TESIS DOCTORAL

2020

THE RISE OF THE LEARNING MACHINES

EZEQUIEL LÓPEZ RUBIO

PROGRAMA DE DOCTORADO EN FILOSOFÍA

EMANUELE RATTI, NOTRE DAME UNIVERSITY
DAVID TEIRA SERRANO, UNED

Acknowledgments

To Rosa, for giving me her tender love and cheerful support, which go beyond my wildest dreams.

To Ada, for illuminating my days from the very moment that I heard her heartbeat for the first time.

To Meli, for keeping our friendship against all the odds.

To my parents, for giving me the affection and values that I will keep forever.

To my brothers, sisters-in-law, nephews, and nieces, for all the moments that we have shared.

To my uncle Ezequiel and my aunt Paca, because they believed in me and helped me when I needed it.

To Juanmi and Fanny, for their true friendship.

To Pepe Muñoz, for supporting me in every imaginable way.

To David and Emanuele, for sharing their wisdom and enthusiasm with me.

To my colleagues, because it is a privilege and a pleasure to work together.

To my students, who have taught me so much.

Contents

Introduction

"S'il se trouvait un perroquet qui répondît à tout, je prononcerais sans balancer que c'est un être pensant..."
[If they find a parrot who could answer to everything, I would claim it to be an intelligent being without hesitation]
– *Denis Diderot (1746), Pensées Philosophiques, XX. Œuvres complètes de Diderot, Texte établi par J. Assézat et M. Tourneux, Garnier, 1875-1877.*

Since its inception in the mid-twentieth century, Artificial Intelligence (AI) has experienced fundamental changes in its core concepts and assumptions. After six decades, it has found its place among the scientific disciplines which generate the most significant public excitement and interest. It is routinely found in the lists of fields that researchers would like to work on, along with molecular biology (Russell & Norvig, 2009). It has also inspired an essential part of the science fiction genre. This interplay between AI and fiction has driven some of the public attitudes and expectations for it. In particular, this might be one of the causes that many people assume that the capabilities of current intelligent systems match those seen in popular culture. In any case, there is some truth to it since AI has already changed some aspects of our lives, and it is likely that it will continue to do so. There is a general awareness about this, and even some fear about its potential adverse consequences (Hawking, Russell, Tegmark, & Wilczek, 2014).

It is widely acknowledged that AI is a branch of Computer Science. However, its foundations were laid on ideas that come from Antiquity and belong to almost all fields of science. Its ultimate goal is to replicate human intellectual activity on a machine, so any human experience is somehow related to it. AI is aimed to reproduce all human knowledge. Hence the study of its nature poses a considerable challenge. It is necessary to identify which scientific disciplines and traditions contribute to each AI subfield or school of thought. This way, a good understanding of this broad subject could be achieved.

At this point, it is essential to highlight that our focus is not on the question of whether AI would attain its ultimate objectives, but how it struggles towards them. In other words, we are interested in scientific methodologies and their interplay. This way, we depart from the most common topics in the philosophy of AI. However, a summary of them is outlined next, along with their connections to our purposes.

The primary debate is about the possibility of a strong AI, as started by Searle (Searle, 1980). The strong AI hypothesis asserts that a computer can have a mind;

such a machine would be called a general artificial intelligence. On the other hand, the weak AI hypothesis states that it can only act like it had one (Russell & Norvig, 2009). The Chinese room argument says that we could put a human with no knowledge of Chinese inside a room with an extensive supply of books that specify how to produce appropriate answers in Chinese to questions in Chinese. The argument continues by stating that such a room would have no real understanding of Chinese, just like any computer acting as if it understood that language. Aside from the question of whether the argument is correct, it is important to notice that it assumes a specific model of the human mind. In that model, mental activity is seen as a symbol manipulation process, where understanding and intention are defined with respect to abstract concepts such as words and sentences. AI followed this model for some decades. Later developments have led to the abandonment of this mental model in AI research. Consequently, the most important philosophical questions about AI are related to the reference model of the human mind that AI researchers choose.

The concept of levels of abstraction in Computer Science is related to this choice of the elementary pieces of information that AI should process. Computer systems contain billions of information processing elements that interact in various ways. Several levels of abstraction are defined to manage such complexity. For example, we can describe a car at four different levels, at the very least. At a high level, a car is a wheeled vehicle that carries people and cargo. At a middle level, a car is a mechanical device that is made of four wheels, an engine to provide propulsion, five seats for the driver and four passengers, etc. At a low level, we would have to describe every single nut, bolt, and cable that the car contains, along with its precise function. At a very low level, the mechanical and chemical interactions among molecules and atoms, governed by the laws of physics, would come into play.

As seen, high level descriptions are short and easy to understand, but not specific. On the other hand, low level descriptions are very long and precise since they include all the details, but they are somewhat difficult to grasp because one can not see the forest for the trees. Computers have many more elements than any mechanical device, so several abstraction levels are necessary. At the lowest level, that might be called the electronic level, there are logic gates which process binary bits (0/1). Then, at the architectural level, there are functional units such as adders, multipliers, or registers that store 32 or 64 bits. At the operating system level, there are hard disks, printers, random access memories, and processes that use them. At the application level, there are word processors, database managers, and Internet browsers. At the network level, there are computers (hosts) that exchange information. One of the biggest challenges of AI is that the abstraction levels of the human brain are not so well understood, and choosing the correct level seems to be essential to its success.

The most popular approach to AI is nowadays data driven AI, also called subsymbolic, since it processes data and not symbols (Goertzel, 2012). It is rooted in connectionism, which holds the view that most of the information processing in the

human brain is done without symbols (J. Anderson et al., 2004; Smolensky, 1988). A good deal of data driven AI falls under the denomination of computational intelligence (van Eck, Waltman, den Berg, & Kaymak, 2006). Its reference model is statistics. Under this paradigm, the starting point is to establish a practical and concrete problem that admits an objective and quantitative performance measure. The key concept is that of machine learning (Kaelbling, Littman, & Moore, 1996; Schapire, 1990; Xu & Wunsch II, 2005). The machine learns from data to enhance its performance progressively. The ultimate goal is that the machine can choose the most suitable actions to maximize the performance, no matter if its internal workings are not similar to the mental processes of humans.

Subsymbolic AI also called connectionist AI, appeared as an alternative to overcome some of classic AI limitations such as brittleness concerning programming or data errors, inability to learn from experience, and the lack of a plausible connection to biological structures. Connectionism also avoids Searle's Chinese room argument against symbolic AI since no symbols are manipulated (Davenport, 2013).

Machine learning has achieved considerable success in solving specific problems. Among the most popular applications, we can mention expert systems, bioinformatics, machine translation, speech recognition, optical character recognition, computer vision, and medical diagnosis (Cristianini, 2010). However, the learning process is not easy to understand by a human (black boxes), and meanings and symbols are absent from the process. Overall, its focus on specific applications is far away from the general AI ideal, with cognitive abilities similar to those of human beings (Cass, 2011; Norvig, 2012). The main questions are whether there are limits to learning in terms of performance and whether machines could perform better than humans in some tasks. The present thesis focuses on the analysis of machine learning from the perspective of the philosophy of science.

The structure of this thesis is as follows. The hypotheses and the theoretical framework are detailed in chapter 1. In particular, the main hypothesis and the goals to be attained are described in Section 1.1. Then the theoretical framework that is considered for this work is outlined in Section 1.2. The methodology of data driven science is explained in Section 1.3. This methodology leads us to the distinction of two aspects in scientific theories, as seen in Section 1.4. Then, two kinds of data science are described in Section 1.5. After that, the published papers which support the thesis are included. General conclusions are extracted in chapter 2. Finally, a report of the publications which support this thesis is given in chapter 3.

1 Hypotheses and theoretical framework

"For if every instrument could accomplish its own work, obeying or anticipating the will of others, like the statues of Daedalus, or the tripods of Hephaestus [...] If, in like manner, the shuttle would weave and the plectrum touch the lyre without a hand to guide them, chief workmen would not want servants, nor masters slaves."
– Aristotle, Politics, Book I, Part IV.
Translated by Benjamin Jowett, Batoche Books, 1999.

1.1 Hypotheses

In November 1970, a featured article in the National Geographic magazine reviewed how deeply computers had transformed a myriad of human activities (White, 1970). Significantly enough, scientific research was not among them. Computers had two main advantages. The first one was that they made calculations very fast, and the other one was that they could store large amounts of data and retrieve them quickly. This did not mean much for the scientific method. Scientists had been using calculators and libraries for centuries, so there was no indication that computers could change the criteria that scientists followed in their activities.

At that time, AI was in the classic period. The sources of classic AI can be traced to mathematical logic, which was already a mature field. AI soon developed its own methodology, which departed largely from logic due to the applied science challenges that AI had to face, which are not present in mathematical logic. It must be pointed out that the classic approach became generally accepted by the AI community until the 1980s. Therefore it provided the foundation for normal research activities during the classic period, which comprises the 1960s and 1970s. The physical symbol system hypothesis summarizes this approach (Newell & Simon, 1976, p. 116):

> A physical symbol system has the necessary and sufficient means for general intelligent action.

where a physical symbol system is to be understood as symbol manipulation software.

Classic AI approaches processed non numerical, discrete symbols by inference rules based on logic. This is why artificial neural networks are sometimes called 'subsym-

bolic', which means that numbers (data) are at a lower level than symbols (concepts). An early split between both approaches is widely acknowledged (Nilsson, 2010, section 2.3):

> With a firm belief in the symbol system hypothesis, some people began programming computers to attempt to get them to perform some of the intellectual tasks that humans could perform. Around the same time, other researchers began exploring approaches that did not depend explicitly on symbol processing [...]. A split between symbol-processing methods and what has come to be called "brain-style" and "nonsymbolic" methods still survives today.

Nilsson places a turning point in the 1980s when the Good Old Fashioned AI (GOFAI) lost the confidence of AI researchers. Interestingly, he thinks that the criticism of outsiders was not essential to this change (Nilsson, 2010, chapter 25):

> Frustrated with AI's slowdown, people with different approaches to AI eagerly stepped forward to claim that what AI needed was more of this or that alternative to AI's reigning paradigm – the paradigm John Haugeland called "good-old-fashioned AI" or GOFAI. GOFAI, of course, had as its primary rationale Newell and Simon's belief that a "physical symbol system has the necessary and sufficient means for intelligent action." But GOFAI seemed to be running out of steam during the 1980s, making it vulnerable to challenges by AI researchers themselves – challenges that had to be taken more seriously than those of Searle, Dreyfus, Penrose, and others outside of the field.

As pointed out by Nilsson, one of the most influential critiques of AI from philosophy came from Herbert Dreyfus (Dreyfus, 1972, 1992). He argued that AI had failed to produce an artificial general intelligence because researchers assumed that all human knowledge could be reduced to symbolic formalizations and rules. An early response to this attack was that of Buchanan (Buchanan, 1972), who conceded that the richness of human experience might be difficult to formalize, but argued that there was no way to find out which framework was better. However, neither of them could foresee the new perspectives which ultimately replaced classic AI. Consequently, it can be said that Dreyfus was able to spot the shortcomings of the classic approach, but neither he nor his detractors set forth any practical solutions.

This state of things changed when computers were programmed to enhance their performance when carrying out some task by learning from data. Then the digital era put an exponentially growing amount of information at the disposal of those programs. In a matter of decades, many scientific datasets have grown so large that a human can no longer analyze them without the help of learning algorithms. But most of those algorithms do not produce explanations that scientists can understand. Even worse, they do not care much about causes when they are asked to make predictions. Despite this violation of the established laws of scientific inquiry, some

scientists have embraced them. There is a widespread perception that something essentially novel has happened to science (National Science Board, 2005, p. 49):

> It is exceedingly rare that fundamental new approaches to research arise. Information technology has ushered in such a fundamental change. Digital data collections are at the heart of this change.

A scientific discipline has emerged from these transformations, which has many names. Through this thesis, it will be called machine learning, although other denominations will also be discussed. There is an urgent need for an epistemological study of its status since it is transforming the way that scientists create and validate models. The data revolution was not evident until the very end of the 20th century, so many established theories of science do not address its specific characteristics. Algorithmic predictors are validated due to their adequate performance (Agarwal and Dhar 2014, p. 446; Dhar, 2013, p. 66). However, in many cases, good machine learning predictive models do not provide explanations.

The central hypothesis of this work is that machine learning has challenged some fundamental concepts in the philosophy of science. The specific goals are the following:

- Introduce a new concept of model simplicity in machine learning, i.e., computational simplicity. It is further argued that such a concept is more encompassing and closer to actual machine learning practices than the classic ones. This is carried out in (López-Rubio, 2020).
- Investigate the tradeoff between prediction and explanation in the application of machine learning to molecular biology. This is developed in (López-Rubio & Ratti, 2019).
- Propose a novel kind of computational functionalism, called neural computational functionalism. Neural computational functionalism addresses the specificities of a new sort of machine learning, namely deep learning. This is done in (López-Rubio, 2018).

1.2 Theoretical framework

Machine learning is a branch of the artificial intelligence subfield of computer science that is devoted to the development of computer systems, which can improve their performance progressively as more samples of the problem at hand are provided. Performance is quantitatively measured. Data science is the discipline which studies the acquisition, storage, manipulation, and exploitation of digital data. Machine learning is employed in many engineering applications not directly related to data science, such as computer vision, robotics, speech recognition, and automated system control. Nevertheless, its techniques, i.e., the learning algorithms, are at the heart

of most data science procedures (Jordan & Mitchell, 2015, p. 255). This is because learning algorithms typically enhance their performance as more data are available.

Machine learning spans a wide range of techniques, but most of them are black boxes (Frické, 2015, p. 655). A machine learning algorithm is called a black box, or that it is opaque when it does not yield an explanation of its output that can be understood by a human. That is, the vast majority of machine learning algorithms do not generate an understandable model of reality. The balance between explanation and predictive accuracy will be later be explored in this thesis, as detailed in (López-Rubio & Ratti, 2019).

Most machine learning algorithms fall into three categories: supervised, unsupervised, and reinforcement learning. Supervised learning algorithms require a set of input (independent) variables, a set of output (dependent) variables, and a dataset comprised of examples of observed inputs and their corresponding outputs. They build models of the relationships among those variables which predict the outputs given the inputs as accurately as possible. Unsupervised learning does not distinguish between input and output variables but aims at finding out the structure of a dataset. In some cases, this involves the automatic discovery of clusters in the data, and in other cases, the algorithms find correlations or dependencies among variables. Reinforcement learning intends to modify the behavior of a man made system which receives rewards from its environment according to its behavior so that the rewards are as high as possible.

Among these three kinds of machine learning, reinforcement learning is the least relevant for this work, since it is associated with engineering applications. Its goal is the enhancement of the system behavior and not the knowledge of the environment. On the other hand, supervised learning can be directly associated with the scientific quest of producing models of reality which yield correct predictions. Unsupervised learning does not generate predictors, but it is used by scientists to examine the structure of datasets, which would otherwise be too complex to be studied. My main concern in this thesis will be the automated prediction by supervised learning, although I will also discuss scientific applications of unsupervised learning. Many machine learning algorithms are found under the name "statistical learning", in particular when mathematical proofs of their properties accompany them. This highlights the location of data science at the crossroads between statistics and computer science. Data science looks for algorithms that are both statistically sound and computationally efficient.

The meaning of the term "data science" has changed over the decades (Press, 2013). In the 1960s, it was seen as a branch of statistics which was aware of the capabilities of computers. Nowadays, it aims to automatically discover complex relations among variables, also called patterns, even if these patterns do not convey a deep comprehension of the process under study. Dhar offers the following definition (Dhar, 2013, p. 64):

> Data science is the study of the generalizable extraction of knowledge

from data.

In this definition, "generalizable" means that the patterns which are automatically extracted from the available data are expected to occur in other situations. At this point, it must be highlighted that the discovered patterns can be employed to build a predictor by supervised learning or reveal the structure of the dataset by unsupervised learning. Given these considerations, data science helps scientists from other disciplines in two distinct ways. Either computers are employed to automatically generate models of phenomena, such that those models can yield accurate predictions of their behavior; or computers automatically discover the dependencies among the variables involved in a phenomenon, to enable scientists to study its structure.

From the standpoint of the philosophy of science, some further distinctions are relevant. Data-driven and hypothesis-driven science are two distinct approaches to the scientific method, which are frequently conceived as an opposing pair (Leonelli, 2012). The received hypothesis-driven approach starts with a scientific theory that explains a certain phenomenon and makes testable predictions. Then experiments are designed to evaluate whether the prediction is fulfilled in a controlled environment. On the other hand, the data-driven scheme starts with a dataset that might have been collected under real world conditions with limited or no experimental control. Then the data-driven scheme uses data science methods to analyze the dataset. Consequently, data science must be distinguished from data-driven science, since data-driven science is what scientists from other disciplines do when they carry out their research activities under the data-driven methodology with the help of data science methods. This means that data science methods can also be used under the hypothesis-driven methodology, although they play a less important role. The term "data analysis" is rather generic, since it is associated to both approaches (D. R. Anderson, Burnham, & Thompson, 2000, p. 913; D. R. Anderson, Burnham, Gould, & Cherry, 2001, p. 311; Breiman, 2001, p. 210; Cleveland, 2001, p. 21; Waller & Fawcett, 2013, p. 77; Sundararajan, Provost, Oestreicher-Singer, & Aral, 2013; Van Horn & Toga, 2014, p. 328; Frické, 2015, p. 653), whereas "big data analysis" is related to the novel one (Chiang, Goes, & Stohr, 2012, p. 2; Stephens et al., 2015, p. 6; Dolinski & Troyanskaya, 2015, p. 2576).

Exploratory research is a third kind of research that has some resemblances to the data-driven approach. Namely, exploratory research is not strongly guided by a theory that must be put to the test (Haufe, 2013, p. 366). In other words, the experiments are not always designed to test theories (Schickore, 2016, p. 21). In some cases exploratory research can be seen as a hybrid of hypothesis- and data-driven research which poses a wide range of possible hypotheses, then tries to rule out as many of them as possible with the help of data science methods, and finally designs experiments to evaluate the remaining ones (Ratti, 2015, p. 200). However, exploratory research is not always related to data-driven science or data science. It can be employed to search for regularities in the physical world when no established theory exists, which has been done well before computers were available (Schickore, 2016, p. 21).

At this point, an critical remark must be made. It has been argued that data-driven science dates back to the times of Linnaeus (Müller-Wille & Charmantier, 2012, p. 6) so that the main differences with current times are the size of the datasets, the fact that they are analyzed by researchers from other disciplines, the change towards quantitative judgment criteria, and the widespread sharing of data (Strasser, 2012, p. 86-87). I contend that there is another fundamental difference in current data-driven science. Computers are not only used as massive data storage devices or fast calculating machines. They are also employed as learning machines that exploit the data to automatically generate algorithmic models of phenomena with predictive capabilities. This way of using computers is entirely new, and it was not possible in the early stages of computer science until machine learning developed in the 1960s and 1970s.

1.3 Methodology of data-driven science

Data are always collected within a theoretical framework, so predictive model generation is only a part of the overall task of theory construction and evaluation. Leo Breiman has argued that there are two statistical cultures, the classic one which emphasizes the explainability of the models, and a modern one which is more concerned about the predictive accuracy (Breiman, 2001, p. 209). This theory leads us to a fundamental methodological question that underlies data science and its supporting role in other scientific disciplines: is it admissible, from the point of view of the scientific method, that explanations of phenomena are not the main research goal? In other words, what is more important, the accuracy of the predictions, or the understanding of nature? There is no unique answer. It has been argued that data science produces scientific theories since it generates models that make good predictions (Dhar, 2013, p. 66; Agarwal & Dhar, 2014, p. 446). However, hypothesis-driven science assumes that explanations of phenomena are a key ingredient (Burian, 2007, p. 289; Karaca, 2013, p. 94; Schickore, 2016, p. 21). It has also been said that the ultimate goal of science is the comprehension of nature, so that black boxes are not allowed (Efron, 2001, p. 219). Nevertheless, the role of hypotheses has decreased in importance in current science, which means that hypothesis-driven science is not the only possible strategy (Krohs, 2012, p. 53).

Against those who argue that black boxes are inadmissible, the rejoinder would be that the process under study is too complicated for a human to understand it, at the current state of the scientific theory of the domain at hand. Otherwise, an understandable model derived from the theory would fit the data equally well or better than black boxes. The scientist must resist the temptation to oversimplify the problems (Breiman, 2001, p. 204):

> With data gathered from uncontrolled observations on complex systems involving unknown physical, chemical, or biological mechanisms, the a priori assumption that nature would generate the data through a

> parametric model selected by the statistician can result in questionable conclusions that cannot be substantiated by appeal to goodness-of-fit tests and residual analysis. Usually, simple parametric models imposed on data generated by complex systems, for example, medical data, financial data, result in a loss of accuracy and information as compared to algorithmic models.

