic of ecological prediction permitted me to show the relevance of theoretical work in scientific practices [7,8,15–19] against the BDD that tend to leave out knowledge improvement goals or even to conceive scientific theorization as obsolete. Theoretical work is essential to precisely express what we mean by prediction (a key issue for environmental sciences), which can refer, not only to an anticipation, but also to a corroboration activity. Scientific theoretical work allows us, then, to precisely express what could actually be expected from big data analysis when one talks about the improvement of models' predictive power, according, for example, to the available scientific knowledge about the predictability of the system under study. Indeed, without this philosophical and theoretical work, the risk is that we make big data "black boxes" [17] of which it could be hard to distinguish the potential fallacy, because of the complexity of the tools and the amount of various data involved, possibly produced from numerous methods and scientific cultures, making it hard for non-specialists of the field to assess or see the method considered [17]. Under the question of big data and scientific promises, it seems then that we find some features of the old Popperian question of demarcation between scientific and non-scientific knowledge [20]. However, this will be revisited under the following pragmatic form: that of understanding how scientific practices geared toward knowledge improvement goals distinguish and articulate themselves from/with action improvement goals, in connection with economic and political fields, for instance [7,8,17].

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: I would like to say all my gratitude to my thesis supervisor, Sébastien Dutreuil, for his multiple reviews, his precious advice, and his warm and constant support, during the writing of this article, but also since the beginning of my thesis work (and before).

Conflicts of Interest: The author declares no conflict of interest.

References

1. Devictor, V.; Bensaude-Vincent, B. From ecological records to big data: The invention of global biodiversity. *Hist. Philos. Life Sci.* **2016**, *38*, 13. [CrossRef] [PubMed]
2. Schmitt, E. Explorer, Visualiser, Décider: Un Paradigme Méthodologique Pour la Production de Connaissances à Partir des Big Data. Ph.D. Thesis, Université de technologie de Compiègne, Compiègne, France, 2018.
3. Kitchin, R. Big Data, new epistemologies and paradigm shifts. *Big Data Soc.* **2014**, *1*. [CrossRef]
4. Anderson, C. The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*, 23 June 2008.
5. Hey, T. *The Fourth Paradigm: Data-Intensive Scientific Discovery*; Hey, T., Hey, A.J.G., Stewart, T., Tolle, K.M., Eds.; Microsoft Research: Redmond, DC, USA, 2009.
6. Pierre-Benoît, J. On the economics of techno-scientific promises. In *Débordements. Mélanges offerts à Michel Callon*; Akrich, M., Barthe, Y., Muniesa, F., Mustar, P., Eds.; Presses des Mines: Paris, France, 2013; 407p.
7. Devictor, V. La Prise en Charge Technoscientifique de la Crise de la Biodiversité. Ph.D. Thesis, Université Paris 1 Panthéon-Sorbonne, Paris, France, 2018.
8. Maris, V.; Huneman, P.; Coreau, A.; Kéfi, S.; Pradel, R.; Devictor, V. Prediction in ecology: Promises, obstacles and clarifications. *Oikos* **2018**, *127*, 171–183. [CrossRef]
9. Beck, J.; Böller, M.; Erhardt, A.; Schwanghart, W. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecol. Inform.* **2014**, *19*, 10–15. [CrossRef]
10. Moudrý, V.; Devillers, R. Quality and usability challenges of global marine biodiversity databases: An example for marine mammal data. *Ecol. Inform.* **2020**, *56*, 101051. [CrossRef]
11. Scheffer, M.; Carpenter, S.; Foley, J.A.; Folke, C.; Walker, B.R. Catastrophic shifts in ecosystems. *Nature* **2001**, *413*, 591–596. [CrossRef] [PubMed]
12. Kéfi, S.; Dakos, V.; Scheffer, M.; Van Nes, E.H.; Rietkerk, M. Early warning signals also precede non-catastrophic transitions. *Oikos* **2012**, *122*, 641–648. [CrossRef]
13. Perretti, C.T.; Munch, S.B.; Sugihara, G. Model-Free Forecasting Outperforms the Correct Mechanistic Model for Simulated and Experimental Data. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 5253–5257. [CrossRef] [PubMed]
14. Jabot, F. Why Preferring Parametric Forecasting to Nonparametric Methods? *J. Theor. Biol.* **2015**, *372*, 205–210. [CrossRef] [PubMed]
15. Houlahan, J.E.; McKinney, S.T.; Anderson, T.M.; McGill, B.J. The priority of prediction in ecological understanding. *Oikos* **2016**, *126*, 1–7. [CrossRef]
16. Mouquet, N.; Lagadeuc, Y.; Devictor, V.; Doyen, L.; Duputié, A.; Eveillard, D.; Loreau, M. Predictive ecology in a changing world. *J. Appl. Ecol.* **2015**, *52*, 19. [CrossRef]
17. Leonelli, S. *Data-Centric Biology: A Philosophical Study*; The University of Chicago Press: Chicago, IL, USA; London, UK, 2016.
18. Clark, J.S.; Gelfand, A.E. A future for models and data in environmental science. *Trends Ecol. Evol.* **2006**, *21*, 375–380. [CrossRef]
19. Marquet, P.A.; Allen, A.P.; Brown, J.H.; Dunne, J.A.; Enquist, B.J.; Gillooly, J.F.; Gowaty, P.A.; Green, J.L.; Harte, J.; Hubbell, S.P.; et al. On Theory in Ecology. *BioScience* **2014**, *64*, 701–710. [CrossRef]
20. Popper, K. *The Logic of Scientific Discovery*; Repr. 2008 (twice); Routledge Classics; Routledge: London, UK, 2008.
21. Laudan, L. The Demise of the Demarcation Problem. In *Physics, Philosophy and Psychoanalysis*; Cohen, R.S., Laudan, L., Eds.; Springer: Dordrecht, The Netherlands, 1983; Volume 76, pp. 111–127.
22. Hacking, I. *Concevoir et Expérimenter: Thèmes Introductifs à la Philosophie des Sciences Expérimentales*; Bourgois, C., Ed.; Épistémè Essais: Paris, France, 1989.
23. Knorr-Cetina, K. *Epistemic Cultures: How the Sciences Make Knowledge*; Harvard University Press: Cambridge, MA, USA, 1999.
24. Giere, R.N. *Scientific Perspectivism*; University of Chicago Press: Chicago, IL, USA, 2006.
25. Callebaut, W. Scientific Perspectivism: A Philosopher of Science's Response to the Challenge of Big Data Biology. *Stud. Hist. Philos. Sci. Part C* **2012**, *43*, 69–80. [CrossRef] [PubMed]
26. Dupré, J. *The Metaphysics of Biology*; Elements in the Philosophy of Biology; Cambridge University Press: Cambridge, UK, 2021.
27. Levins, R. The strategy of model building in population biology. *Am. Sci.* **1966**, *54*, 421–431.