Some data sets are so large and complex that it would be almost impossible for a human to find significant patterns without the help of algorithms which automatically find models to fit the data (Dhar, 2013, p. 68). Moreover, experimental design imposes that uncontrollable sources of variation are reduced as much as possible, but this means that many experiments are conducted under oversimplified, unrealistic conditions. On the other hand, data science can exploit real world observations to generate predictors that work in practice (Dhar, 2013, p. 72-73; Agarwal & Dhar, 2014, p. 444), although the lack of experimental control means that the underlying processes might not be adequately understood. Experimental data are preferable when the goal is an explanation because causal relations can be ascertained without interferences. At the same time, accurate prediction requires observational data free of experimental control when those interferences significantly affect the outcome of the process in real situations (Shmueli & Koppius, 2011, pp. 562-563). A middle way is to employ the algorithms to discover interesting patterns in observational data, then find a theory which explains them, and finally test it with specifically designed experiments (Agarwal & Dhar, 2014, p. 446; Dolinski & Troyanskaya, 2015, p. 2577; Peters et al., 2014, p. 3; Ratti, 2015, p. 200).

Therefore, my answer to the above questions is twofold:

- If there is a developed scientific theory from which interpretable models of phenomena can be derived whose predictions match the data, then the problem falls outside of the field of algorithmic models. Even if the predictions do not match the data, but there is hope to solve this issue by a suitable modification of the theory (Cox, 2001, p. 218), then interpretable models are preferable because they can give the scientists the insights required to carry out such modification successfully.
- If there is little hope to obtain a scientific theory with strong predictive power, or there is a need to get accurate predictions without having to wait until such a theory is found, then data science and its black boxes can help. This case is more likely to happen in disciplines where the system under study is complex or chaotic. For example, molecular biology, epidemiology, and social sciences are amenable to black boxes because the underlying processes involve a vast number of relevant variables with highly nonlinear dependencies among them. Moreover, it is not feasible to measure all of these variables, i.e., many pertinent variables are hidden. This means that interpretable models would likely be oversimplifications of limited predictive performance. Existing understandable theories would still be employed whenever the focus is on the

general understanding of the processes, and not in the generation of accurate predictions.

In other words, the role of machine learning in contemporary science crucially depends on the complexity of the process under study, since this complexity is related to the possibilities to obtain an interpretable model that captures the intricacies of the problem at hand.

Another methodological problem is the multiplicity of models of very different structure which fit the data well (Tukey, 1977, p. viii; Breiman, 2001, p. 203-204). Which one gives a faithful account of reality? In the traditional hypothesis-driven methodology, the model which best fits the data would be chosen as the one which yields the best explanation of phenomena; the others would be rejected. But when the differences among the fitting errors are not statistically significant, it can be said that none of the models entirely reflect the underlying mechanism of the phenomenon under study. Under these circumstances, we can no longer dismiss black box procedures since one could have two or more white boxes that fit the data equally well, each providing an alternative explanation corresponding to a partial perspective (Callebaut, 2012, p. 76). Explanatory and predictive models can be regarded as different perspectives of reality. The former focus on the causal relationships among the variables of a problem and the laws behind those relationships, while the latter focus on the relation among the input and output variables. These perspectives overlap to the degree that they employ the same variables to describe a phenomenon. Let us consider a specific example. Decision trees are among the best known white box algorithmic prediction methods. A person can easily interpret them with basic knowledge about the problem domain at hand (Breiman, 2001, p. 207). Therefore, they are suitable for a perspective from the explanatory point of view. However, if our emphasis is in prediction, an ensemble of decision trees (a random forest) usually delivers higher predictive performance at the expense of the interpretability of the model, even though the variables are the same that a single decision tree employs.

1.4 Conceptual and mechanistic aspects of scientific theories

It is widely acknowledged that observations and hence data are theory laden (Callebaut, 2012, p. 74; Frické, 2015, p. 652). This has sometimes led to downplay the role of data (Leonelli, 2015, p. 814) or reject the possibility of data-driven science (Pietsch, 2015, p. 908; Ratti, 2015, pp. 198-199). In what follows, I defend a way to resolve this confrontation. Two parts can be distinguished in scientific theories. The first one, which might be called the conceptual one, contains the abstract concepts that are regarded as relevant to the phenomenon under study, their operationalization into measurable variables, and the experimental and observational procedures required to collect scientific data. This part is common to both data-driven and

hypothesis-driven science. Their differences only arise in the second part, which I call "mechanistic". This part establishes how the variables are related and how predictions can be obtained from those relations. Here hypothesis-driven research derives the mechanism from a set of hypotheses, while data-driven research learns the mechanism from data. Some philosophers of science have already posed this split of scientific theories into two parts. Wolfgang Pietsch argues that data-driven science is theory-laden externally, but not internally, which means that it does not make hypotheses about the internal structure of the phenomenon (Pietsch, 2015, p. 913-914). On his part, Emanuele Ratti distinguishes background assumptions that both kinds of research must follow (Ratti, 2015, p. 202). These assumptions are the conceptual part of a theory, which guide the data collection process but are not relevant to the mechanistic part in data-driven research. Laura Franklin highlights that there is a theoretical background upon which particular hypotheses or mechanisms are postulated by hypothesis-driven science, but not by exploratory experimentation (Franklin, 2005, pp. 892-893).

Let us be more precise. I will focus on prediction tasks. For the purposes of this discussion, a problem is a set of input variables (available at the time that the prediction must be delivered), a set of output variables (to be predicted, not available at prediction time), a set of samples (previously observed input-output pairs), and a set of real world situations where the dependency among inputs and outputs can be assumed to be the same. For example, predicting the price of a house in Los Angeles is a problem. The set of input variables could be the population density, the surface area of the house, the distance to the nearest hospital, the distance to the beach, the criminality rate of the area, etc. The set of output variables would only contain the price of the house. The samples would be collected all over Los Angeles, and it would be assumed that the relation among the input variables and the price of the house is the same across the city. The key difference between data science and hypothesis-based science is that a machine learns data science models for a single problem, and they can not be applied to any other problem. For example, we can not apply a model for Los Angeles house prices to Miami, because the relations among the variables are not the same. One could say that a model could be learned that predicts the price of the houses for both cities by taking samples from both of them. But this would lead to a reduction of the prediction performance because it would increase the complexity of the relationship that has to be learned. It would give better results to learn two models, one for each city. However, one can not reduce the size of the problem indefinitely, because, at some point, the number of available samples would be too small to learn. That is, one can not reduce the scope of the housing problem to a tiny neighborhood in Los Angeles because the sample would be too small to learn any significant relationship among the variables. Here "significant" means that the relation is not a random coincidence due to the small number of collected samples. Therefore, in machine learning, there is a fundamental balance: the bigger the scope of the problem, the more samples that can be collected (which increases the significance of the discovered relations),

but the more complex the relations which must be found. If we were to apply a general (hypothesis-driven) economic theory, we would say, for example, that all other things being equal, larger houses are more expensive. This principle would apply to any city. But the predictive performance of a model based on these general principles would likely be lower than that of a data-driven one. Now we can see the difference between data-driven and hypothesis-driven science: the former tries to solve each problem separately without relying on theoretical considerations to restrict the range of possible dependencies among the variables. On the other hand, hypothesis-driven science aims to find general principles about the dependencies among the variables that apply to many related problems, and it is the task of the scientist to derive specific models for each problem from those general principles. If the relations among variables for a certain set of problems are simple enough that scientists can understand them, create a general theory that comprises those relations, and derive models for specific problems, then hypothesis-based science has the key advantage that its theories can be applied to many problems without collecting so many data for each one. Only the data necessary for ascertaining the boundary conditions and a few problem specific parameters would be needed. Otherwise, data science is the only way to go, at the expense that a large amount of data must be collected, and a model has to be learned for each problem. This is the kind of balance that underlies the debate about the data-driven and hypothesis-driven approaches to cancer research (Golub, 2010; Weinberg, 2010). If the goal of this research were the understanding of the biological causes of cancer, then data-driven methods would be less used because they do not lead to a general theory of cell mechanisms associated with this disease. Here it must be remembered that some deny that such a theory can be found (Burian, 2007, p. 305). Data science is used because the real goal is to cure cancer, no matter if we fully understand its causes. Therefore, it is enough if data science can find one or more suitable target proteins to halt a specific cancer process. There are some disciplines such as metabolomics where the extreme complexity of the systems under study is widely acknowledged by the scientific community, so that oversimplified models which try to understand the influence of every single variable are rejected by researchers (Levin, 2014, p. 565). In those disciplines, scientific activity is focused on algorithmic models that capture the overall relation among hundreds of variables and their relative importance. Still, it can not give a detailed account of the role of each variable in the system (Levin, 2014, pp. 557, 561).

At this point, it is advantageous to clarify what is meant when it is said that an algorithmic model is a black box. This does not mean that it is some sort of magic crystal ball which gives the correct answers, but no one knows why. On the contrary, algorithmic prediction models are rooted in a mathematical background, which explains why they can generate good predictions from data. For example, in the Bayesian classification framework, it is possible to mathematically prove the convergence of a classifier to the maximum possible performance as the dataset size grows to infinity (Dalton & Dougherty, 2013a, 2013b; Dalton & Yousefi, 2015). The

issue here is that these models transform and combine the input variables of the problem in complex ways. Hence the prediction mechanism can not be translated in terms of the original variables to produce an understandable explanation. The situation is similar in unsupervised learning: complex relations among variables can be found, but their exact form can not be determined. As mentioned before, in some disciplines this complexity in the models can be an inescapable consequence of the complexity of the phenomena under study.

In general terms, data science is more important in those disciplines more oriented to prediction, as opposed to explanation (Ratti, 2015, p. 211). Following Shmueli's notation (Shmueli, 2010, p. 293), for a given problem there is an abstract level of reasoning where the existence of an abstract function $\mathcal{F}$ is postulated, which relates some abstract concepts $\mathcal{X}$ to other abstract concepts $\mathcal{Y}$, $\mathcal{Y} = \mathcal{F}(\mathcal{X})$. Then an operationalization must be carried out to translate these abstract concepts into measurable variables X, Y linked by the function f, which captures the exact relation among them, $Y = f(X)$. The abstract level of reasoning and its operationalization belong to the conceptual part of the scientific theory. Therefore, they are required by both the hypothesis-driven and the data-driven research approaches. The last step is the construction of a model $\hat{f}$, which is an approximation of f because it can not capture all the details of the phenomenon under study. Data science comes into play whenever the abstract function $\mathcal{F}$ is too complex to be ascertained by a human (Pietsch, 2015, p. 915; James, Witten, Hastie, & Tibshirani, 2013, p. 19), or it can not be used to derive a suitable predictive model $\hat{f}$. Data science proceeds by learning $\hat{f}$ from examples of pairs X, Y (James et al., 2013, p. 17). Data-driven research considers that $\hat{f}$ is an acceptable model of reality until a better one is found, even if $\hat{f}$ does not provide an explanation about how f works. The criteria that data science advocates to determine which model of reality is the best will be investigated below. The traditional, hypothesis-driven approach works differently. First of all, the abstract function $\mathcal{F}$ is inferred from a general theory. Then a predictive model $\hat{f}$ with few adjustable parameters is built. This typically involves some simplifications of f which preserve the explanatory value of $\hat{f}$, i.e., $\hat{f}$ can be employed to explain how f works. Finally, statistical parameter estimation or hypothesis testing is used to evaluate how well the data fits the model $\hat{f}$. Point estimations or confidence intervals can be obtained for the few adjustable parameters of the model. Alternatively, if an interesting phenomenon should affect some model parameters, then statistical hypothesis testing can assess whether the null hypothesis that there is no effect can be rejected with some confidence.

Now it is time to consider how data science compares models $\hat{f}_i$ to choose one. Automatable criteria are preferred since it is common that the computer must make this kind of choice many times during the learning process. Most model performance criteria try to find a balance between the complexity of the model and how well it fits the data (Bishop, 2006, pp. 32-33). At the heart of this issue is the bias-variance tradeoff. A well known performance criterion is the expected error that is committed when using a model $\hat{f}$ for prediction. This error is the sum of three non negative

quantities (James et al., 2013, pp. 33-35; Hastie, Tibshirani, & Friedman, 2009, pp. 223-228): the irreducible error, which stays the same no matter how accurate the model is; the bias, which is caused by the difference between the true function f and its approximation $\hat{f}$; and the variance of the learning method, which is caused by the oscillations of $\hat{f}$ as the dataset is changed. Thus the model which yields the best predictions is that which attains the minimal sum of bias and variance. This is hard to achieve because, in most cases, models with low bias have high variance and vice versa. Low complexity models suffer from high bias because they are oversimplifications of the true function f; this effect is called underfitting. On the other hand, complex models exhibit a large variance because they focus on irrelevant details of the data (overfitting). Before the advent of big data, there was no way to escape this dilemma (Geman, Bienenstock, & Doursat, 1992, p. 51). The availability of large datasets means that the variance can be diminished as the dataset size increases, which helps in reducing the overall prediction error. The key lies in the complexity of the true function f, which can be seen as a measure of how abrupt and intricate the changes in the output variables Y are, with respect to changes in the input variables X. Some functions f are simple enough that an accurate and understandable model $\hat{f}$ can be fitted to match f closely with few data. This implies that both the bias and the variance are low, even for a small dataset size. However, as science struggles to study more complex phenomena, simple functions f are more scarce, and this brings opportunities for data science techniques that can reduce the variance of highly complex predictive models by harnessing large data volumes.

A formal argument in favor of using machine learning algorithms to approximate complex f functions is that some algorithmic models such as neural networks (Hornik, 1991; Hornik, Stinchcombe, & White, 1989) and fuzzy systems (Kosko, 1994; Wang, 1992) have been proved to approximate any continuous function to within any degree of accuracy, given enough data samples and computational resources. At this point, it is worth noting that most continuous functions can not even be described, i.e., they do not have a finite definition because the infinite cardinal of continuous functions is larger than the infinite cardinal of finite definitions. This suggests that there could be some problems whose functions f can not be described in a comprehensible way, but can be adequately approximated by an algorithmic predictor.

Explanation, as opposed to prediction, searches for an understandable model $\hat{f}$ with the minimum possible bias, while the minimization of the variance is not essential (Shmueli, 2010, p. 293). Consequently, for simple f, there is no conflict between explanation and prediction. Only for complex f, it is necessary to choose between them because the variance is not negligible. This conflict is counter-intuitive for many researchers because science has not been able to address this kind of complexity until machine learning algorithms were available. A way to reconcile these results with intuition is to think of explanation and prediction as two goals that coincide at the ideal limit of infinite data, where the variance is reduced to zero. These considerations suggest that the bias-variance tradeoff is a fundamental principle of

data science.

One of the challenges which are frequently made against data science is related to the overfitting problem that was mentioned above. Hypothesis-driven research establishes that models must be proposed before experiments are done. Choosing a model after the data have been collected is regarded as a bad practice, called data dredging (Thompson & Higgins, 2002, p. 1559; D. R. Anderson et al., 2001, p. 314; D. R. Anderson et al., 2000, p. 921). In statistical literature, data dredging is the extraction of hypotheses from some datasets, which are then tested for statistical significance on the same dataset. It is seen as a risky procedure because statistical hypothesis testing is based on the assumption that the hypothesis to be tested is determined prior to seeing the data. Data mining has been criticized as a form of data dredging (Fayyad, Piatetsky-Shapiro, & Smyth, 1996, p. 39). Indeed data science frequently chooses from alternative models after they have been fitted to data, so this challenge must be seriously considered. The critical problem here is that the model that best fits a given dataset is not necessarily the best performing for other data. That is, a model can be adjusted to small details that are particular to the dataset, and do not appear in different situations. As said before, this is called overfitting, and it is the issue that has rightfully been denounced by critics of data dredging (Frické, 2015, p. 657). However, data scientists are well aware of this inconvenience. It can be considerably alleviated if the dataset is randomly split into three prior to processing (Breiman, 2001, p. 204): the training set is used to fit the candidate models, the validation set is used to evaluate the models in order to choose one of them, and the test set is used to obtain an estimation of the performance of the chosen model when used to predict unseen data. There are many variants of this procedure, but they are collectively known as cross-validation (James et al., 2013, p. 176; Hastie et al., 2009, p. 241). An alternative procedure is to choose the model which conveys the maximum information about the phenomenon under study, according to information theoretic criteria (Bishop, 2006, p. 33). The validity of these approaches depends on two factors. First of all, the collected data must be representative of the phenomenon under study, so that nothing fundamentally new arises at prediction time (Frické, 2015, p. 655). Secondly, a large amount of data is required to make an informed selection among many complex models, so that the choice is not affected by random fluctuations (Shmueli & Koppius, 2011, p. 563). Therefore big data techniques are required to learn complex models since their future performance can not be adequately estimated unless the sample size is large (Dhar, 2013, p. 72; Dolinski & Troyanskaya, 2015, p. 2576).

1.5 Two kinds of data science

Next, it is time to explore the specificities of data science according to the kind of problem that it is applied. There are two kinds of datasets:

- Unstructured datasets. In this context, only short or medium term predictions are possible due to the time variability of the systems under study. The data are highly unstructured, i.e., it comes in various formats and from different sources. Furthermore, in most cases, they come from observations and not experiments. They are frequently qualitative, and in many cases, their meaning is not clear or homogeneous over all the samples, for example, in the responses to automated surveys, or the usage of a particular term in Twitter messages. Under these circumstances, it is acknowledged that all models are very far from reality. Therefore any model that generates good predictions is seen as a working approximation among many other possible models. One of the most pressing difficulties contributing to the complexity of modeling social processes is that many important variables are not measured, as the systems under study are open and include a large number of individuals, each with their behavioral patterns. However, data science contributes to the understanding of these systems by taking advantage of the ever increasing amount of data to build quantitative models that can be evaluated objectively. This kind of datasets is prevalent in the social sciences.

- Structured datasets. Long (ideally infinite) term predictions are desired since the aim is to find general laws of nature. That is, an interpretable model that produces reliable predictions in all circumstances is the overall goal. The data are typically more structured, and their meaning is prespecified by a precise and widely acknowledged ontology. Substantial efforts are made to measure all relevant variables. Here the problem of black box models is of paramount importance since scientists wish to extrapolate the models to other situations, and this can only be done by understanding how these situations differ from the one that the original data were collected. The data are almost always quantitative, and they come from measurements made under experimental conditions or at least computer simulations where the sources of error are identified and the size of the errors are controlled. These datasets are more likely to be found in the natural sciences.

These two kinds of datasets call for different methodologies and requirements, which defines a split between two types of data science. But the machine learning techniques are the same for both of them, which can lead to confusion. We will call the first one 'soft data science' and the second one 'hard data science', following the traditional distinction between soft and hard sciences (Hedges, 1987; Hong, 2013; Mryglod, Kenna, Holovatch, & Berche, 2013; Smith, Best, Stubbs, Johnston, & Archibald, 2000). Of course, the boundaries between the two are not clear cut, but this split may clarify the discussion because the role of data science in both contexts is fundamentally different (Dhar, 2013, p. 71; Zhu & Xiong, 2015). Social sciences and physics stand at the extremes (Dhar, 2013, p. 69), while biology and medicine lie in the middle field. The complexity of biological systems can not be completely reduced to the straightforward application of the underlying physical laws to their subsystems (Green, 2015, pp. 81-82; Krohs, 2015, p. 101). This leaves the alter-

native of looking for good predictors, even if the basic principles of those systems are not well understood (Braun & Marom, 2015, p. 71). In other words, long term predictions and general principles are sometimes attainable, but in many cases, the systems are too complex to derive interpretable predictive models from abstract concepts.

Minds and Machines (2018) 28:667–688
https://doi.org/10.1007/s11023-018-9480-7

Computational Functionalism for the Deep Learning Era

Ezequiel López-Rubio[1,2]

Received: 12 April 2018 / Accepted: 28 September 2018 / Published online: 5 October 2018

Abstract
Deep learning is a kind of machine learning which happens in a certain type of artificial neural networks called deep networks. Artificial deep networks, which exhibit many similarities with biological ones, have consistently shown human-like performance in many intelligent tasks. This poses the question whether this performance is caused by such similarities. After reviewing the structure and learning processes of artificial and biological neural networks, we outline two important reasons for the success of deep learning, namely the extraction of successively higher level features and the multiple layer structure, which are closely related to each other. Then some indications about the framing of this heated debate are given. After that, an assessment of the value of artificial deep networks as models of the human brain is given from the similarity perspective of model representation. Finally, a new version of computational functionalism is proposed which addresses the specificity of deep neural computation better than classic, program based computational functionalism.

Keywords Computational functionalism · Artificial intelligence · Machine learning · Neuroscience

1 Introduction

The question whether the functions of the human brain could be implemented on a computer has captured the imagination of philosophers of science and computer scientists alike over the decades. There are two lines of research in this respect. The neuroscience approach intends to simulate the biological structures of the brain as closely as possible. This is mainly aimed to gain biological knowledge about the fundamentals of information processing in the brain. It can help to understand and

✉ Ezequiel López-Rubio
ezeqlr@lcc.uma.es

[1] Departamento de Lenguajes y Ciencias de la Computación, Universidad de Málaga (UMA), Bulevar Louis Pasteur 35, 29071 Málaga, Spain

[2] Departamento de Lógica, Historia y Filosofía de la Ciencia, Universidad Nacional de Educación a Distancia (UNED), Paseo de Senda del Rey 7, 28040 Madrid, Spain

eventually cure some diseases, and it is also relevant to understand human behavior. In contrast to this, the artificial intelligence approach focuses on the design of computer systems which are able to replicate the abilities of human brains, even if the internal workings of the computer systems are fundamentally different from those of the brain. For artificial intelligence, the biology of the brain has been seen as an inspiration, rather than a prescription to be followed. In particular, machine learning is a subfield of artificial intelligence which deals with the design of algorithms whose performance improves over time by interacting with the environment. These algorithms share the ability to learn with animals, whereas they do not intend to replicate the learning processes involved in biological systems.

This state of things has been shaken to the roots by the emergence of deep learning. It deals with some types of artificial neural networks, which are made of fundamental processing units called artificial neurons. Artificial neurons are arranged into layers, which are connected to form an artificial neural network. Deep learning is characterized by the development of networks with many layers. It has made a large impact in machine learning (Yu and Deng 2011, p. 145; Yamins and DiCarlo 2016, p. 362; Silver et al. 2017) and artificial intelligence as a whole. It is seen as a revolution which will change these fields substantially (Voosen 2015). From the point of view of philosophy of science, deep learning artificial neural networks are unique among other machine learning algorithms. This is because artificial deep networks have a much closer similarity to biological neural networks than any previously considered machine learning models. Therefore, it is relevant to wonder if there is something essentially new in them, i.e. whether they are closer to human intelligence than previous machine learning models.

It has been recently argued that, while artificial neural networks are appealing because they have some analogies to biological networks, they do not exhibit any fundamental advantages with respect to other artificial intelligence models (Parnas 2017, p. 30). It is further said that the use of words like 'intelligence' or 'learning' with blurry or changed meanings creates the illusion that artificial intelligence algorithms are close to human intelligence (Parnas 2017, p. 31). This calls for precision in the evaluation of the merits of all artificial intelligence models, and in particular deep learning, which carries a positive aura due to its analogies to biological networks. This is challenging, since there is no agreed definition in psychology about human intelligence, nor a unique procedure to measure it (Legg and Hutter 2007, p. 392). Intelligence has been informally defined as the ability of an agent to attain its goals in a wide range of situations (Legg and Hutter 2007, p. 402), but this leaves the unknown of the relevant situations and goals that should be tested in order to ascertain whether a machine achieves human-like performance.

Some part of this confusion about the nature of computer intelligence can be traced to the misinterpretation of the foundational Alan Turing's paper about the possibility of building thinking machines (Turing 1950). Turing proposed that, in order to have an indication that a machine possessed intelligence, a test might be carried out to check whether humans could not distinguish the machine from another human in a conversation. But Turing did not mean that we should aim to build machines that could fool a human into believing that they are also human. Sadly, some researchers mistakenly assumed that any machine which passed Turing's test

was to be declared intelligent (Parnas 2014). In particular the ELIZA conversation program (Weizenbaum 1966), which did not exhibit any kind of intelligence nor was it meant to do so, was erroneously seen as a significant step towards machine intelligence. This misinterpretation was created by the illusion that such machines possess something close to human general knowledge and common sense. These lessons from the past teach us that the definition of general intelligence is too complex and disputable, so it will not lead us to any practical evaluation criteria. Consequently, in what follows the evaluation of deep learning will be based on its performance on specific tasks, so that performance measurement criteria are less disputed.

Deep learning has already demonstrated human-like or even higher performance in some vision tasks. It attains human level performance for handwritten digit recognition (Cireşan et al. 2012b). As seen in Cireşan et al. (2011, 2012a), Stallkamp et al. (2012), it recognizes traffic signs with an average classification error which is half of humans'. Moreover, deep learning based Go players clearly outperform the most skilled human masters (Silver et al. 2017). This level of success calls for an explanation. Some would say that deep learning networks have a much higher performance than previous machine learning algorithms because they are more similar to the biological neural networks which carry out those tasks (Marblestone et al. 2016, p. 4; Kruger et al. 2013, pp. 1865–1866). It can not be denied that deep learning is inspired by a biological analogy, but it is necessary to ascertain the extent and meaning of this analogy. In this work, we intend to study this problem. As a result of such study, it is found that artificial deep networks might work as computational models of the human brain which are more realistic than previous ones. Therefore we propose a new kind of computational functionalism which dwells on the success of artificial deep learning. Such kind of computational functionalism advocates explanations based on the computational capabilities of artificial and biological neural structures and synaptic weights, rather than classic computation by Turing machines.

The structure of this paper is as follows. Section 2 outlines deep learning artificial neural networks. It specifies the learning mechanisms that they employ, in the context of general machine learning. Then Sect. 3 addresses the fundamental problem of the interpretation of the features used in artificial and biological neural networks. The relations among these sets of features may hold the key to understand why artificial deep learning has such an impressive performance. However, the learning mechanisms in artificial and biological networks do not share many resemblances. That is, the relation of biological adaptation and machine learning is not clear, which might undermine the case for the similarity of both kinds of networks. This issue is investigated in Sect. 4. The possible novelty of artificial deep learning in terms of the modeling of the human mind resides in the role of the multiple layers in their deep structure. A possible account of this role is given in Sect. 5, which deals with the theories which state that deep structures of neural layers are successful because they adequately approximate the hierarchical structure of physical reality. Some indications about the most relevant aspects of the debate regarding the merits of artificial deep networks are given in Sect. 6. Then we advocate a similarity view of model representation in order to address the question whether artificial deep networks are good models of the human brain (Sect. 7). After that, it is recognized

that classic forms of computational functionalism can not capture the way that artificial deep networks compute, which leads us to propose a new version of computational functionalism which we call neural computational functionalism (Sect. 8). Finally, Sect. 9 concludes this paper.

2 How Deep Networks Learn

Our investigation must start by putting artificial deep networks in their context, as a first step towards the assessment of their value as models of the human brain. Artificial deep networks belong to the machine learning branch of artificial intelligence, which is one of the main subfields of computer science. As such, artificial deep networks are computer models of data. They are aimed to solve a certain problem, and they improve their performance as they learn from data. Their learning is accomplished by executing learning algorithms which might not have a biological counterpart. Moreover, the basic computing units of artificial networks, i.e. the artificial neurons, are much simpler than biological neurons. There are three main approaches to machine learning, which are associated to three kinds of problems (Trappenberg 2014): supervised learning, unsupervised learning, and reinforcement learning. Deep networks have been employed for the three of them, but their role varies.

Supervised learning intends to approximate a function from a set of examples. That is, given some example patterns of possible inputs and their desired outputs, the task is to learn a model of the underlying function such that the difference between the desired and predicted outputs are as small as possible for previously unseen patterns. The supervisor is an entity which compares the desired and predicted outputs, and provides a measure of the differences. The supervisor plays a key role in supervised learning, since these error measures are required to adjust the model to the patterns. Typical supervised deep learning networks have several layers of neurons between the inputs and the outputs, so that the information flows from one layer to the next one. This is called a feed forward network architecture. Each layer computes a nonlinear transformation of the information provided by the previous layer, so that irrelevant details of the input are progressively removed (see Sect. 5). This way, supervised deep networks may capture highly complex relations among the input and output variables.

Unsupervised learning is characterized by the absence of a supervisor, i.e. there are no desired outputs. It is aimed to discover the structure of an input dataset. Deep networks can be trained in an unsupervised way to learn the probability distribution of the input. This is the case of Deep Belief Networks (Hinton et al. 2006). Furthermore, some layers of deep feed forward networks can also be trained in an unsupervised way, so that these layers capture the relevant features (see Sect. 3). This is done by training an autoencoder, i.e. a network whose desired output coincides with the input so that the information must flow through an intermediate layer with few neurons. The presence of this narrow intermediate layer forces the network to learn a set of relevant features that it uses to reconstruct the input as faithfully as possible.

The third kind of machine learning problems is called reinforcement learning. It intends to enhance the behavior of an agent that must make decisions in

an environment that gives it rewards or penalties for its actions, where penalties can be regarded as negative rewards. Again there is no supervisor, because the rewards are generated by the environment without human intervention. This means that reinforcement learning arises in many problems that living beings must solve to survive and thrive. In this sense it can be said that it is the most natural of the three kinds of learning that we have just discussed. On the contrary, supervised learning has no direct biological counterpart, since the supply of desired outputs and the existence of a supervisor have no correlates in nature. Interestingly enough, deep learning approaches to reinforcement learning do include a supervisor (Mnih et al. 2015, p. 529). There is a replay memory which stores all past experiences of the agent. This way it is possible to compute how good the available actions at a certain state of the problem are, based on recorded rewards in similar circumstances. Then the supervisor compares the actual output of the deep network with the output that it should generate according to past experiences. This means that recollection of past memories could work as a supervisor for biological organisms, provided that the biological mechanism to compute the desired outputs is found. If no such mechanism is found, then supervised learning can only exist in an artificial environment where a supervisor intentionally provides adequate training patterns composed of inputs and their associated desired outputs. Therefore, although supervised deep learning networks have a clear connection with the structure of the physical world, and they could be valid models of it (Lin and Tegmark 2016b), their learning procedure seems to be intrinsically distinct from anything which can be related to biological evolution or human culture (Yamins and DiCarlo 2016, p. 364). Unsupervised deep learning networks do not have this inconvenience, although the relation of their learning procedures to natural processes is yet to be determined.

Much of the success of deep learning comes from a two-stage approach to training (Yu and Deng 2011, p. 145). First a deep network is trained from a wide range of situations, often in an unsupervised way. Then the network is retrained for a more specific task. This is known as fine tuning, and it is often supervised. The first stage can be conceived as an adaptation to the features of the data that best summarize the general structure of the problem, while the second stage aims to maximize the predictive performance of the network. The success of this approach indicates that there are features whose significance is quite broad (Bonfiglioli and Nanni 2016, p. 92), irrespective of specific classes or categories. In other words, the learned features comprise a good representation of the data and the underlying process that generated them. It has been argued that this means that the features are associated to the causal process behind the observed data (Hinton 2014, pp. 1091–1092). The learned features summarize nonlinear, high order correlations among the input and output variables. While some of these correlations might not be associated to the real causes of the phenomenon under study, there is a good chance that many of them are. This is enough since deep learning algorithms are able to find out which features are the most significant ones, so that spurious (non-causal) correlations are filtered out in the learning process.

3 Features in Artificial and Biological Networks

As seen in Sect. 2, the ability of deep networks to discover significant features is fundamental in their workings. This ability is so important because the discovered features have two main advantages. First of all, features comprise relevant information so that noise, errors, and unimportant characteristics of the data are discarded. Therefore the networks focus on those aspects of the input which are related to the problems that they solve. This can be regarded as a procedure to extract a concise representation of reality made of valuable information from a large volume of raw data which is mostly made of unessential details. Secondly, the multiple layer architecture of these network implies that this feature extraction is carried out by steps, so that deeper layers encode more concise and relevant representations. In particular, deeper representations are increasingly invariant with respect to irrelevant transformations. For example, the position of an object in a scene is not relevant to determine which kind of object it is, so it is expected that deeper layers of a deep network trained for object recognition encode object representations which are invariant with respect to such positional information.

It is well known that the brain is able to learn and use significant features in order to extract from a given situation the information which is relevant to solve a problem. That is, the nervous system discards useless information from the senses, so that it can manage a myriad of different cases which differ in irrelevant details. Given the much larger information storage capacity of the brain compared to the genome, the remarkable flexibility of humans and other animals to adapt their behavior to new situations, and the huge number of features to be learned, we must accept that features can not be completely innate (Hinton 2014, pp. 1078–1079). Biological evolution finds neural structures which are capable of discovering relevant features, but it can not hardwire into the genome all the features that a living being must extract and use to thrive successfully. In particular, it has been observed that low level features are mostly innate, while high level features are more dependent on learning from the environment (Kruger et al. 2013, p. 1851). For example, complex features are learned which are specifically tailored to recognize particular objects (Quiroga et al. 2007, p. 2003). This poses the question of how the central nervous system adapts to the environment in order to find out high level features which are significant to solve a problem. Moreover, some learned features are significant to solve a wide range of problems. For example, recognizing letters and other written symbols enables a person to carry out many different tasks. Hence there is the possibility that biological feature learning is decoupled to a certain degree from the specific problem that an individual is trying to solve when the feature learning is accomplished. In fact, such possibility could be favored by evolution, since finding features that can be applied to many problems saves some of the energy required to learn features for new problems (Kruger et al. 2013, p. 1864), and therefore makes the individual more fitted.

Traditional machine learning approaches to feature management were based on the manual specification of a large set of candidate features. That is, the

machine learning researcher interviewed some experts in the domain of the problem to be addressed, so that she could get some understanding of its structure. Then the experts and the machine learning researcher carefully designed a set of potentially significant features, along with the procedures to compute them from the raw data. Finally, those tentative features were supplied to a feature selection algorithm which automatically chose the most useful ones in order to solve the specific problem at hand. Deep learning means a revolution in two ways. On one hand, it learns the features itself (for example by autoencoding, see Sect. 2), which is a giant advancement with respect to manually specifying them. This frees the computer scientist from the need to find sets of candidate features and select an optimal subset of them. On the other hand, if the deep learning network learns the features in an unsupervised fashion (which is the case of autoencoders), then those features have the potential to be employed for several different problems. This is because each problem is associated with a set of desired outputs which is supplied by the supervisor in supervised learning, while unsupervised learning is not attached to any such set of desired outputs. Successes in deep reinforcement learning mean that deep learning might handle interaction with the physical world by finding salient features from the environmental rewards without the need of a supervisor, just like living beings (Mnih et al. 2015, p. 532).

At this point, given the architectural similarities between artificial and biological networks, and their ability to learn without a supervisor, we must wonder whether the features employed by biological brains and the features discovered by deep learning are similar. It has been argued that, if biologically and artificially learned features are similar, then artificial deep networks could help to understand the biological ones (Schmidhuber 2015, p. 100). This suggests that artificial deep networks could be appropriate models of their biological counterparts. An argument in favor of similarity is that deeper layers of both biological and artificial deep networks encode higher level features which are invariant to irrelevant transformations (Kruger et al. 2013, p. 1850). However, there are important differences in other aspects such as color constancy for visual information, where current machine learning approaches use local low level operations, while the biological vision system employs several local and global procedures which belong to various complexity levels (Kruger et al. 2013, pp. 1854–1855, 1860). If they are not similar, then it is possible that artificially learned features perform better than biologically learned features. Moreover, it is also possible that biologically learned features can be read out by machine learning algorithms, i.e. these algorithms can accomplish tasks when provided with the biologically learned features, which has already been observed in some cases (Kiani et al. 2007). At least artificial deep learning features have become widely used in natural language processing (Yu and Deng 2011, pp. 148–149). These features have proven themselves useful to discover highly complex relations among words in datasets spanning several centuries of literature (Bonfiglioli and Nanni 2016, p. 93). This abstract task is beyond the capabilities of any single human, since no one can read such a large amount of text and then find out the relations among all the words that occur. Furthermore, deep learning based machine translation systems have demonstrated a performance which is quite similar to that of humans (Wu

et al. 2016). However, the limits of deep learning when applied to natural language processing are not clear yet, since it seems that the parts of language where rules are more influential remain elusive to neural networks, while cognitive science is still more successful (Manning 2015, p. 706). It has been demonstrated that the visual cortex has a hierarchical structure which has some similarities with convolutional neural networks (Hong et al. 2016, p. 619), so that the neurons in the highest layers of both hierarchies are associated with object categories that are recognized (Yamins et al. 2014, p. 8619). Moreover, it has been found that convolutional neural networks trained for one visual task also perform well for other tasks (Yamins and DiCarlo 2016, p. 363), which suggests that features have certain independence with respect to the task to be carried out. This agrees with the operation of the human visual cortex, where object categorization is carried out at the higher levels of the neural layer hierarchy, whereas the information about category orthogonal properties also increases in those higher levels (Hong et al. 2016, pp. 613, 620).

The situation might be summarized as follows. Both biological and artificial deep networks manage the complexity of the tasks by finding relevant features of progressively higher level of abstraction. Biological networks do not possess such a regular structure as artificial ones. This can be explained by the fact that they have not been designed, but selected by evolution. In any case, their resemblances are remarkable with respect to features. On one hand, artificial networks have already been used as models of biological ones by neuroscience researchers, which means that artificially learned features can model biological ones. On the other hand, artificial networks can read out biologically learned features, i.e. biological features can be employed by artificial networks. Consequently there is a possibility for a two way flow of features between the biological and the artificial. Moreover, both artificial and biological features show a certain degree of independence with respect to the specific task that they were learned for. This further suggests that both kinds of features are compatible. Finally, there are tasks such as the discovery of relations among words in a large literary corpus that are beyond human capabilities due to our time and memory limitations, but can be addressed with machines. In Table 1 the similarities and differences among biological and artificial deep networks are listed.

Table 1 Similarities and differences among biological and artificial deep networks

Similarities	Differences
Hierarchy of layers	No supervisor in biological networks
Filtering irrelevant details	Artificial neurons are too simple
Levels of abstraction	Irregular structure of biological networks
Invariance to irrelevant transformations	Learning mechanism differences
Features independent from the task	Biological networks are evolved and not designed

4 Adaptation in Biological Networks

Both biological and artificial networks rely on learning in order to enhance their problem solving performance. However, the connection between the biological and artificial modes of learning is problematic, since artificial deep learning algorithms are not designed to match biological processes, and at first sight they do not have many resemblances. This connection must be clarified, since the possibility of using artificial deep networks as realistic models of the human brain depends on how similar artificial and biological neural adaptation processes are. In this section possible ways to establish some links are investigated.

As mentioned in Sect. 2, among the three kinds of learning discussed there, reinforcement learning is the most biologically plausible one. This is because there is no need for a supervisor to provide the desired output. One of the most influential theories of learning in cognitive biology is the Rescorla–Wagner rule (Fitch 2014, pp. 338, 343; Bassett and Mattar 2017, pp. 258–259; Holland and Schiffino 2016, p. 207). It assumes that organisms build predictive models of reality, i.e. models that yield predictions about the reward (unconditioned stimulus) which should follow a given observed event (conditioned stimulus). Those models are adjusted by evaluating the error of the predicted reward with respect to the actual reward from the environment. Here the environment acts like a supervisor, since it provides the desired outputs (the actual rewards) for the reward prediction model built by the organism. Hence, the Rescorla–Wagner model combines reinforcement with supervised learning. Fitch argues that error backpropagation, which is at the heart of deep learning algorithms, is biologically plausible at the neuron level, although not at the network level (Fitch 2014, pp. 340–341). For example, dopamine flows might increase or decrease in response to positive or negative reward prediction errors, respectively (Fitch 2014, p. 345). However, recent developments on machine learning suggest that training algorithms that only employ local information, which are thereby more biologically plausible, might be used to replace backpropagation in artificial deep networks (Hassabis et al. 2017, p. 254). On their part, Basset et al. argue that the reinforcement mechanism can work at the network level (Bassett and Mattar 2017, pp. 259–260). In any case, the focus on prediction of rewards is common to both lines of thought. The Pearce-Hall theory varies in several ways from Rescorla–Wagner, but it also considers that reward prediction errors drive learning (Bassett and Mattar 2017, p. 208). Therefore, supervised learning is biologically plausible in a global reinforcement learning context driven by rewards from the environment. The association of stimuli to rewards must be genetically determined, since they form the basis for this learning scheme (Fitch 2014, p. 336).

If biological learning is driven by the errors committed by a reward prediction model, then the kind of models that can be built by biological neurons is of paramount importance. Models stored in the human brain would be more complex than those of other animals, since humans would have an innate proclivity to discover tree structures (Fitch 2014, pp. 351–352), which are required for many typically human cognitive capabilities (Dehaene et al. 2015, p. 13).

Human language is a paramount example of a tree structured problem. Linguistic elements can have a strong correlation even if they are separated by many other elements in a text, and this can not be modeled by shallow networks (Lin and Tegmark 2016a, pp. 8–10). In other words, while the overall reward prediction mechanism might be the same, the range of possible reward prediction models that the organism can build influences the kind of cognitive abilities that can be learned. This is because complex reward patterns can not be adequately predicted by simple models.

Perhaps the most relevant example of the need for complex models is human vision. The adaptation of the human vision system can not be explained by reinforcement learning alone. Visual features are too complex to be learned by simple reward signals. Moreover, the genetic information stored in the DNA, which comprises the bulk of what is passed from one generation to the next, can not hold the complexity of the situations that humans must evaluate as explained in Sect. 3. There must be some unsupervised feature learning in place, just like in the deep learning case. The genetic information enables the unsupervised feature learning process, but the particular features to be learned are not genetically determined, at least at the highest levels of the visual pathway. Experiments with ferrets suggest that genes are responsible for preparing the brain to extract useful features from the sensory signals, but the exact nature of those features and signals is not hard coded in the DNA (Merzenich 2000, p. 821). In other words, each type of sensory inputs generates a different structure in the developing brain (Melchner et al. 2000, p. 872). The principles of predictive model construction and reinforcement learning based on prediction error are common to many sensory modalities (Dehaene et al. 2015, p. 15), so that specific neural structures might arise in response to the particular features of each kind of sensory inputs. Therefore the division of the normal brain into regions can be changed, which means that biological neurons are not genetically programmed to recognize a specific kind of features. It is more likely that the genome drives a general tendency to hierarchical organization into layers. This hierarchical structure increases the efficiency of the learning process because the features to be extracted at each layer are only slightly more complex than the features of the previous layer, so that the adaptation can be accomplished with relatively few samples (Kruger et al. 2013, pp. 1864–1865). This mechanism would work equally well for different kinds of sensory inputs to the lowest layers of the structure.

As seen above, current theories of biological learning integrate the three kinds of learning that were introduced in Sect. 2. There is a global reinforcement learning mechanism, where the distinction between positive and negative rewards is genetically determined. The maximization of the rewards is done by learning a predictive model of such rewards, so that the adjustment of the model is carried out by supervised learning, possibly including error backpropagation. Finally, complex predictive models are hierarchies made of several layers, and the discovery of relevant features at each layer is carried out by unsupervised learning.

5 Capturing the Hierarchical Structure of the Universe

Prior to deep learning, artificial neural networks typically had very few layers, i.e. they were shallow. Why deep and not shallow? This question deserves a careful consideration, since its answer will indicate whether deep learning represents a truly new step in the modelling of human mind, which can be recruited by computational functionalism. The answer that is presented in this section is related to the modeling of physical reality by levels of description.

While shallow neural networks can approximate any function, they may require large amounts of neurons and training data (LeRoux and Bengio 2008, p. 1631). It has been suggested that deep learning is better because deep networks have a hierarchical structure that mimics the structure of physical reality (Lin and Tegmark 2016b, p. 9; Patel et al. 2015, pp. 4–5). That is, at a lower (microscopic) level of description there are many nuisance variables which are not relevant at higher (macroscopic) levels, where a much smaller number of degrees of freedom are significant. Higher levels in the hierarchy filter out the nuisance information which abounds in the lower levels. In other words, deep networks are particularly efficient at approximating the kind of functions that occur in practice, while the set of all possible functions is much larger (Lin and Tegmark 2016b, p. 2). Real world functions are associated to sparse, low order polynomial Hamiltonians (Lin and Tegmark 2016b, pp. 3–5). That is, functions that describe real entities conform a set which is relatively small and easy to describe when compared to the set of all possible functions.

Shallow architectures can not adequately capture many natural processes which exhibit intricate correlations (Levine et al. 2017, pp. 25, 29), in particular long range correlations which obey a power law with respect to distance (Lin and Tegmark 2016a, pp. 1–2). On their part, deep architectures can generate context free grammars which exhibit the required power law correlation decay (Lin and Tegmark 2016a, pp. 8–10). Here distances measured on the input to the neural system are considered. In vision this would be the distance in pixels between two visual features which are correlated. For example, in a picture of a face it could be the distance in pixels or centimetres between both eyes, which are correlated because they are likely to be similar in color and shape. In natural language processing the distance could be the number of words between two words which are correlated. For example, the subject and the verb of a sentence are correlated because they must agree in number (singular or plural). The previously mentioned findings suggest that this kind of long range correlations can be learned more efficiently by deep architectures.

The successive layers in the network structure attain progressively higher levels of invariance with respect to unimportant transformations in the input data, which can not be done efficiently by shallow architectures (Lin and Tegmark 2016b, pp. 9, 12; Mehta and Schwab 2014, pp. 1, 7). The efficiency of deep networks is exponentially higher than that of shallow networks, i.e. an exponentially larger shallow network is required as the complexity of the correlations to be learned grows (Levine et al. 2017, p. 26). These higher levels of invariance

can also be regarded as more compressed representations of the input data which retain the information with respect to the desired output (Tishby and Zaslavsky 2015, pp. 1–3; Khadivi et al. 2016, pp. 3–4).

Another key factor is that deep networks can approximate compositional functions (those that can be expressed as compositions of constituent functions of small numbers of variables) exponentially better than shallow ones, so that the former are not affected by the curse of dimensionality when approximating those functions (Poggio et al. 2017, p. 7). It turns out that many tasks in vision and language understanding are compositional, and this explains the advantage of deep networks with respect to shallow ones in computer vision and natural language processing (Poggio et al. 2017, p. 13; Kruger et al. 2013, p. 1865). Invariance and compositionality together mean that deeper layers contain more abstract representations of the input data (Schmidhuber 2015, p. 103), which are useful for new tasks (Hassabis et al. 2017, p. 250). While deep neural networks are universal approximators like shallow ones (Poggio et al. 2017, p. 3), there are functions that can not be implemented exactly by either of them (Khadivi et al. 2016, p. 5).

However, the primate visual processing system is not perfectly hierarchical nor the information always flows in the forward direction, since it contains shortcuts and feedback connections along with specialized layers which do not follow the hierarchical structure. Moreover, its architecture remains relatively plastic during the life of the individual (Kruger et al. 2013, pp. 1865–1866). This means that if the human brain is the pinnacle of intelligence, then present day artificial deep learning networks are far from that goal from the structural point of view. Nevertheless, receptive field overlapping is a characteristic of deep convolutional neural networks which partly explains their performance in vision tasks (Sharir and Shashua 2017, p. 2), while it has been recognized as one of the key ingredients of the endstopping characteristic of the human vision system (Fitch 2014, p. 344). Endstopping is the capability of some neurons to activate in response to a short bar of a certain orientation while they do not activate when the bar is too long, even if the orientation is still the same.

Weight sharing is a key characteristic of artificial deep networks (Hinton 2014, p. 1081), but it does not relate directly to any biological structure since biological neurons do not share their synapses. Weight sharing means that the same number is used as the synaptic weight for several synapses corresponding to different artificial neurons. Therefore there are aspects of deep learning which depart largely from biology. This is acknowledged by machine learning researchers (Hinton 2014, pp. 1095–1096). However, it has been argued that synaptic weights at different locations of the brain which process the same kind of information could converge, so that weight sharing could be a realistic approximation to biological synapses (Yamins and DiCarlo 2016, pp. 357–358). Weight sharing is more important in the lower layers of artificial networks, which are genetically determined and not learned in biological networks (Kruger et al. 2013, p. 1851), so maybe this difference among biology and machine learning is not as significant as it seems. Michael Jordan thinks that neuroscience is far from finding the way that the brain learns, and deep learning methods are not related to biological learning in any significant way (Gomes 2014).

Spiking neurons have not been successfully integrated into deep learning architectures to this date (Schmidhuber 2015, p. 100). Moreover, physical constraints which exist in the human brain on the communications among neurons are not honored in artificial deep learning neural networks (Schmidhuber 2015, p. 103). However it has been argued that compositionality is the essential ingredient that ensures the success of deep networks, and not weight sharing (Poggio et al. 2017, p. 12). If this is true, then deep networks with a more biologically plausible structure could be developed. This would suggest that both artificial deep neural networks and biological neural networks are good at solving the same kind of problems, namely those that can be expressed as compositional functions (Poggio et al. 2017, p. 16).

6 The Debate

The debate about the relation between deep learning and the human brain has seen the clash between the enthusiasts, who see artificial deep learning networks as a revolutionary model which has unprecedented similarities with biological neural networks (Sect. 4) or even the structure of the universe (Sect. 5); and detractors, who think that there is nothing essentially new in deep learning, as compared to other machine learning models. This debate can be regarded as an evaluation of the merits of artificial neural networks as computational models of the human mind, whether these networks can provide the foundation for a new kind of computational functionalism. The situation should be clarified by specifying what enthusiasts and detractors really mean. Next the most relevant problems which arise in the debate are considered.

It is important that argumentations clearly identify their scope, i.e. the range of brain functions that they consider. On one hand, enthusiasts should specify which high level brain functions they claim that deep learning is able to simulate properly. Aside from computer vision and natural language processing tasks, artificial deep learning networks have shown a standard performance when compared to earlier machine learning approaches. In other words, if deep architectures form the basis of the human brain, then we expect to see substantial performance improvements for a wide range of intelligent tasks. On the other hand, detractors should not use non perceptual human capabilities such as conscience or feelings as examples to try to undermine the possibility of deep learning simulating human perceptual functions. Maybe perceptual tasks are accomplished by deep neural networks, while other tasks are carried out by other architectures. Moreover, some non perceptual capabilities such as conscience are poorly understood with respect to perceptual ones, so it is difficult to ascertain whether they might be carried out by deep learning.

The performance goals of brain models are fundamentally different for artificial intelligence and neuroscience. Neuroscience aims to obtain models which are as close as possible to biological reality, as it aims to explain the mechanisms that underlie the human brain. However, artificial intelligence usually has technological goals, i.e. its aim is the construction of machines which carry out intelligent tasks. Artificial intelligence scientists see the similarity between artificial and biological deep networks as an advantage because they know that the human brain performs

such intelligent tasks correctly, and they expect that artificial deep networks will inherit those abilities from their biological counterparts. In this sense, deep learning research is a form of reverse engineering because it tries to extract solutions from an already working system (the human brain). Consequently, the goals of artificial deep learning models depend on whether they are employed for neuroscience or artificial intelligence purposes.

Performance evaluation is another key aspect of the deep learning debate. Quantitative evaluation measures should be agreed so that the merits of deep learning are assessed. Qualitative assertions should be avoided, since their evaluation is more controversial. If qualitative evaluation is unavoidable, then it is advisable that the evaluation is blind, i.e. the evaluators do not know whether they are evaluating a human or a machine. This is particularly important for non perceptual capabilities, where quantitative evaluation is more difficult. Moreover, detractors should compare the performance of deep learning with respect to a competing theory. Saying that deep learning is not perfect is a truism. The real question is whether deep learning is a better model than earlier ones. This appeal to performance is fundamental because in most cases deep learning machines are built with a technological goal, rather than a purely scientific one.

Thirdly, discussions about the meaning and relevance of deep learning must address the hierarchical structure of the brain. As the human brain is organized by levels of architecture and function, the architectural and functional similarities and differences should be discussed level by level. From the functionalist perspective that is advocated in this work, it is irrelevant to say that deep learning is not similar to the brain because deep learning is implemented on silicon chips while biological neurons are made of biomolecules. On the other hand, it is also meaningless to say that deep learning and the brain are similar because both of them can recognize traffic signs. Furthermore, biological and artificial learning mechanisms must be compared. It must be ascertained whether deep reinforcement learning (based on computing the desired output from past memories) can be associated with biological mechanisms in the brain. Otherwise, the success of deep supervised learning networks does not directly translate to a model of biological learning.

These difficulties in the evaluation of the value of deep learning can be alleviated by framing the entire debate in philosophy of science terms. Our proposal is to employ similarity terms to this end, as detailed next.

7 Modeling with Artificial Deep Networks

The question whether deep learning is a good model of the human brain can be seen as an instance of the scientific representation problem. Here we advocate a similarity conception of model representation (Giere 2009; Mäki 2009, 2011; Weisberg 2013), as opposed to the structuralist view (Fraassen 2008; Bueno et al. 2002; Bueno and French 2011; Bartels 2006). Artificial deep networks are computational models (Weisberg 2013, pp. 13–15) of biological ones. This way, deep learning would be a good model of the target system, i.e. the human brain, if there are agents (computer scientists) who use deep learning to represent the brain by proposing the hypothesis that they are similar

in some architectural and functional respects for the purpose of performing intelligent tasks such as object recognition or natural language processing. In Michael Weisberg's terms, it could be said that such architectural and functional respects are given by the modeler the highest weights in the weighted feature matching process between the target and the model (Weisberg 2015, pp. 299–301). The architectural similarity would include some straightforward mappings from the target system to the computational structures which form the model (Weisberg 2013, pp. 29–31): biological neuron to artificial neuron, synapse to connection between two artificial neurons, biological neural layer to artificial neural layer, and the like. The learning algorithm could represent the adaptation of the biological neural network to its environment, although the similarity among biological and artificial learning is less clear (Sect. 4).

Functional similarity between artificial and biological neural networks can be understood in two ways. A *black box* approach means that we compare the inputs and outputs of the overall networks, without having any interest in their internal workings. On the other hand, a *white box* approach means that the similarity must also extend to the kind of information which flows in their interior. The white box approach is more demanding, that is, one can have a black box similarity without a white box similarity. An example of white box similarity is the use of similar features (see Sect. 3). Similarity among the features employed in biological and artificial networks is to be assessed irrespective of the coding of the features. Biological and computer coding of information are substantially different, but this is irrelevant to the functions that both kinds neural networks implement. A possible way to carry out such similarity assessment is to evaluate the mutual information between a biological and an artificial feature. There is a range of degrees between black and white box similarities. From the current state of research, it is likely that the similarities among biological and artificial features extends from the highest level of description, i.e. the overall inputs and outputs, to a certain intermediate level of description, while the lowest levels such as the electrical signals in the biological synapses do not match well with their artificial counterparts. Such lowest levels would be excluded from the modeling process, i.e. they would have zero weight in Michael Weisberg's terms. In other words, artificial deep networks are not designed to match all biological details, but they are expected to extract intermediate and high level features which are similar to biological ones. If this is achieved, then it is further expected that artificial networks will be able to solve the same kinds of problems as biological ones. This would be enough for artificial intelligence researchers to say that artificial deep learning is a good model of the human brain. On their part, neuroscientists would not be satisfied unless the lowest levels of information processing in the brain are accurately modeled. Artificial intelligence and neuroscience modeling goals should be seen as alternative, rather than conflicting, approaches.

8 Neural Computational Functionalism

It could be argued that, even if artificial and biological networks build models which are somehow similar, they still differ in the way that the necessary computations are carried out. Classic computational functionalism (CCF) depicts the brain as a

Turing machine and the mind as the software running on the machine (Piccinini 2010, p. 276). Neither biological nor artificial neural networks run any software, if software is understood as a sequence of program instructions. Therefore, CCF is unable to provide similar models of the brain. This calls for a new form of computational functionalism which is more realistic. This way, a more credible account of the nature of mental states would be obtained. In this section such a proposal is detailed, that builds on the success of artificial deep networks.

The role of the program (software) of a digital computer is played in a deep learning network by the set of its synaptic weights. This is because the synaptic weights store the information that the network requires to perform its function. Therefore the synaptic weights provide a partial explanation of the capabilities of the network. While universal Turing machines and digital computers are amenable to explanations by program execution (Piccinini 2010, p. 277), biological and artificial neural networks allow explanations by synaptic weights:

> An *explanation by synaptic weights* of a capacity C possessed by a (biological or artificial) neural system S is a set of weights W for C such that S possesses C because S operates according to its stored weights W.

With the help of explanations by synaptic weights, it is possible to revive computational functionalism in spite of the criticism that the human brain does not execute any program (Piccinini 2010, p. 278). The versatility of the human mind is partly explained by its power to operate according to a set of stored weights. The procedure by which such weights are set up is not relevant to the concept of explanation by synaptic weights. This implies that both innate and learned weights are allowed for biological networks. There is another characteristic of neural networks which can partially explain their capabilities, namely their architecture, i.e. the arrangement of the neurons in layers and the connections among them. The innate structure of biological neural systems is regarded as a legitimate explanatory concept (Piccinini and Scarantino 2011, p. 16). Since the architecture of artificial neural networks is specified before their synaptic weights are determined, it is possible to compare the innate structure of biological networks to the prespecified architecture of artificial networks. Both of them can explain the capabilities of their respective networks, to a certain extent.

Since synaptic weights do not constitute a program, we do not need to assume that artificial deep learning networks are running a program in order to establish their similarity to the biological networks. But artificial networks are still computing devices, so that a version of computational functionalism which does not rely on an analogy among Turing machines and human brains is possible. In other words, artificial deep learning networks compute, but they do not execute programs (Piccinini 2010, p. 291). Therefore, the success of artificial deep learning networks is an argument against the classic, program based version of computational functionalism (Piccinini 2010, p. 296). This leads us to propose a new version of computational functionalism:

> *Neural computational functionalism* (NCF): the mind is the set of synaptic weights of the brain.

This is to be interpreted in the sense that: (a) the brain stores synaptic weights in its neural structures, (b) some of those neural structures are organized in a hierarchy of layers, (c) those synaptic weights determine the computation of significant features of progressively higher level as we traverse the neural hierarchy, (d) those features ultimately determine behavior. Neural computational functionalism agrees with the classic computational functionalism in that minds are multiply realizable (Piccinini 2010, p. 297), since deep networks can be implemented on different kinds of hardware. NCF would be a specific case of the generic computational functionalism advocated in Piccinini (2010), p. 300, which allows any kind of computing mechanism. It must be noted that NCF does not require that all neural structures are hierarchical, i.e. explanations by synaptic weights can be developed for single layered structures. However, deep networks exhibit computation of features of progressively higher level which account for key capabilities of neural systems such as vision.

Another fundamental difference between the classic version of computational functionalism and NCF is that the former relies on digital computation, which manipulates discrete variables which encode symbols, while the latter manipulates continuous variables which do not encode symbols. Analog and neural computation are different kinds of non digital computation, and they must be clearly distinguished. Analog computers are a kind of computing devices which are designed to solve differential equations, so the range of functions that they can implement is rather small (Piccinini and Scarantino 2011, p. 11; Piccinini and Bahar 2013, p. 461).

In contrast to this, artificial deep networks can implement a much broader class of functions because they carry out compositions of functions (see Sect. 5) which typically include sharp nonlinearities. It has been shown that neural activity in the human ventral cortex can be modeled by artificial hierarchical deep neural networks (Hong et al. 2016, p. 619). Even shallow neural networks are universal in the sense that they can approximate any continuous function with a compact domain and range (Maass 1997, p. 1659), and deep networks inherit this capability.

Artificial neural networks can not be modeled by Turing machines because they do not manipulate symbols. The Blum–Shub–Smale model (BSS, Blum et al. 1989) is the reference model of general computation with continuous variables, and it is adequate to formalize the kind of computations that artificial neural networks carry out. This means that, even though artificial neural networks are implemented on digital computers, their reference model is different because neural networks are not designed to manipulate symbols. Therefore, digital computers are employed to carry out discrete variable approximations to continuous variable computations. It has been argued that biological neural networks manipulate variables which are not discrete or analog, since the voltages are continuous variables but their most relevant features are the sharp spikes (Piccinini and Bahar 2013, pp. 466–467). In artificial neural networks, threshold-like nonlinearities are typically used in the activation functions, so both kinds of neural networks manipulate continuous variables which exhibit abrupt changes. The obtained values are then treated as instances of continuous variables, so there is no symbolic processing (Piccinini and Bahar 2013, pp. 468, 471). Thus it can be said that both artificial and biological networks manipulate continuous variables with frequent abrupt changes whose meaning is not symbolic.

Computation with neural networks is resilient to small changes in the architecture and the weight vectors, so there is not a problem with margins of error associated to discrete digits (Piccinini and Bahar 2013, p. 473). Sharp nonlinearities (threshold-like) in the activation functions of neural networks make the implemented functions robust, because a wide range of inputs result in a narrow range of outputs, i.e. small errors in the inputs are mostly ignored by the neuron. But at the same time the implemented function is still continuous. Moreover, biologically-oriented spiking neural network models demonstrate that the shape of the of the response and threshold functions do not significantly affect the class of functions that can be computed (Maass 1996, p. 4).

Artificial neural neurons use continuous output values to play the role of spike rates or inter-spike intervals of biological neurons (Piccinini and Scarantino 2011, p. 12; Maass 1997, pp. 1660–1661). This suggests that biological and artificial neurons might have different response and threshold functions, and still compute the same functions. Overall it can be said that artificial neural networks are best viewed under the BSS model perspective, i.e. machines which manipulate continuous values with a frequent use of threshold-like functions which are not associated to any symbolic processing. Therefore, artificial deep networks could be good models of the human brain even if Turing machines and digital computation are not.

It could be argued that NCF is rather simplistic, since there are control parameters other than synaptic weights that affect the operation of neural systems. For example, in the Rescorla–Wagner model of biological networks (Fitch 2014, p. 338) there is a learning speed parameter, and there are also learning rate parameters in most artificial networks. NCF can be extended to accomodate these details, so that all those control parameters P are considered along with the synaptic weights W. Then an explanation is a tuple (W,P) rather than just the set of synaptic weights W. This way, (W,P) is a partial explanation of the capabilities of the network which has a higher explanatory power than the partial explanation W.

Furthermore, there are aspects of biological nervous systems such as the glial cells which are not considered by NCF. At this respect it must be emphasized that explanations provided by computational functionalism either in the CCF or NCF form are always partial. For NCF this means that there are capabilities of nervous systems that can only be explained by biological factors not related to synaptic weights (or control parameters).

Given the above considerations, it is possible to outline some common principles of biological and artificial neural computation:

- Computation on continuous variables. The BSS model of computation is adequate for both biological and artificial neural networks, while Turing machines are not. There are no discrete or discretized variables, i.e. no alphabet of symbols.
- Tolerance to noise. In biological networks this is provided by the stochasticity of neural trains (Piccinini and Bahar 2013, p. 474). In artificial networks it is achieved by using threshold-like nonlinearities in activation functions of artificial neurons, so that a wide range of inputs result in a narrow range of outputs, i.e. small quantities of noise in the inputs are mostly ignored by the neurons.

- Tolerance to faults/redundancy. Each input synapse has a small contribution to the result, so that if a synapse (or neuron) fails or a synaptic weight gets corrupted, then the computed function degrades gracefully.
- Feature extraction. Meaningful features are extracted from large volumes of data (see Sect. 3). Such features are relevant for many problems. Biological features are compatible with artificial networks, as shown in Kiani et al. (2007).
- Hierarchy of representation levels. Successive layers of neurons extract features of progressively higher level (see Sect. 5).

Artificial deep networks have already been proposed as computational models of some parts of the brain, thereby yielding partial explanations of their function (Hassabis et al. 2017, p. 254). In the case of the ventral visual stream, a correspondence among the layers of the artificial networks and those of the biological networks has been done such that the response patterns in the artificial layers are predictive of those of the biological layers (Hong et al. 2016, p. 619). That is, functional layer-to-layer correspondences have been found between artificial and biological deep networks. The computational capabilities of the neural networks is what matters, and not the details of the implementation of the computations.

The similarities between biological and artificial neural computation imply that NCF is much closer to biological reality than CCF. CCF is built on the Turing machine metaphor of the brain, which is not adequate because Turing machines are good models of computers which manipulate symbols by running programs, while they are not good models of brains. Moreover, NCF can be validated by evaluating the extent to which artificial and deep networks agree in their information processing mechanisms. Therefore, the success of artificial deep learning suggests that CCF should be replaced by NCF.

9 Conclusion

The debate about the meaning of deep learning is becoming more heated as deep learning networks attain human-like or even higher performance in more tasks. They stand among the machine learning models which are closest to the human brain (Sect. 2). It seems that their success is largely explained by their ability to learn significant features whose utility goes well beyond the specific problem that they were learned from (Sect. 3), which is also the case of biological networks. Furthermore, biological networks exhibit the three main kinds of learning that exist in machine learning in general, and artificial deep networks in particular (Sect. 4). It is also plausible that deep architectures are required in order to capture the structures that naturally appear in physical reality (Sect. 5). This would suggest that deep networks may attain performance levels which are unattainable by shallow architectures. The scope and the evaluation of deep networks should stand at the core of the debate about them (Sect. 6). We have argued that a better understanding of the value and meaning of deep networks can be obtained by studying them both from the similarity conception of model representation (Sect. 7). In particular, artificial deep learning can be regarded as a good model of the human brain if they are similar in the

features that they extract from the input information and the kind of problems that can be solved with the help of such features. Finally, we have criticized the limitations of the classic, program execution version of computational functionalism. This has led us to propose a new version of computational functionalism, namely neural computational functionalism. It states that the mind is the set of synaptic weights in the brain, which determine the computation of significant features suitable for carrying out intelligent tasks. Neural computational functionalism is more realistic than the classic version, which means that the former is a more credible conception of the nature of mental states.

Acknowledgements The author wishes to thank the editor and the anonymous reviewers for their constructive feedback on the manuscript. He is also grateful to David Teira (Universidad Nacional de Educación a Distancia, Madrid, Spain) and Emanuele Ratti (University of Notre Dame) for their valuable comments. Finally, he is indebted to José Muñoz-Pérez, José Luis Pérez-de-la-Cruz and Lawrence Mandow (Universidad de Málaga, Spain) for sharing with him their views on Artificial Intelligence.

References

Bartels, A. (2006). Defending the structural concept of representation. *THEORIA An International Journal for Theory, History and Foundations of Science*, *21*(1), 7–19.

Bassett, D. S., & Mattar, M. G. (2017). A network neuroscience of human learning: Potential to inform quantitative theories of brain and behavior. *Trends in Cognitive Sciences*, *21*(4), 250–264.

Blum, L., Shub, M., & Smale, S. (1989). On a theory of computation and complexity over the real numbers: NP-completeness, recursive functions and universal machines. *Bulletin of the American Mathematical Society*, *21*(1), 1–46.

Bonfiglioli, R., & Nanni, F. (2016). History and philosophy of computing. In *From close to distant and back: how to read with the help of machines* (pp. 87–100). Springer, Cham.

Bueno, O., & French, S. (2011). How theories represent. *The British Journal for the Philosophy of Science*, *62*(4), 857–894.

Bueno, O., French, S., & Ladyman, J. (2002). On representing the relationship between the mathematical and the empirical. *Philosophy of Science*, *69*(3), 452–473.

Cireşan, D., Meier, U., Masci, J., & Schmidhuber, J. (2011). A committee of neural networks for traffic sign classification. In *The 2011 international joint conference on neural networks* (pp. 1918–1921).

Cireşan, D., Meier, U., Masci, J., & Schmidhuber, J. (2012a). Multi-column deep neural network for traffic sign classification. *Neural Networks*, *32*, 333–338.

Cireşan, D., Meier, U., & Schmidhuber, J. (2012b). Multi-column deep neural networks for image classification. In *Proceedings of the 2012 IEEE conference on computer vision and pattern recognition (CVPR), IEEE Computer Society, Washington, DC, USA, CVPR '12* (pp. 3642–3649).

Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., & Pallier, C. (2015). The neural representation of sequences: From transition probabilities to algebraic patterns and linguistic trees. *Neuron*, *88*(1), 2–19.

Fitch, W. T. (2014). Toward a computational framework for cognitive biology: Unifying approaches from cognitive neuroscience and comparative cognition. *Physics of Life Reviews*, *11*(3), 329–364.

Giere, R. N. (2009). An agent-based conception of models and scientific representation. *Synthese*, *172*(2), 269.

Gomes, L. (2014). Machine-learning maestro Michael Jordan on the delusions of big data and other huge engineering efforts. IEEE Spectrum 20 Oct 2014.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(2), 245–258.

Hinton, G. (2014). Where do features come from? *Cognitive Science*, *38*(6), 1078–1101.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527–1554.

Holland, P. C., & Schiffino, F. L. (2016). Mini-review: Prediction errors, attention and associative learning. *Neurobiology of Learning and Memory*, *131*, 207–215.
Hong, H., Yamins, D. L. K., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, *19*, 613–622.
Khadivi, P., Tandon, R., & Ramakrishnan, N. (2016). Flow of information in feed-forward deep neural networks. arxiv:1603.06220v1.
Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology*, *97*(6), 4296–4309.
Kruger, N., Janssen, P., Kalkan, S., Lappe, M., Leonardis, A., Piater, J., et al. (2013). Deep hierarchies in the primate visual cortex: What can we learn for computer vision? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1847–1871.
Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds and Machines*, *17*(4), 391–444.
LeRoux, N., & Bengio, Y. (2008). Representational power of restricted Boltzmann machines and deep belief networks. *Neural Computation*, *20*(6), 1631–1649.
Levine, Y., Yakira, D., Cohen, N., & Shashua, A. (2017). Deep learning and quantum entanglement: Fundamental connections with implications to network design. arxiv:1704.01552.
Lin, H. W., & Tegmark, M. (2016a). Critical behavior from deep dynamics: A hidden dimension in natural language. arxiv:1606.06737.
Lin, H. W., & Tegmark, M. (2016b). Why does deep and cheap learning work so well? arxiv:1608.08225.
Maass, W. (1996). Lower bounds for the computational power of networks of spiking neurons. *Neural Computation*, *8*(1), 1–40.
Maass, W. (1997). Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, *10*(9), 1659–1671.
Mäki, U. (2009). MISSing the world. Models as isolations and credible surrogate systems. *Erkenntnis*, *70*(1), 29–43.
Mäki, U. (2011). Models and the locus of their truth. *Synthese*, *180*(1), 47–63.
Manning, C. D. (2015). Computational linguistics and deep learning. *Computational Linguistics*, *41*(4), 701–707.
Marblestone, A. H., Wayne, G., & Kording, K. P. (2016). Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, *10*, 94.
Mehta, P., & Schwab, D. J. (2014). An exact mapping between the variational renormalization group and deep learning. arxiv:1410.3831v1.
Merzenich, M. (2000). Seeing in the sound zone. *Nature*, *404*, 820–821.
Mnih, V., Kavukcuoglu, K., & Silver, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*, 529–533.
Parnas, D. L. (2014). On the significance of Turing's test. *Communications of the ACM*, *57*(12), 8.
Parnas, D. L. (2017). The real risks of artificial intelligence. *Communications of the ACM*, *60*(10), 27–31.
Patel, A. B., Nguyen, T., & Baraniuk, R. G. (2015). A probabilistic theory of deep learning. arxiv:1504.00641v1.
Piccinini, G. (2010). The mind as neural software? Understanding functionalism, computationalism, and computational functionalism. *Philosophy and Phenomenological Research*, *81*(2), 269–311.
Piccinini, G., & Bahar, S. (2013). Neural computation and the computational theory of cognition. *Cognitive Science*, *37*(3), 453–488.
Piccinini, G., & Scarantino, A. (2011). Information processing, computation, and cognition. *Journal of Biological Physics*, *37*(1), 1–38.
Poggio, T., Mhaskar, H., Rosasco, L., Miranda, B., & Liao, Q. (2017). Why and when can deep—but not shallow—networks avoid the curse of dimensionality: a review. arxiv:1611.00740.
Quiroga, R. Q., Reddy, L., Koch, C., & Fried, I. (2007). Decoding visual inputs from multiple neurons in the human temporal lobe. *Journal of Neurophysiology*, *98*(4), 1997–2007.
Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.
Sharir, O., & Shashua, A. (2017). On the expressive power of overlapping operations of deep networks. arxiv:1703.02065.
Silver, D., Schrittwieser, J., & Simonyan, K. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*, 354–359.
Stallkamp, J., Schlipsing, M., Salmen, J., & Igel, C. (2012). Man versus computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, *32*, 323–332.

Tishby, N., & Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. arxiv:1503.02406.

Trappenberg, T. P. (2014). Growing adaptive machines. In *A brief introduction to probabilistic machine learning and its relation to neuroscience* (pp. 61–108). Springer, Berlin.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *49*, 433–460.

van Fraassen, B. C. (2008). *Scientific representation: Paradoxes of perspective*. Oxford: Clarendon Press.

von Melchner, L., Pallas, S. L., & Sur, M. (2000). Visual behaviour mediated by retinal projections directed to the auditory pathway. *Nature*, *404*, 871–876.

Voosen, P. (2015). The believers. *Chronicle of Higher Education* 61(24).

Weisberg, M. (2013). *Simulation and similarity: Using models to understand the world*. New York: Oxford University Press.

Weisberg, M. (2015). Biology and philosophy symposium on simulation and similarity: Using models to understand the world. *Biology & Philosophy*, *30*(2), 299–310.

Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arxiv:1609.08144v2.

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Yu, D., & Deng, L. (2011). Deep learning and its applications to signal and information processing. *IEEE Signal Processing Magazine*, *28*(1), 145–154.

Synthese
https://doi.org/10.1007/s11229-019-02271-0

Data science and molecular biology: prediction and mechanistic explanation

Ezequiel López-Rubio[1,2] · **Emanuele Ratti**[3]

Received: 14 March 2018 / Accepted: 28 May 2019

Abstract

In the last few years, biologists and computer scientists have claimed that the introduction of data science techniques in molecular biology has changed the characteristics and the aims of typical outputs (i.e. models) of such a discipline. In this paper we will critically examine this claim. First, we identify the received view on models and their aims in molecular biology. Models in molecular biology are mechanistic and explanatory. Next, we identify the scope and aims of data science (machine learning in particular). These lie mainly in the creation of predictive models which performances increase as data set increases. Next, we will identify a tradeoff between predictive and explanatory performances by comparing the features of mechanistic and predictive models. Finally, we show how this a priori analysis of machine learning and mechanistic research applies to actual biological practice. This will be done by analyzing the publications of a consortium—The Cancer Genome Atlas—which stands at the forefront in integrating data science and molecular biology. The result will be that biologists have to deal with the tradeoff between explaining and predicting that we have identified, and hence the explanatory force of the 'new' biology is substantially diminished if compared to the 'old' biology. However, this aspect also emphasizes

Both authors have contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11229-019-02271-0) contains supplementary material, which is available to authorized users.

Ezequiel López-Rubio
ezeqlr@lcc.uma.es

Emanuele Ratti
mnl.ratti@gmail.com

1 Departamento de Lógica, Historia y Filosofía de la Ciencia, Universidad Nacional de Educación a Distancia (UNED), Paseo de Senda del Rey 7, 28040 Madrid, Spain

2 Departamento de Lenguajes y Ciencias de la Computación, Universidad de Málaga (UMA), Bulevar Louis Pasteur 35, 29071 Málaga, Spain

3 Reilly Center for Science, Technology, and Values, Department of Philosophy, University of Notre Dame, Notre Dame, USA

Published online: 31 May 2019

the existence of other research goals which make predictive force independent from explanation.

Keywords Biology · Data science · Machine learning · Explanation · Prediction

1 Introduction

In the last few years, the intersection of data-intensive science and biology have attracted the interest of philosophers, historians and STS scholars (Leonelli 2011, 2016; Strasser 2011; Ratti 2015; Boem and Ratti 2016; Stevens 2013, 2015, 2017),[1] in particular where cognitive strategies and modes of knowledge production are concerned. However, how data science have shaped the desiderata and the aims of bioscience has not received enough philosophical attention. The impression is that data science changed something profound about the epistemic units of analyses, the conceptual tools and the aims of biological science itself. This concern applies in particular to molecular biology (Alberts 2012; Golub 2010; Callebaut 2012), which is the field investigated in this article. In particular, it is not clear which are the effects that employing data science techniques have for the status of typical outputs—like models and their aims—of molecular biology. In this paper, we will identify the changes that data science has stimulated in molecular biology when the characteristics and goals of models and modeling are concerned.

After having identified the received view on models in molecular biology as *mechanistic* and *explanatory* (Sect. 2), we will identify the aims and scope of data science, and machine learning in particular (Sect. 3). From this we will be able to draw a comparison between the mechanistic and the machine learning perspective. We will show the existence in machine learning of *a tradeoff between prediction and (mechanistic) explanation*, such that when predictive performances increase, the possibility of elaborating a (mechanistic) explanation necessarily decreases (Sect. 3.3). This may help to elucidate a common intuition according to which mechanistic models explain in terms of cause-effect relations (represented via diagrams), while predictions may be derived also from statistical correlations. Finally, we will see whether our philosophical analysis of machine learning, (mechanistic) explanation and prediction applies to the practice of biology (Sect. 4). To achieve such an aim, we will analyze the publications of The Cancer Genome Atlas Consortium (n = 43). This consortium has been one of the first big-science projects at the forefront of the integration between machine learning and biology. The techniques it developed and the results it published have been highly influential for anyone who wants to engage in data science and biology. This means that the analysis of the publications of this consortium may have a significant degree of applicability to the entire field of molecular biology. What this analysis will show is that biologists have to deal with the tradeoff between explanation and prediction that we have identified. This has two consequences. First, it limits the

[1] It should be emphasized that the effects of the use of computers in biology have been studied well before the studies we refer to. The use of computers and computational models has certainly boosted in the last few decades (see for instance Keller 2002, Chapter 8), but these have been mostly used as 'crutches' and instrumentally to achieve the aims that are traditionally ascribed to biology.

explanatory force of the 'new' biology, such that models in data-intensive biology will hardly be explanatory from a mechanistic point of view. Second, it emphasizes the increasing importance of other non-explanatory aims related to this scientific field.

2 Mechanistic models and explanation in molecular biology

In order to understand whether data science has changed the status of models in molecular biology, we should first state clearly the received view on this 'status'.

Molecular biology is a field that investigates the processes and dynamics of biological phenomena at the level of complex organic molecules. Molecular biology aims at making sense of biological phenomena conceived as the result of the activities of those molecules. Complex and less complex organic molecules include nucleotides, polypeptides, lipids, etc. while the activities include binding, releasing, or more complex operations as phosphorylating among others. When we say 'make sense' we mean that molecular biologists aim at explaining a specific class of biological phenomena as nothing *but* the activities, interactions and the organization of those macromolecules. By 'organization', we rely on a basic meaning spelled out by Levy and Bechtel (2013), namely *organization as 'causal connectivity'*, in the sense of "an internal division of causal labor whereby different components perform different causal roles" (p. 243). In other words, molecular biology *explains* biological phenomena by *describing* how these are produced by the way macromolecules interact. These descriptions are given in form of mechanistic models. To sum up, it is common in the literature—especially the so-called 'mechanistic philosophy' (Machamer et al. 2000)—to state that molecular biology aims at (a) *explaining* phenomena (b) in terms of *mechanistic models*. Therefore, models here are mechanistic models, which function as explanations. The explanatory dimension is usually interpreted by practitioners as the main goal of molecular biology (Alberts 2012; Weinberg 1985). However, one may say there are a plenty of mechanistic models that are not necessarily explanatory, or they are not used for their explanatory capacity, but in other ways. For instance, it is possible to use an abstract schema in the form of a mechanism sketch as a starting point to investigate a possible mechanism (Craver and Darden 2013, Ch 3 and 5). The schema itself is not explanatory, but it is a very useful resource to uncover the nature of phenomena. Therefore, a mechanistic model to be mechanistic does not need to be explanatory; however, the emphasis on the need "to transform black boxes (…) into gray boxes (…) into glass boxes" (Craver and Darden 2013, p. 31) often spelled out in term of 'completeness' of mechanistic description (relatively to the purpose at hand) is an indication that biologists still see good mechanistic models as being explanatory.

Before we go further in the received view on molecular biology, three things need to be specified.

First, appealing to mechanistic explanations in the form of mechanistic descriptions do not only have epistemic motivations. In other words, it is not necessarily the case that mechanistic explanations are superior epistemically to other forms of explanation.[2] Actually, there are also pragmatic reasons for these explanatory standards. Among

[2] We do not commit to any particular thesis on this aspect.

the many, one prominent aspect is that mechanistic descriptions provide a useful way for human beings to think about control. In molecular biology material manipulations are important, and mechanistic descriptions provide easier ways to think about those manipulations. There are also historical reasons for this preference towards mechanistic models in biology. As Sloan (2000) has emphasized, mechanistic programs in life sciences after the nineteenth century (and in molecular biology in the twentieth century) have been a product "of the incursion of physical sciences into the work of medical faculties" (p. 7). While entering into details of these pragmatic and historical aspects is beyond the scope of this work, it is interesting to notice that acceptance of epistemic standards do not have necessarily (or only) epistemic reasons.

Next, while molecular biologists commit to the idea of mechanism, they often have in mind a notion of mechanism that may not be as precise as the one specified by mechanistic philosophers. However, our impression is that the word 'mechanism' is not merely a term which replaces the word 'cause'. Biologists are not only interested in identifying causes and causally relevant entities. Appealing to mechanisms means that we also try to find out a more fine-grained causal connectivity between causally relevant entities.

Finally, we must spend a few words on models. We will refer to models in two ways throughout this work. On one hand, the level of abstraction of a model is the degree of detail in the representation of the target (real) system. For example, one may have a model of the circulatory system where the smallest elements are blood vessels. This would be regarded as a high level model, as compared to a model of the circulatory system where the smallest elements are the red and white cells. Therefore, a model can be high or low level, also called coarse grained versus fine grained, as compared to other models, depending on the smallest (atomic) elements that it considers (Gerlee and Lundh 2016, pp. 32–33). On the other hand, the size of a model is the number of elements that it contains (Gerlee and Lundh 2016, p. 48, simple vs. complex models, also p. 73). For example, in a model of the nervous system where the smallest elements are the neurons, the system size could be the number of neurons. That is, a large model is one with many atomic elements, irrespective of its level of abstraction (high or low). In this work, we are interested in comparing handcrafted, mechanistic, small size models versus automatically generated, machine learned, large size models. We do not address the distinction between high level (systemic) versus low level (local) models.

2.1 Mechanistic models and desiderata of biological explanations

A mechanism is usually understood as "entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (Machamer et al. 2000, p. 3). More succinctly, a mechanism "is a structure performing a function in virtue of its component parts, component operations, and their organization" (Bechtel and Abrahamsen 2005, p. 423). Entities figuring in a mechanistic model are those taken to be responsible for the production of the phenomenon under scrutiny. Moreover, activities or operations may be organized simply in terms of temporal sequence, even though "biological mechanisms tend to exhibit

more complex forms of organization" (Bechtel and Abrahamsen 2005, p. 424). In addition, mechanisms may involve more than one level of organization.

Outlining the dynamics and the organization of parts and activities of a mechanism is, in molecular biology, providing an explanation of the phenomenon one is investigating. There is a consensus according to which "to give a description[3] of a mechanism for a phenomenon is to explain that phenomenon, i.e., to explain how it was produced" (Machamer et al. 2000, p. 3).

To simplify a bit,[4] a mechanistic model is explanatory when it describes the causal structure of the world, namely how a phenomenon is produced by *constitutively causally relevant entities* (Craver 2007; Kaplan and Craver 2011). Entities that are supposed to be causally relevant must be clearly identified (i.e. they must be real) and their organization—in the sense of 'causal connectivity'—must be clearly spelled out.[5] It is important to emphasize that being explanatory is not a yes/no issue, but rather it is by degree; the depth of explanatory power of mechanistic models will vary a lot, depending on the circumstances.

An important indication of the 'depth' of explanatory mechanistic models is their being predictive, in the sense of "the ability to answers questions about how the mechanism would behave if different parts of the mechanism were to work differently, or if they were to break" (Craver and Darden 2013, p. 91). If the model is explanatory, then we must be able to anticipate what would happen if things were slightly different. These predictions are usually qualitative, and they can be corroborated through interventionist strategies. Importantly, interventions must be concrete, not only possible; the predictive force (in the qualitative sense) of a mechanistic model is to be measured by material and concrete interventions—interventions that we can do right now. If a mechanistic model affords predictions that hold in the real world, then clearly the model itself is on the right track to more explanatory power. However, being predictive is only a necessary condition for being explanatory—in fact, we can imagine plenty of examples of very uninformative and general predictions holding true for very abstract and general mechanistic models. But still being predictive is necessary—if predictions do not hold, then it is likely that we just got the phenomenon wrong.

2.2 Understanding and explaining

While there are many accounts of mechanistic explanations and desiderata for mechanistic models, the remark we just did about organization and richness of details may not apply to all of them. Here we specifically focus on Craver's and similar accounts.

An important aspect of mechanistic explanations that needs to be flagged (especially for the present work) is the fact that mechanistic models must be *intelligible* in

[3] 'Description' does not mean necessarily a linguistic description; actually, the preferred means for expressing mechanistic description are diagrams (see Bechtel and Abrahamsen 2005 for preliminary arguments on this matter).

[4] It is not necessary to recall in detail the complex structure of mechanistic explanations; for a full-fledged account of this issue, see (Craver 2007).

[5] There is an interesting debate on what 'relevant' might possibly mean and the difference between leaving out relevant details from the model (i.e. incomplete models) and abstracting from irrelevant details (i.e. abstraction). See in particular (Levy 2014; Levy and Bechtel 2014; Love and Nathan 2015).

order to be explanatory. Biologists *elaborate* such descriptions, and how entities and activities are organized should be clearly discerned by modelers, especially because modelers are *prior* to the model. For a model to exist modelers have to build it, and to elaborate it, modelers have to understand the model itself. This issue is raised here and there in the literature, but it has been rarely tackled properly. A crucial example here is Craver (2006). He mentions this problem when he talks about the notion of 'complete description of a mechanism'. In particular, he says that "such descriptions [i.e. complete descriptions of mechanisms] would include so many potential factors that they would be unwieldy for the purpose of prediction and control and *utterly unilluminating to human beings*" (p. 360, emphasis added). In other words, Craver is saying that too complex mechanistic models would not explain, because such a complex model would be *non-intelligible* (though he does not use this exact word). The process of building a mechanistic model where too many details appeared would be unsuccessful, because the modeler would be unable to find out the causal structure and connectivity between components. Therefore, too many details are likely to prevent the construction of mechanistic models. This is why Craver praises the ability of modelers to understand which details should be taken into account. Hence, it seems there is a connection between the ability to explain and the ability to understand models.

2.2.1 'Build-it' test and intelligible models

Let's explore Craver's intuitions. What is the relation between (mechanistic) explanation and understanding? The debate on scientific understanding has experienced a significant growth in the last few years. Due to lack of space, it is not possible to summarize the main themes of this debate. However, we can draw some distinctions that are important for the topic at hand. First, we should distinguish between *understanding of natural phenomena*, and *understanding of the models about phenomena* (Rice 2016). One can understand a model or a theory about x, but if the model/theory for some reasons is not adequate, then one does not understand x, as in the case of phlogiston theory and the phenomenon of combustion. While most of the debate on understanding tries to flesh out the notion in the former sense, here we are interested in the latter. In particular, we make use of De Regt's framework (2009, 2015, 2017), though modified according to the present context.

De Regt aims at deciphering the epistemic importance of what he calls *pragmatic understanding*, i.e. the *intelligibility of a theory*, defined as being able *to use* the theory.[6] Since he grounds his analysis in the framework of models as mediators between theories and the world provided by Morgan and Morrison (1999), according to his perspective a theory is being used especially for the construction of models. Here we avoid any reference to theories, and we focus on models.

When it comes to 'mechanistic models', their intelligibility is related to the use we make of them. Being able to use a model covers a plethora of things that a scientist can do with a model, and in particular what Craver and Darden (2013) call the *build-it test*. Mechanistic models have been understood, among the other things, as 'recipes for

[6] De Regt's characterization of understanding is *much more rich* than this, but for the purpose at hand this definition is sufficient.

construction'. Recipes describe a set of operations performed on certain ingredients in a way that they will produce a specific thing (e.g. a cake). Mechanistic models afford something similar; they describe how specific entities, if interacting in a specific and organized way, can result in a specific phenomenon. Therefore, Craver and Darden point out that the *successful* ability to modify at will a specific experimental system on the basis of the 'instructions' provided by mechanistic models can be interpreted as a sign that the mechanistic model is somehow describing the phenomenon in a plausible way, i.e. the model explains. Therefore, by 'building' the phenomenon according to the model, we understand if we have an adequate explanation. This is especially important when it comes to models that only highlight that some entities are causally relevant, but without specifying how the entities influence each other to produce the phenomenon; a *how-possibly* model including only a catalogue of entities and activities is not enough, because we would not have instructions whatsoever as to how to make them interacting. Similarly, a *how-possibly* model depicting in the right way the organization of the mechanism, but failing to identify the relevant entities and activities will be useless as well, because we would not know which entities and activities need to be modified.

But the test is not useful only as a 'confirmation tool'. Actually, it is very useful also when we elaborate the model itself. By doing the *build-it test*, we obtain useful hints as to how we are on the right path—this is related to aspect about prediction highlighted before, and the concrete and material interventions required. But in order to elaborate such a test, the model needs to be *intelligible*. By developing Craver's concerns (2006) about models that are too detailed, if in the sketch of the mechanism there are too many entities and activities, then it becomes really complex for the scientist to find out how they are organized or, to use Levy and Bechtel's jargon (2013), it is very difficult to uncover *causal connectivity*. How do we start to stimulate or inhibit entities if we have too many of them? Which entity is to be ascribed causal relevance if we have a long list of potential ones? This means that if we do not strike the right balance between relevant entities and activities and the number of variables that the human mind can deal with, then it is unlikely that the elaboration of a mechanistic model will be successful. This means that too complex (in terms of number of variables, i.e. entities) how-possibly models are unlikely to be turned into mechanistic explanations. To use the terminology introduced above, if the size of the mechanistic model is overwhelming, it is unlikely to become explanatory—it will be remain a how-possibly model.[7] This argument has been spelled out in greater details in another work we have co-authored (Ratti and López-Rubio 2018). In that article, we translate De Regt's account of intelligibility in the mechanistic context, and by relying on studies of cognitive psychology, we make the argument that too complex models cannot be possibly intelligible because of cognitive limits of human agents. They define complexity as a function of the number of model components included in the actual model. Too complex models are

[7] Please note that it does not mean that large models cannot be used to draw general predictions that can be also verified experimentally. For instance, if you model protein–protein interactions with network science, you will obtain a very large model that cannot be turned into an explanation—it is impossible to draw the exact causal narrative connecting all the entities. However, network science tools identify central hub, and one can draw the very general prediction that, if I knock-down a central hub, then the network—and hence the biological phenomenon—will be disrupted. However, this prediction does not help any researcher in elaborating a mechanistic explanation.

not intelligible, and they cannot be turned into explanations. But one may rely on more liberal accounts of mechanistic models and explanations, thereby making the case that very complex models can be somehow explanatory. However, if we are too liberal in saying what mechanistic models are explanatory and what not—and even what counts as a mechanistic model in the first place—, then we may lose the very reasons why we needed a mechanistic account for models in the first place.

To sum up, mechanistic models have to *intelligible* to the modeler in order to be explanatory, where by 'intelligible' we mean De Regt's preliminary notion of pragmatic understanding, but remodeled for the present context, i.e. *understanding a model is being able to successfully use it*. If we consider such a test as a model-developing tool, there are important consequences. In order to transform a how-possibly model into a how-actually model (or at least a how-plausibly), we need to be able to experimentally stimulate (e.g. inhibitory and excitatory strategies, see Bechtel and Richardson 2010) the phenomenon under investigation, and we do this on the basis of the strategies suggested by the model itself. Therefore, if the size of a *how-possibly* or *how-plausibly* model is too big (e.g. too many components)—and hence it is unintelligible—then there is no way to successfully turning it into an explanatory model which specifies how the process occurs.

3 Data science, machine learning and the value of predictions

In the previous section we have identified the characteristics and aims of models and modeling in molecular biology. We have emphasized that molecular biologists look for explanations of biological phenomena, where explanations have to be understood as mechanistic models. Such descriptions stress the causal structure/connectivity holding between components producing the phenomenon we want to explain. We have also emphasized that in order to build explanatory models and to 'test' the adequacy of models, these have to be *intelligible* to the user, where this is measured by the ability of the user to use the model in various ways. This means that explaining and understanding are strictly related, at least in mechanistic models. Therefore, the question about the relation between data science and biology is exactly the question of whether data science modifies the picture just sketched. In order to understand this, in this section we introduce data science, its aims and what motivates them.

3.1 Data science, algorithmic models and machine learning

Dhar defines data science as "the study of the generalizable extraction of knowledge from data" (2013, p. 64). In this definition, *data* is a set of samples, such that each sample is a realization of the variables whose joint probability distribution underlies the observations. Also, *generalizable* means that the patterns automatically extracted from the available samples are expected to occur in other samples of the same process under study. *Knowledge* is not meant in any philosophical sense. Since the idea is to extract patterns between variables, knowledge here simply refers to such patterns.

In data science, patterns are understood in terms of predictions, and they are extracted from samples starting from a *problem*. For the purpose of this work, *a problem* is a set of *input variables* (available at the time that the prediction must be delivered), a set of *output variables* (to be predicted, and hence not available at prediction time), a set of samples (previously observed input–output pairs), and a set of *real-world situations* where the dependency among inputs and outputs can be assumed to be the same.[8] For example, evaluating a mortgage application from a potential borrower is a problem. The inputs could be the address, year built and price of the real estate property, whether the applicant has previously succeeded in repaying a previous loan, her annual income, etc. The outputs could be whether money should be lent to the applicant, and if so, the maximum amount, for how many years and the interest rate. *Data science aims to discover the quantitative relations between inputs and outputs that can possibly hold also in the set of real-world situations assumed to be similar to the one depicted in the samples*. Such relations between inputs and outputs—in the form of a *predictive model*—are obtained by elaborating and applying algorithms. For our purposes, *an algorithm* is a sequence of actions such that, given a certain kind of input, calculates an output in a finite time, after the execution of a finite number of steps using a finite amount of memory.[9] *Prediction* here is the computation of the values of the output variables for an input whose associated output is unknown (though, as we will see, model construction by algorithms is based on cases where similar pairs of input–output are available). In what follows, 'prediction' will have this meaning unless specified otherwise.

Data science may include several disciplines. Here we are interested in *machine learning* (ML). This is the discipline that is central to the so-called 'data revolution' or 'data deluge' (Leonelli 2012, p. 3). ML enables the automated construction of predictive models, which is the reason of its central role. Very broadly, ML is a branch of data science devoted to the development of computer systems that can improve their performance for a given task progressively as more samples are provided, where performance is quantitatively measured. Usually this performance improvement comes from the execution of an algorithm (a *learning algorithm*[10]) that builds and refines a model. Since such models are built by applying algorithms, we will call them *algorithmic models*. An algorithmic model is a model quantitatively depicting the relation between variables, and it is built starting from a set of samples and an

[8] There are similarities with the notion of problem representation in a problem space as famously indicated by Newell and Simon (even though here we refer to the formulation made by Bechtel and Richardson). Bechtel and Richardson's concept of problem is an instantiation of the four elements of the problem representation (initial state, goal, defining moves, path constraints). For machine learning, a problem is an instantiation of an underlying input–output relation (a function) which has been sampled in order to obtain the dataset to be supplied to the learning algorithm.

[9] While there are algorithms that do not halt for some inputs or may require an unlimited amount of memory, they are not used in data science, so we will ignore them in what follows.

[10] Algorithms in machine learning can be supervised, unsupervised and reinforcement learning. We will refer to supervised and unsupervised algorithms because they are relevant to 'discover' quantitative relations between inputs and outputs. On the other hand, reinforcement learning algorithm are the less relevant here, which are used especially in engineering rather than natural or social sciences.

algorithm.[11] The idea is that we infer a quantitative relation between already available pairs of inputs and outputs that will be used in similar cases to generate a predicted output where only the input is available.

ML uses a learning algorithm to build an algorithmic model from a set of *training samples*, called the *training set*. The time sequence is as follows: first a set of data is collected, then an algorithm is chosen to be run on those data, and finally the algorithm is run on the input data, so that the algorithm generates an algorithmic model. Several learning algorithms can be run on the same training set to yield alternative algorithmic models. Then, a set of *validation samples* which were not available at training time is used to measure the performance of each model, so that the algorithmic model which performs best on the validation set is selected. After that, a third test set, which is disjoint with both the validation and test sets, can be employed to carry out a *final assessment* of the selected model. These training, validation and test procedures are usually repeated many times (*replications*) with different random splits of the available data into training, validation and test sets. Through this process, the statistical confidence on the performance of the selected model is improved, as measured by suitable statistical tests.

3.2 What are algorithmic models of machine learning about?

Technically speaking, ML algorithmic models are *predictive models.*[12] When we apply a selected algorithmic model[13] to an input that was not used to train the algorithm in order to produce a predicted output, then we are generating a new prediction. Predictions refer to the local context of a given problem. That is, a model which has been learned by a learning algorithm to predict whether a mortgage borrower will repay the loan is not expected to perform well if used to try to predict whether a student will repay a student loan.

An algorithmic model is designed to predict the behavior of a single target system, which is defined *as an aspect of reality which is associated to a probability distribution of possible cases from which data samples can be drawn at random, in order to obtain the training and validation datasets, and from which the test cases also come*. In addition, the role of machine learning can be further extended by automatically *discovering classes*[14] which would be later predicted. In biomedicine this is pretty common, and classes of (for instance) mutated genes and associated diseases might be later used in order to predict prognoses for new patients (Akbani et al. 2015, pp. 1687–1688).

[11] Here we must remark that each learning algorithm builds a different kind of algorithmic model. Some learning algorithms build complex models composed by many submodels which are combined by a consensus subalgorithm. These are called ensemble algorithms and models.

[12] "The generalization performance of a learning method relates to its prediction capability on independent test data. Assessment of this performance is extremely important in practice, since it guides the choice of learning method or model, and gives us a measure of the quality of the ultimately chosen model" (Hastie et al. 2009, p. 219).

[13] The algorithm produces the model, which in turn is applied to test new data.

[14] Given a universe of objects, a class is a subset of the universe whose elements share some features which makes the class relevant for some scientific or technological purpose.

Algorithms used in ML build algorithmic models of problems whose predictive accuracy typically enhances as more data are supplied.[15] We have reported in another work (Ratti and López-Rubio 2018) the example of the algorithm PARADIGM (Vaske et al. 2010). This algorithm is used in biology to predict the state of molecular biological entities that for some reasons cannot be measured directly. In particular, PARADIGM is aimed at inferring how genetic changes in a patient influence genetic pathways. Pathways are collections of molecular entities that are causally relevant to higher cellular processes. Because of biological complexity, pathways are large and complex. Therefore, it is impossible to keep in mind all the possible entities and their relations into mind—there are databases that keep track of all the details. PARADIGM integrate several genomic datasets and will predict whether a patient will have specific pathways disrupted given the status of some measured entities (e.g. specific genes mutated). The model generated by PARADIGM will specify the expected state of several biological entities which are not experimentally measured, by drawing the information from databases. We have here the characterization of ML through the idea of a problem; there is the input (i.e. the measurements on patients), the samples (i.e. the database), and the output (i.e. the status of entities given inputs and samples). Intuitively, the larger the inputs and samples, the more precise the algorithm performances will be, because by adding more information the algorithm will be able to perform more precise calculations on the expected status of unmeasured entities.

Models generated by ML are *black boxes* (Frické 2015, p. 655). 'Black-box' is a central concept to understand ML's focus on predictions. By 'black-box' model we mean models that *while being precise in associating inputs with outputs, they are nonetheless unable to clearly depict the (causal) relations between them* (Hastie et al. 2009, pp. 351–352).[16] Consider an example. Let's say that we have a model elaborated on the basis that anytime there is a specific set of customer characteristics, it is very likely that the customer population with such characteristics will buy a specific type of product. Let's say that we can calculate quite precisely the 'very likely'. This model is said to be 'black-box' because even if it can clearly point out that there is a quantitative relation between specific inputs (i.e. the set of customer characteristics) and outputs (i.e. the probability of buying a specific product), still it is not at all able to uncover the causal connectivity between the characteristics of the customers and the mental processes which lead them to buy the product. In other words, we know *how to associate* inputs and outputs, but we do not know *what is in between*. Machine learning algorithms generate models that 'black-box' the relation between inputs and

[15] Abundance of data facilitates a better model selection and assessment, because the estimations of the performance of a model are more accurate (Hastie et al. 2009, p. 222). Infrequent but relevant cases can only be observed in sufficient numbers if the number of samples is large enough (Junqué de Fortuny et al. 2013, p. 216). For an important kind of algorithmic models such as maximum likelihood estimators, there are formal proofs that under mild conditions they are both asymptotically unbiased, i.e. the bias reduces to zero as the number of samples tends to infinity; and asymptotically efficient, i.e. the variance reduces to the lowest possible for any model as the number of samples tends to infinity (Sugiyama 2015, pp. 140–143). In some fields such as natural language processing, as the number of samples increases, a point is reached where all relevant cases are sufficiently covered in the dataset, so that memorization of examples outperforms models based on general patterns (Halevy et al. 2009, p. 9).

[16] Please note that this concept of 'black box' differs from Latour's, who refers to ready-made computer systems which are assumed to perform their function correctly (Latour 1987, pp. 3–4).

outputs, though they are very specific in saying that *there is* a relation. This association between inputs and outputs without saying what's in between may be interpreted in a mechanistic framework as the fact that such black-boxed models are not explanatory.[17]

However, one may say that the algorithmic black-boxed models generated by ML may be turned, in principle, into 'white-box' models, namely models where the causal connectivity between components is uncovered. This is usually what we do with how-possibly mechanistic models. Let's imagine that we have a straightforward correlation between the mutations of certain genes and specific phenotypes. Let's hypothesize that we are working with a very well known human gene, such as *PTEN* or *KRAS*. Since we already know the processes that one of these genes can possibly trigger, we can elaborate a causal narrative (see Ratti and López-Rubio 2018)—though idealized and/or abstract, of course—connecting such genes with the phenotypes we are observing. We can refine the causal history by instantiating well-known experimental strategies for discovering mechanisms, such as the ones mentioned in (Bechtel and Richardson 2010; Craver and Darden 2013). By depicting the causal connectivity between components, we try to make sense of why there is a relation between inputs and outputs and hence we say what's in between inputs and outputs. In other words, by connecting inputs and outputs we usually obtain an *explanation*. However, an algorithm used in the context of ML does not do anything like that; it does not yield an explanation, it just points out a connection between two entities that must be explained by experimental strategies for discovering mechanisms.

3.3 Predictivism in machine learning

The appeal to predictive performances is motivated by specific epistemological reasons. Even if ML is oriented towards predictions (a quantitative relation between an input and an output), one may think that, at least in principle, this does not prevent this discipline from being oriented towards mechanistic-like explanations as well. However, this is not the case; ML is *only* about predictions. This is not a sort of *predictivism* (in the sense of Kaplan and Craver 2011); predictivism in ML is *unavoidable.* In other words, a predictivist would say that a scientific model which conveys accurate predictions counts as an explanation (Kaplan and Craver 2011, p. 605), while ML focuses on producing models with a good predictive performance without addressing the question of whether the obtained models have an explanatory value.[18] In any case, ML models are not required to be simulations of the real world processes that they are associated to. In order to explain why ML focuses on prediction, first we will say when and how we need machine learning. Next, we will say why the problems that ML tackles can be approached just by appealing to predictions.

[17] Any how-possibly model is in a sense a black-box model, because it establishes that there are components that are clearly involved in a phenomenon, but we do not know exactly how.

[18] There are some machine learning methods like Bayesian networks (Spirtes et al. 1993) which can learn causal connections among variables from data. They can ascertain that some variables are causes of other variables, but they cannot say anything about the specific mechanism that is behind such causal connection. In other words, Bayesian networks by themselves cannot produce any explanations, since the specific mechanisms must be found by the scientist.

3.3.1 When and how we need machine learning

We need ML when black-boxes are the best options to discover something about target systems.[19] This is more likely to happen in disciplines where the system under study is complex or chaotic. For example, epidemiology and social sciences are amenable to black boxes because the underlying processes involve a huge number of relevant variables with highly nonlinear dependencies among them.[20] Such complexity makes it very difficult to organize variables in a coherent mechanistic-like model. Either we simplify the model or we tackle complexity and settle for black-boxes. This is because even if we could elaborate intelligible models of chaotic and complex systems, these would likely be oversimplifications with limited predictive performance in practical settings. For example, if you want to predict the probability that a person will develop osteoporosis, you may consider a simple model which considers that such probability is a linear function of the age of the person. This is certainly intelligible, and it could provide a first approximation to the problem. But if you need better predictions and a richer (in terms of details) model, you will have to opt for a more complex and (necessarily) less intelligible model where all the risk factors such as age, gender, family history, bone structure, body weight, broken bones, ethnicity, smoking, alcohol, medications, and previous related diseases interact in nonlinear and non obvious ways. Moreover, if the mechanism under investigation involves a complex relation among variables, then an input–output relation cannot be written down in intelligible terms. This can happen even if the number of variables is small, i.e. for small size models, depending on the intrinsic complexity of the relation among the variables in the system. Under these circumstances, only unintelligible algorithmic models can yield good predictions.

3.3.2 Machine learning performances and the bias-variance tradeoff

We need ML when complex systems are our targets, and ML is oriented towards predictive models. What is the best predictive model in this context[21]? This is a pretty technical matter; there are theoretical difficulties in obtaining a model with high predictive performances. This difficulty can be explained by appealing to the so-called *bias-variance tradeoff*. Let us see what this is about.

Following Shmueli's notation (Shmueli 2010), for a given problem there is an abstract level of reasoning where the existence of an abstract function F is postulated, which relates some abstract concepts X to other abstract concepts Y, where $Y = F(X)$. Then, an operationalization must be carried out to translate these abstract concepts into

[19] The real-world process and the model are categorically different. For trained machine learning models, they might not even be structurally similar, depending on the kind of algorithm that is used to learn the model. There are algorithms that aim to learn the structure of the biological interactions, which can be regarded as white box algorithms, while other algorithms do not try to yield a model which resembles the target biological system.

[20] Well known cases include stock market prediction, modeling the spread of communicable diseases on a population, and recommendation systems for online marketing.

[21] Please note once again that predictions here do not necessarily overlap with predictions in the mechanistic context.

measurable variables X, Y linked by the function f which captures the exact relation among them, $Y = f(X)$. The last step is the construction of a model f', which is an approximation of f because it cannot capture all the details of the phenomenon under study. Here the phenomenon must be associated to a single target system from which samples (X, Y) can be drawn at random. In other words, a phenomenon is associated to a specific probability distribution $p(X, Y)$ which can be sampled by observing the target system. Data science comes into play whenever the abstract function F is too complex to be ascertained by a human[22] (Pietsch 2015, p. 915; James et al. 2013, p. 19), or it cannot be used to derive a suitable predictive model f'. Data science proceeds by learning f' from examples of pairs X, Y (James et al. 2013, p. 17) and it considers f' as an acceptable model of reality until a better one is found, even if f' does not provide an explanation about how f works.[23]

How does data science compare models f_i' to choose the best one? In this context, most model performance criteria try to strike a balance between the complexity of the model and how well it fits the data (Bishop 2006, pp. 32–33). At the heart of this issue is the *bias-variance tradeoff*. A well-known performance criterion for predictive models is *the expected error* which is committed when using a model f' for prediction. This error is the sum of three non-negative quantities (James et al. 2013, pp. 33–35; Hastie et al. 2009, pp. 223–228); *the irreducible error*, which stays the same no matter how accurate the model is; *the bias*, which is caused by the difference between the true function f and its approximation f'; and *the variance of the learning method*, which is caused by the oscillations of f' as the training dataset is changed. *The best model is the one for which the sum of bias and variance is very low*. However, this is very difficult to obtain.

Imagine that we wish to predict the ambient temperature tomorrow at 0:00 GMT for all locations on the Earth surface, given the temperatures measured today at 0:00 GMT in all weather stations in the world, so that the problem has one input variable for each weather station. We might use a model f_1' which considers that the average temperature is the same on all points of the Earth surface, and it is computed as the average of the recorded temperatures on all stations. This model would have a high bias, since it is far from the real function f. For example, it would grossly overestimate the temperature at the polar regions, and it would underestimate at the deserts. We may also use another model f_2' which divides the Earth surface into 510 patches of about 1 million km^2 (the surface of the Earth is about 510.1 million km^2), and then predicts the temperature at each patch to be the average of the recorded temperatures on all stations within the patch. The model f_2' would have a smaller bias than f_1', since the overestimations and the underestimations would be less severe. However, f_2' would have a higher variance than f_1', because patches for f_2' contain fewer stations, so that any random changes in the temperature recorded at one station (training dataset

[22] Some data sets are so large and complex that it would be almost impossible for a human to find significant patterns without the help of algorithms which automatically elaborate models to fit the data (Dhar 2013, p. 68).

[23] A difference between the desiderata of models f' for explanation-based and data science approaches is that for the former we want f' to have a causal structure that is similar (not understood in the philosophy of science technical sense) as much as possible to the phenomenon we are investigating, while for data science the predictive performances for new cases is the main goal.

change) would cause a larger oscillation in the temperature prediction for the patch that the station belongs to, since those random changes happen independently from one station to the others. There could be a third model f_3' which divides the Earth surface into 510,000 patches of about 1000 km^2. Now f_3' would have an even smaller bias than f_2', since it would represent even smaller details of the temperature distribution. But f_3' would have a very high variance, since there would be very few weather stations in each of these small patches, so that the computed average on each patch would be even more sensitive to random changes in the temperature recorded at a station, i.e. dataset changes. As we go from f_1' to f_2' and then to f_3' both the model complexity and the variance increase, while the bias diminishes. This is the tradeoff. However, data science techniques *can reduce the variance of highly complex predictive models (while keeping bias low as well) by harnessing large numbers of samples.*[24]

The computational techniques that are used to manage large numbers of samples are also suitable for large numbers of input variables. Therefore, a typical way to enhance the predictive accuracy of a model is to increase the number of input variables that are supplied to the algorithms. The complexity of the learned models usually increase as the number of input variables grows, but this is not a pressing concern for machine learning because both the computer hardware and software are capable of managing large numbers of input variables and samples at the same time.

3.3.3 Why we cannot obtain explanations when using machine learning

Let us summarize what has just been said:

1. ML generates algorithmic models which are predictive models in nature.
2. Algorithmic models are necessary to study very complex systems.[25]
3. The best (predictive) algorithmic model is (by definition) the model which yields the best predictions, and this is that which attains the minimal sum of bias and variance. Therefore, we want to keep both bias and variance low.
4. Keeping both bias and variance low is hard to achieve because of the bias-variance tradeoff. However, through ML, the variance of complex models could be reduced while keeping the bias low by *supplying large amounts of training samples*.
5. The same computational techniques that are used to manage large numbers of samples are also employed to increase the number of input variables, with the aim of further enhancing the predictive performance.

And here come the troubles; *supplying larger and larger numbers of input variables means also making the model even more complex* (i.e. the size of the model increases).

[24] Let's consider a variation of the example of temperatures. We may say that the temperature tomorrow at 0:00 GMT in a particular weather station will be the average of the temperatures recorded on the same day of the year at 0:00 GMT in the same weather station, computed over the available weather data. We can diminish variance as follows. As the number of years with available data (the number of samples) increases, the variance diminishes because the output that *f*' produces for unseen test data will be less sensitive to oscillations in the training dataset, i.e. extremely cold or extremely hot years in the historic record which is used for training.

[25] This makes sense only if we assume that complex systems must be approached by taking into account the contribution of each of their components. There might be other approaches to complexity (e.g. systemic or holistic approaches) which may not require the attitude we are describing here.

The argument can be summarized as follows. If we want to generate reliable models about complex systems we have to use algorithms (as defined above). In this context, a model will be a predictive model, because algorithms of data science obtain exactly that. We do not want just models; we want *good models* in terms of their predictive performance (which, again, is not the same as predictive performances in the mechanistic context). If we want a better predictive model (i.e. one for which the sum of variance and bias is low), we have no choice but to make the model more complex by supplying larger and larger amounts of training data. Since this usually means more variables, *ML works better when the number of variables is overwhelming*, but this also means that in order to connect inputs and outputs we have to identify the causal connectivity or causal structure between an increasing number of components.

This worry can be framed in terms of the relation between explanation and intelligibility that we have delineated in Sect. 2.2 and argued in (Ratti and López-Rubio 2018). An algorithmic model is very unlikely to be 'turned' (or 'upgraded') into an explanatory (mechanistic) model. This is because machine learning algorithms *do not generate intelligible models* (in the sense delineated in Sect. 2.2); there are too many variables, and we do not know how to use the model in order to instantiate a series of 'built-it' tests (in case we are in a mechanistic context) or other procedures such that we can elaborate clearly the causal connectivity (i.e. the organization) between the components. Therefore, it seems that when we use machine learning because of the magnitude of the data sets, at the same time *we cannot even in principle* obtain explanatory (mechanistic) models, but only predictive models. Algorithmic models obtain predictions, and predictions are all we can obtain from them.[26]

A consequence of this is that, under the circumstances where ML is employed, prediction and explanation stand in a *relation of tradeoff*[\]. The tradeoff[27] consists in the fact that the more we use machine learning because we have larger volumes of data, the less our chances of elaborating (mechanistic) explanations will be. That is, the underlying biological processes are so complex that intelligible mechanistic models are oversimplifications of reality. Only more complex, less understandable data-based models can capture the complexity of the underlying biology. Mechanistic models in molecular biology are not very good at prediction (in the machine learning sense) because their bias is very high, i.e. they are far from representing the real target system due to the unavoidable oversimplification that they entail. As explained before,

[26] It can be argued that there are algorithms which learn the structure of the biological system under investigation, so that we can relate somehow the work of algorithms to mechanistic descriptions. For example, the algorithm PARADIGM mentioned above does learn the structure of the interactions among the entities. But it cannot ascertain the specific mechanisms which underlie behind the interactions. So the lack of intelligibility comes from the inability of algorithms to learn those specific mechanisms rather than a dissimilarity between the learned interaction structure and the real one. There are other algorithms such as Prediction Analysis of Microarray (PAM, Tibshirani et al. 2002), which do not intend to learn the biological structure in any way, since they are aimed to prediction only. When algorithms like PAM are used, it means that scientists are not particularly interested in the structures, but in the predictions. In other words, when the biological network under investigation is too complex to obtain an explanatory mechanistic model, then the only option is to use a black box prediction algorithm which does not intend to learn the structure of the analyzed network.

[27] Let us clarify again: this tradeoff between explaining and predicting is a consequence of the way machine learning deals with the bias-variance tradeoff. In other words, this tradeoff and the bias-variance tradeoff are different tradeoffs.

any predictive model must keep both bias and variance low to yield good predictions. Mechanistic models are expected to yield the best predictions when the complexity of the modeled system is small, so that there are no significant oversimplifications in them.

Before proceeding, it is important to stress something important about the tradeoffs (bias-variance, and explanatory abilities-predictive performances) we talked about. In philosophy of science, there has been an interesting discussion on tradeoffs in modeling for a few decades. The discussion originated from Levins' influential paper on model-building in population biology (1966), and it has been recently developed by Weisberg in several papers (for instance Weisberg 2006; Matthewson and Weisberg 2008). Even though Levins does not use the word 'tradeoff', it seems he identified a sort of three-way tradeoff between three different features of models. These features are precision, realism, and generality. By following Weisberg's careful analysis (2006), generality can be defined as the number of target systems to which a model can apply to. Next, realism is how well a model represents its target system. Finally, precision is the "fineness of specification of parameters" (Weisberg 2006, p. 636). The type of tradeoffs identified by Levins are

(1) you cannot increase both realism and precision without sacrificing generality,
(2) you cannot increase both generality and precision without sacrificing realism
(3) you cannot increase both realism and generality without sacrificing precision.

Are these tradeoffs similar to the tradeoffs developed here? We do not see any resemblance between the tradeoff we have identified between explanatory abilities and predictive performances, and Levins' tradeoffs. The tradeoff between bias and variance might deserve a little bit more of discussion. High precision can be interpreted as low variance, since variance measures the variability of parameters as they are learned from training data. High realism might mean low bias, since bias can be pretty much defined in the same way Weisberg characterizes Levins' notion of realism. However, generality is much more difficult to interpret in the context of machine learning. In fact, as it has been specified above, the scope of ML models is not that broad, since the models are learned with specific datasets to solve specific problems. Hence, generality will be usually low. Given the fact that Levins' theory of tradeoff is a three-way tradeoff, and the tradeoff between bias and variance is a tradeoff between two features, we find it difficult to find a straightforward correspondence between them.

4 Biology and machine learning

Because ML masters complexity efficiently, it is likely to be used more and more in the next decades in several scientific fields, including biology. How does molecular biology deal with such a tradeoff? It is time to make explicit the comparison between mechanistic models and machine learning predictive models, especially in light of the tradeoff between explanation and prediction.

In the case of mechanistic models, explanation and understanding (in the sense of 'intelligibility' as specified above) are strictly related; one desideratum of *explanatory* mechanistic models is that they spell out the organization (understood as 'causal connectivity') between the entities and processes producing the phenomenon of inter-

est. Pinpointing the organization implies that we can 'use' (as specified above) the model in many ways, in particular via 'built-it' tests. Therefore, in order to elaborate an explanation in this context, the model must be intelligible to us, both in the phase of construction and confirmation. Hence, explanation and understanding are strictly related to the task of grasping how entities and activities are organized.

On the other hand, predictive models of machine learning are more efficient in their performances when the number of variables increases, i.e. larger size models are preferred.[28] If we use the language of mechanistic models, this means that the number of 'entities' and 'activities' (variables) included in the model has to be high for machine learning algorithms to generate efficient models. But since the number of variables increases, the difficulty of finding out the organization between a high number of variables increases as well; the more machine learning is efficient, the less these models are intelligible. In other words, the more predictive models of machine learning are 'good' models according to the standards of the discipline, the less intelligible they are, and hence the less they can be turned into explanations. *As soon as predictive performance increases, explanatory abilities (and hence power as well) decrease*.

This general analysis suggests that data science (in particular ML) has to impact the aims and goals of any science where explanatory mechanistic models are central. But is this really the case with molecular biology? In order to understand this, we fully scrutinize an important 'big science' project of molecular biology that has been at the forefront in applying machine learning to biological questions. This project is *The Cancer Genome Atlas*[29] (TCGA).

TCGA has started in 2005 and ended in 2016. The main goal of the project was to study the biology of cancer through a massive use of sequencing technologies and bioinformatics. The project background is rooted in the evolutionary view of cancer as genomic instability, which in turn is grounded in the view that there are certain genes (called cancer genes—be they oncogene or tumor suppressor (Morange 1998; Weinberg 2014)) that play a prominent causal role in the development of tumors. The framework of TCGA is the typical reductionist and mechanistic framework that has been guiding molecular biology for at least 60 years (Weinberg 2014; Tabery et al. 2015). The structure of TCGA studies is centered around the use of sequencing technologies to accumulate big data sets (usually about genes, mutations or structural variations both at the genetic and epigenetic levels) which are then analyzed by computer scientists, who rely heavily on the use of data science (and machine learning in particular). In most cases, such studies identify a certain number of genes, mutations and/or structural variations that are then associated in various ways to specific phenotypes[30] (i.e. specific types of tumors). In what follows, we will provide an epistemic analysis of *all* studies published by TCGA, in order to evaluate whether relying on machine learning has, in at least this highly representative case, changed the goals and priorities of molecular biologists.

[28] For network modeling in molecular biology, the number of variables is fixed by the number of detected compounds, so that there is no flexibility to choose the size of the model to be learned.

[29] https://cancergenome.nih.gov.

[30] For a more thorough exposition of the structure of TCGA studies, see Ratti (2015).

The motivation to use TCGA is straightforward. TCGA is a pioneering project in the development of sequencing technologies and bioinformatics tools to adapt biology to the new big data trend, and its studies and its data portal are extensively used as 'models' (both in technical and non-technical senses!) for anyone in contemporary biology who wants to deal with big data sets, especially in cancer biology research.

4.1 The screenings

The information about TCGA studies can be found in Supplementary Table 1. In total, we scrutinize 43 publications. 31 out of 43 are official publications (1–31) of TCGA.[31] The rest of publications (32–43) are part of an initiative called *TCGA Pan-cancer Analysis,*[32] which is a meta-analysis of data sets accumulated by TCGA across different types of cancer. We report different characteristics of the screenings analyzed.

First, we indicate the magnitude of the input analysis, which is usually the number of cancer samples subjected to next-generation sequencing and computational analysis. Numbers may vary considerably, especially in light of the availability of samples for a given cancer. For instance, study 16 has a low number of samples (n = 66) because of the nature of tumor analyzed (i.e. chromophobe renal cell carcinoma) for which the availability is substantially low, especially if compared to more common cancers such as breast cancer (see for instance study 21, where sample is n = 817). When we report the number of samples, we usually specify the number of samples subjected to specific analyses (e.g. whole-exome sequencing, whole-genome, etc.). Studies 32–43 have a high number of samples because they are meta-analyses of different data sets of TCGA consortium.

Next, we indicate the magnitude of the output, namely the amount of mutations, genes or structural variations that are strongly associated with the input. In most screenings data sets are not limited to the genetic or epigenetic levels, and a focus on the proteomic level is provided. This further level of biological analysis is evidence for the data-intensive turn in biology, but our analysis focuses especially on whole-exome, whole-genome and somatic copy-number alterations data.[33] These are, respectively, the mutational analysis of exons across genomes, the analysis of whole genomes, and the analysis of structural variations. Therefore, each study from 1 to 31 (with the exception of study 15) identifies a quantitative relation between a particular phenomenon (i.e. a type of cancer) and specific molecular events/entities (e.g. mutations, structural variations, epigenetic alterations, etc.).

Thirdly, we create a taxonomy of aims and goals for such screenings. In studies from 1 to 32, some have the only aim of understanding in quantitative terms how tumors are mutated, and which mutations are better associated to them (n = 6), while others in addition to this provide what is called a 'pathway analysis', namely a study to

[31] https://cancergenome.nih.gov/publications.

[32] http://www.nature.com/tcga/.

[33] The reason for doing this is that, when mechanistic sketches in this field are outlined, usually genes and mutations at the genetic level are considered central, and hence we decided to focus only on this important level because the epistemic processes applied to it are similar to the ones applied to other levels. Hence the analysis of other levels may be redundant.

contextualize mutated genes within already characterized biological pathways.[34] (n = 10). Other studies provide also attempts to sketch mechanistic descriptions, though by involving only few of the entities found to be associated to the phenomenon of interest, and in this way they identify specific biological processes that may be important for the development of specific types of cancer (n = 15). Studies from 32 to 43 have different aims. They are mostly computational analyses of big data sets across different types of cancer. The most minimal are focused just on 'oncogenic signatures' across cancers (n = 4), while others try to identify possible high-level abstract mechanisms that may be conserved across tumor types (n = 6). Finally, few studies are devoted to the development of specific algorithms (n = 2).

Fourth, we report the number of machine learning algorithms used. We observe that in general several algorithms are used within the same study, and this indicates the fulfillment of different sub-goals within each screening. These sub-goals are identified in two sheets of Supplementary Table 1, named 'Algorithms' and 'Taxonomy'. 'Algorithms' provides a list of the algorithms used and these are associated to the sub-goals identified. On the other hand, 'Taxonomy' quantifies the number of times such sub-goals have been achieved in the screenings analyzed.

4.2 Predictions, explanations and the aims of molecular biology

If our epistemological analysis about explanation and machine learning in the mechanistic context is correct, then we should observe specific things in analyzing the outputs of a project like TCGA. In particular, we should observe that (a) *there is no fully-fledged mechanistic explanation of cancer* in each and every paper of TCGA and (b) *predictive models serve goals that are different (and independent) from the typical explanatory-based aims of molecular biology*. Let us now see whether these two claims that have been derived in principle hold in the actual practice of some data-intensive biology.

First of all, we hasten to add that when we say 'fully-fledged mechanistic explanation of cancer' we do not mean a 'theory' of cancer. There can be a general theory of cancer in the sense of a family of models explaining how instances of cancer may manifest one or more hallmarks that are usually produced in certain ways.[35] What we mean by 'fully-fledged' mechanistic explanation of cancer is that a biologist, starting from the characterization of a phenomenon (a particular type of cancer in x samples), tries to instantiate those discovery strategies identified by mechanistic philosophers (Craver and Darden 2013; Bechtel and richardson 2010) in order to explain (i.e. elaborate a causal narrative) of how that particular phenomenon has been produced. For example, one may start to identify entities and activities that are likely to be involved in the production of the phenomenon to explain, and then recompose these components into a narrative of how these components produce exactly the explanandum (Glennan 2017).

[34] On the difference between 'pathways' and 'mechanisms' see Ross (2018) and Boniolo and Campaner (2018).

[35] Consider for instance the so-called 'hallmarks of cancer' (Hanahan and Weinberg 2000, 2011).

Let us now turn to the extent to which (a) applies to actual biological practice. On the basis of the analysis of TCGA publications, we argue that, in each paper analyzing a specific cohort of samples of a specific type of tumor, the data gathered is not used to explain mechanistically how that particular cancer has been produced. However, one may say that in the past it has been already observed that data-intensive studies in molecular oncology follows a mechanistic recipe. For instance, Ratti (2015) emphasizes that the structure of the discovery strategies of studies such as TCGA's is similar to the ones emerged in the mechanistic tradition (Bechtel and Richardson 2010). However, the study of Ratti (2015) is not very specific in spelling out the explanatory dimension of data-intensive studies in molecular biology. For instance, consider the phase of validation of computational analysis (Ratti 2015, pp. 206–209). This is the phase when a few mutations/genes that are found associated to cancer are 'experimentally' validated to see if they play a causal role in the phenomenon of interest. Experimental validation—also known as the 'functional test' (Bertolaso 2016)—can be interpreted as being part of the discovery strategies of the mechanistic tradition (Bechtel and Richardson 2010; Craver and Darden 2013).[36] However, even though discovery strategies of the mechanistic tradition aim at explanatory goals, a mere experimental validation of the causal role of an entity is just one piece of a very complex puzzle. Numerous studies of TCGA experimentally validate some cancer genes, and they also mention the general processes these entities may be involved into. For instance, in Supplementary Table 1 we emphasize that studies number 6, 8, 11, 13, 16, 17, 18, 19, 20, 21, 22, 23, 24, 30, 31 investigate *possible mechanisms* of a few genes/mutations. Instead of trying to explain through a mechanistic explanation a phenomenon such as how from certain mutations a specific type of tumor will show this and that phenotypes, these studies focus on a few entities that machine learning analysis strongly associates with tumors. This means that there is not an attempt to explain the phenomenon of interest (i.e. a specific type of tumor), but only an exploration of the information about *few* entities that may play important causal roles in one or more tumors. The attention shifts from the phenomenon to explain to some entities that may be involved.[37] In other studies, rather than explaining a phenomenon, information about mutated genes is contextualised in so-called mechanistic schemas (Levy 2014). This is when TCGA practitioners try to make sense of the causal role of one or more genes by showing that what the genes are supposed to do is consistent (by means of computational analysis) with it being part of an already known biological pathway whose disruption in cancer has been observed in the past, even though the

[36] Please note that experimental validation is different from the validation phase of machine learning procedures, where an algorithmic model is chosen according to its performance on a validation set, which is a subset of the available dataset.

[37] One may also say that this is the point where 'data-intensive' studies meet mechanistic studies, in the sense that studies such as the ones of TCGA are only one step towards the elaboration of mechanistic explanations. However, this observation may miss the importance and the role of bioinformatics tools. These are not just tools aimed at selecting a few entities by means of eliminative inductive strategies, but tools that are used to characterize the complexity of biological systems. In fact, we may use these tools to prioritize a few entities and elaborate a simple mechanistic models, but by doing this we would miss the complexity of biological systems and the other analyses that can be done on biological complexity without necessarily narrowing down just a small and local part of it, as we do when we just focus on a few cancer genes as part of the complexity of a particular type of tumor.

role of the genes in the pathway is not clarified at the level of detail that a mechanistic explanation would require (studies number 4, 5, 7, 9, 10, 12 14, 25, 27, 28 only focus on pathway analysis, while the previously mentioned studies have both the focus on new cancer genes and pathway analysis). Schemas are employed notably in some studies of the Pan-cancer analysis of TCGA (studies number 33, 38, 39, 41, 42, 43), where selected entities that are associated with several types of tumors are then investigated with respect to their possible mechanistic contexts.

Therefore, while no study of TCGA does provide a fully-fledged mechanistic description of the phenomenon of interest, some may provide either sketches or schemas but limited to a few entities. This is consistent with the introduction of ML's methodology in molecular biology and our analysis of explanation, understanding and prediction; *when big data sets with an overwhelming number of variables are analyzed, no heavily vetted mechanistic explanation is attempted*, but rather biologists limit their analyses to a few entities.[38] The fact that studies try to provide evidence for new cancer genes is also related to the biomedical dimension of the project; its findings can be used for later phases of drug discovery (Ratti 2016). However, our claim is that if you do molecular biology with machine learning techniques, and if you want to have the best machine learning performances, then you cannot even in principle elaborate fully-fledged mechanistic explanation.[39] It is important to notice that this is not due technological limitations; the argument spelled out in Sect. 2 and in great detail in (Ratti and López-Rubio 2018) is that the more the size of the model increases, the less the human mind is able to organize the model's components into a causal narrative, which forms the backbone of any mechanistic description with explanatory force. Moreover, we would also like to resist the objection that the explanation—even if it cannot be elaborated right now by humans—is still there, present in the complexity of the biological data sets. Our appeal to the capacity of the agents elaborating explanations implies and advocates a sort of epistemic conception of explanations—the explanations is not contained in the how-possibly model and has to be 'discovered', but it is a product of the cognitive abilities of the agent. Therefore, if there are no cognitive agents that can elaborate the explanation, then there is no explanation at all.

Let us turn to (b). By connecting molecular biology to the mechanistic tradition, philosophers of biology have emphasized its explanatory goals. Even though prediction, control and other aims are recognized, these have value only to the extent that they increase the chances of elaborating explanations. However, it seems that TCGA hardly fits this framework. While it is true that predictive models may function as a way to select entities and activities that can be central in mechanistic descriptions (Ratti 2015), this ignores the fact that predictive models play other important roles, which are not necessarily related to explanation and cause-effect narratives. Among the others, predictive models are useful for disease classification. For these purposes, an unsupervised clustering algorithm is first employed in order to discover clusters of

[38] This aspect may be interpreted as being related to the pathway concept, as Ross (2018) points out when she says that "instead of identifying a particular explanatory target and 'drilling down', these maps [i.e. pathways representation] involve identifying a set of entities in some domain and 'expanding our by tracing their causal connections" (p. 13).

[39] One may argue that mechanistic philosophers' requirements for a good explanation are in tension, but this is beyond the scope of the present paper.

tumors according to their genetic signature. This does not necessarily mean that the members of a cluster share the same oncogenic mechanism, but only that the observed mutations are similar, i.e. no detailed explanation of a common mechanism is proposed for the members of a tumor cluster. After that, new tumors can be classified by supervised learning algorithms into these previously discovered clusters. Again, the classification is not dependent on any explanation of the oncogenic mechanisms, since the algorithm classifies the tumors according to their genetic signatures. In turn, this is related to the use of these studies as powerful tools for diagnosis. In order to diagnose a specific disease, you do not need an explanation of the disease itself. In the case of these screenings, we do not need to connect causally or mechanistically a mutational signature to a specific type of tumor. We just need to know that anytime there is a specific pattern of mutations, then there is a precise quantification of the probability of having a specific type of tumor, even though plenty of associations will turn out to be just spurious correlations (as the study number 34 seems to suggest) or just indirectly related to difference makers. This is part of the aims of TCGA and its focus on the so-called *precision medicine*. By creating more and more specific subgroups of the same tumor by means of machine learning analyses, TCGA aim at elaborating highly specific genome profiles that can be next further tailored around the needs of individual patients or very small groups of patients.

Therefore, the analysis of the TCGA publications is consistent with what one could have inferred from our in principle analysis of molecular biology and machine learning, namely that machine learning pushes molecular biology towards slightly different epistemic aims and concerns. To say this more loud and clear, the introduction of machine learning in molecular biology has introduced a change in the molecular biology system of practice (Chang 2014),[40] by shifting the field from purely explanatory practices to a variety of predictive activities that may be even in a tradeoff relation with explanations.

5 Conclusion

In this article, we have analyzed the claim according to which the introduction of data science in molecular biology has somehow changed its epistemic aims.

First, we described the epistemic aims of molecular biology, and we have identified explanatory goals achieved in terms of the elaboration of mechanistic models/descriptions (Sects. 2, 2.1). We have also connected the task of building explanatory models to the intelligibility of such models, such that extremely complex mechanistic sketches are unlikely to be turned in good explanatory models.

Next, we described the epistemic aims of data science, and machine learning in particular. We have emphasized that machine learning points towards the elaboration of predictive models (Sect. 3.2). Moreover, we have also added that machine learning can produce more effective predictive models as the data volume it analyses grows, both in terms of number of samples and number of input variables. This, together

[40] This of course does not mean that more traditional forms of molecular biology cannot possibly coexist with machine learning-driven molecular biology.

with the remarks about intelligibility of models, led to the identification of a tradeoff relation between the ability to elaborate mechanistic-like explanation and predictive models.

We finally showed that in practice molecular biologists have to deal with such a tradeoff; when they attempt to elaborate mechanistic explanations, they necessarily have to narrow their focus to a few of the variables considered to investigate the phenomenon under scrutiny (Sect. 4). However, the predictive force of such narrow explanations will be very limited if compared to the predictive force of predictive (and non-explanatory) models generated by taking fully into account the complexity of the data sets about a phenomenon.

Acknowledgements The authors would like to thank David Teira, Enrique Alonso, and the participants to the workshop "Making sense of data in the sciences" in Hannover, and in particular Federica Russo and Sara Green, for their valuable comments and suggestions. They are also grateful to the editor and four anonymous reviewers for their constructive feedback.

References

Akbani, R., et al. (2015). Genomic classification of cutaneous melanoma. *Cell, 161*(7), 1681–1696.
Alberts, B. (2012). The end of "small science"? *Science, 337*(6102), 1583.
Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 36*(2), 421–441. https://doi.org/10.1016/j.shpsc.2005.03.010.
Bechtel, W., & Richardson, R. (2010). *Discovering complexity—Decomposition and localization as strategies in scientific research*. Cambridge, MA: The MIT Press.
Bertolaso, M. (2016). *Philosophy of cancer*. Dordrecht: Springer.
Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.
Boem, F., & Ratti, E. (2016). Towards a notion of intervention in big-data biology and molecular medicine. In G. Boniolo & M. Nathan (Eds.), *Foundational issues in molecular medicine*. London: Routledge.
Boniolo, G., & Campaner, R. (2018). Molecular pathways and the contextual explanation of molecular function. *Biology & Philosophy, 33*(3–4), 1–19. https://doi.org/10.1007/s10539-018-9634-2.
Callebaut, W. (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 Sciences, 43*(1), 69–80. https://doi.org/10.1016/j.shpsc.2011.10.007.
Carrier, M. (2014). Prediction in context: On the comparative epistemic merit of predictive success. *Studies in History and Philosophy of Science Part A, 45*(1), 97–102. https://doi.org/10.1016/j.shpsa.2013.10.003.
Chang, H. (2014). Epistemic activities and systems of practice: Units of analysis in philosophy of science after the practice turn. In L. Soler, S. Zwart, M. Lynch & V. Israel-Jost (Eds.), Science after the practice turn in the philosophy, history and social studies of science. Routledge.
Cox, D. R. (2001). Comment to 'statistical modeling: The two cultures'. *Statistical Science, 16*(3), 216–218.
Craver, C. F. (2006). When mechanistic models explain. *Synthese, 153*(3), 355–376. https://doi.org/10.1007/s11229-006-9097-x.
Craver, C. (2007). *Explaining the brain - Mechanisms and the mosaic unity of neuroscience*. Oxford University Press.
Craver, C., & Darden, L. (2013). *In search of mechanisms*. Chicago: The University of Chicago Press.
De Regt, H. W. (2009). The epistemic value of understanding. *Philosophy of Science, 76*(5), 585–597. https://doi.org/10.1086/605795.
De Regt, H. W. (2015). Scientific understanding: Truth or dare? *Synthese, 192*(12), 3781–3797. https://doi.org/10.1007/s11229-014-0538-7.
De Regt, H. (2017). *Understanding scientific understanding*. Oxford: Oxford University Press.
Dhar, V. (2013). Data science and prediction. *Communications of the ACM, 56*(12), 64–73.

Douglas, H. E. (2009). Reintroducing prediction to explanation. *Philosophy of Science, 76*(4), 444–463. https://doi.org/10.1086/648111.
Douglas, H., & Magnus, P. D. (2013). State of the field: Why novel prediction matters. *Studies in History and Philosophy of Science Part A, 44*(4), 580–589. https://doi.org/10.1016/j.shpsa.2013.04.001.
Frické, M. (2015). Big data and its epistemology. *Journal of the Association for Information Science and Technology, 66*(4), 651–661.
Gerlee, P., & Lundh, T. (2016). *Scientific models*. Basel: Springer.
Glennan, S. (2017). *The new mechanical philosophy*. Oxford: Oxford University Press.
Golub, T. (2010). Counterpoint: Data first. *Nature, 464*(7289), 679. https://doi.org/10.1038/464679a.
Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems, 24*(2), 8–12.
Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, *100*(1), 57–70.
Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell*, *144*(5), 646–674.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). New York: Springer.
Hempel, C., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science, 15*(2), 135–175.
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.
Junqué de Fortuny, E., Martens, D., & Provost, F. (2013). Predictive modeling with big data: Is bigger really better? *Big Data, 4*(1), 215–226.
Kaplan, D. M., & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science, 78*(4), 601–627. https://doi.org/10.1086/661755.
Keller, E. F. (2002). *Making sense of life: Explaining biological development with models, metaphors and machines*. Cambridge, MA: Harvard University Press.
Latour, B. (1987). *Science in action*. Cambridge, MA: Harvard University Press.
Leonelli, S. (2011). Packaging data for re-use: Databases in model organism biology. In P. Howlett & M. S. Morgan (Eds.), *How well do facts travel? The dissemination of reliable knowledge*. Cambridge, MA: Cambridge University Press.
Leonelli, S. (2012). Introduction: Making sense of data-driven research in the biological and biomedical sciences. *Studies in History and Philosophy of Biological and Biomedical Sciences, 43*(1), 1–3. https://doi.org/10.1016/j.shpsc.2011.10.001.
Leonelli, S. (2016). *Data-centric biology*. Chicago: University of Chicago Press.
Levins, R. (1966). The strategy of model building in population biology. In E. Sober (Ed.), *Conceptual issues in evolutionary biology* (pp. 18–27). Cambridge, MA: MIT Press.
Levy, A. (2014). What was Hodgkin and Huxley's achievement? *British Journal for the Philosophy of Science, 65*(3), 469–492. https://doi.org/10.1093/bjps/axs043.
Levy, A., & Bechtel, W. (2013). Abstraction and the organization of mechanisms. *Philosophy of Science, 80*(2), 241–261. https://doi.org/10.1086/670300.
Lombrozo, T. (2011). The instrumental value of explanations. *Philosophy Compass, 6*(8), 539–551. https://doi.org/10.1111/j.1747-9991.2011.00413.x.
Love, A. C., & Nathan, M. J. (2015). The idealization of causation in mechanistic explanation. *Philosophy of Science, 82*(December), 761–774. https://doi.org/10.1086/683263.
Machamer, P., Darden, L., & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science, 67,* 1–25.
Matthewson, J., & Weisberg, M. (2008). The structure of tradeoffs in model building. *Synthese, 170*(1), 169–190. https://doi.org/10.1007/s11229-008-9366-y.
Morange, M. (1998). *A history of molecular biology*. Cambridge, MA: Harvard University Press.
Morgan, M., & Morrison, M. (Eds.). (1999). *Models as mediators*. Cambridge, MA: Cambridge University Press.
Pietsch, W. (2015). Aspects of theory-ladenness in data-intensive science. *Philosophy of Science, 82*(5), 905–916.
Press, G. (2013). A very short history of data science. *Forbes*. http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/. Accessed 12 June 2016.
Ratti, E. (2015). Big data biology: Between eliminative inferences and exploratory experiments. *Philosophy of Science, 82*(2), 198–218.

Ratti, E. (2016). The end of "small biology"? Some thoughts about biomedicine and big science. *Big Data & Society,* no. July–December:1–6.
Ratti, E., & López-Rubio, E. (2018). Mechanistic models and the explanatory limits of machine learning. In *[2018] PSA 2018: The 26th Biennial meeting of the philosophy of science association (Seattle, WA; 1–4 November 2018)*. http://philsci-archive.pitt.edu/view/confandvol/confandvolPSA2018.html.
Rice, C. C. (2016). Factive scientific understanding without accurate representation. *Biology and Philosophy, 31*(1), 81–102. https://doi.org/10.1007/s10539-015-9510-2.
Ross, L. N. (2018). *Causal concepts in biology: How pathways differ from mechanisms and why it matters*. [Preprint]. http://philsci-archive.pitt.edu/id/eprint/14432. Accessed March 13, 2018.
Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289–310.
Sloan, P. (2000). Completing the tree of descartes. In P. Sloan (Ed.), *Controlling our destinies—Historical, philosophical, ethical, and theological perspectives on the human genome project*. Notre Dame: University of Notre Dame Press.
Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. New York: Springer.
Stevens, H. (2013). *Life out of sequence—A data-driven history of bioinformatics*. Chicago: Chicago University Press.
Stevens, H. (2015). Networks: Representations and tools in postgenomics. In S. Richardson & H. Stevens (Eds.), *Postgenomics—Perspective on biology after the genome*. Durham: Duke University Press.
Stevens, H. (2017). A feeling for the algorithm: Working knowledge and big data in biology. *Osiris, 32*(1), 151–174. https://doi.org/10.1086/693516.
Strasser, B. (2011). The experimenter's museum—GenBank, natural history, and the moral economies of biomedicine. *Isis, 102*(1), 60–96.
Strevens, M. (2008). *Depth—An account of scientific explanation*. Cambridge, MA: Harvard University Press.
Sugiyama, M. (2015). *Introduction to statistical machine learning*. Burlington, MA: Morgan Kaufmann.
Tabery, J., Piotrowska, M., & Darden, L. (2015). Molecular biology. In E. N. Zalta (Eds.), *The stanford encyclopedia of philosophy (Summer 2018 Edition)*.
Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America, 99*(10), 6567–6572.
Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., et al. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics, 26*(12), i237–i245.
Weinberg, R. A. (1985). The molecules of life. *Scientific American, 253*(4), 48–57. https://doi.org/10.1038/scientificamerican1085-48.
Weinberg, R. (2010). Point: Hypotheses first. *Nature, 464*(7289), 678. https://doi.org/10.1038/464678a.
Weinberg, R. A. (2014). Coming full circle-from endless complexity to simplicity and back again. *Cell, 157*(1), 267–271. https://doi.org/10.1016/j.cell.2014.03.004.
Weisberg, M. (2006). Forty years of "the strategy": Levins on model building and idealization. *Biology and Philosophy, 21*(5), 623–645. https://doi.org/10.1007/s10539-006-9051-9.

European Journal for Philosophy of Science (2020) 10:22
https://doi.org/10.1007/s13194-020-00288-8

PAPER IN PHILOSOPHY OF TECHNOLOGY AND INFORMATION

The Big Data razor

Ezequiel López-Rubio[1,2]

Received: 29 July 2019 / Accepted: 24 March 2020 /

Abstract

Classic conceptions of model simplicity for machine learning are mainly based on the analysis of the structure of the model. Bayesian, Frequentist, information theoretic and expressive power concepts are the best known of them, which are reviewed in this work, along with their underlying assumptions and weaknesses. These approaches were developed before the advent of the Big Data deluge, which has overturned the importance of structural simplicity. The computational simplicity concept is presented, and it is argued that it is more encompassing and closer to actual machine learning practices than the classic ones. In order to process the huge datasets which are commonplace nowadays, the computational complexity of the learning algorithm is the decisive factor to assess the viability of a machine learning strategy, while the classic accounts of simplicity play a surrogate role. Some of the desirable features of computational simplicity derive from its reliance on the learning system concept, which integrates key aspects of machine learning that are ignored by the classic concepts. Moreover, computational simplicity is directly associated with energy efficiency. In particular, the question of whether the maximum possibly achievable predictive accuracy should be attained, no matter the economic cost of the associated energy consumption pattern, is considered.

Keywords Model simplicity · Machine learning · Bayesianism · Information theory · Energy efficiency

1 Introduction

Automated model selection is one of the most relevant features of machine learning. It gives the scientist a powerful tool to quantitatively assess the merits of several

✉ Ezequiel López-Rubio
ezeqlr@lcc.uma.es

1 Departamento de Lenguajes y Ciencias de la Computación, Universidad de Málaga (UMA), Bulevar Louis Pasteur 35, Málaga, 29071, Spain

2 Departamento de Lógica, Historia y Filosofía de la Ciencia, Universidad Nacional de Educación a Distancia (UNED), Paseo de Senda del Rey 7, Madrid, 28040, Spain

possible models in the light of their predictive performance when fitted to large volumes of data. In the current Big Data age, many scientific and engineering endeavors involve the execution of machine learning software to obtain a fitted model, with little or no human intervention. This calls for an analysis of the criteria which are employed in such software to choose one model over another. Simplicity is often employed to justify these selections, but the concept of simplicity has different meanings depending on the school of thought that a machine learning practitioner adheres to. Here we aim to explain such differences and their associated underlying assumptions about the goals of model selection. Furthermore, we describe and discuss the computational concept of simplicity, which is frequently applied in practice but often neglected in the literature. Finally, we claim that the computational concept of simplicity is more appropriate to understand current practices in machine learning than the classic ones. In this work, we focus on model selection for machine learning, and not for general scientific inference.

In what follows we will use the following terminology. A model is a mathematical structure that aims to fit some experimental data. A model contains zero, one or more adjustable (also called learnable) parameters, which are real numbers that must be determined. Therefore, a model with one or more parameters has infinite possible realizations,[1] which we call instantiations. A machine learning algorithm takes a model and a set of training data as inputs and returns a set of values for the adjustable parameters of the model, i.e. an instantiation of the model. We also say that the algorithm fits the model to the data so that a fitted (instantiated) model is obtained.

We start by presenting four classic concepts of simplicity which have been applied to machine learning models (Section 2). Then the relations among machine learning models, the learnable parameters that they contain, and the learning algorithms which are used to adjust the parameters are discussed, along with the concept of a learning system, which contains them (Section 3). After that, the proposed concept of computational simplicity is detailed, and its epistemic advantages with respect to the classic ones (Section 4). The relations among the classic and computational approaches of simplicity are studied in Section 5. Next, some non epistemic justifications of the proposed computational simplicity concept are outlined (Section 6). Finally, Section 7 concludes this work by highlighting the relevance of computational simplicity to state of the art machine learning.

2 Four brands of simplicity for machine learning models

Ockham's razor is invoked by machine learning theorists to prefer simpler models:

> Entities should not be multiplied beyond necessity.

It must be understood as a search strategy to extract good models from data, and not as a statement that Nature must actually be simple (de Rooij and Grünwald 2011, p.

[1] An uncountable infinity, since the parameters are real numbers. In the particular case that all adjustable parameters are constrained to be integers, the number of possible realizations is a countable infinity.

893). It has been argued that, even if it does not lead to the true model, it finds models that yield reliable predictions (de Rooij and Grünwald 2011, p. 894).

As an example, let us consider a typical application of machine learning, namely the modeling of customers of an e-commerce site. Given the complexity of human behavior, it is almost impossible to build machine learning models that capture all the details of this problem. Therefore, machine learning practitioners assume that the true model is not attainable. Nevertheless, under the Ockam's razor principle, it is possible that a simple model can be obtained by learning from customer activity data collected by the web site software. Such a model might be good enough to produce accurate predictions of customer actions like buying a certain product or revisiting the site.

For instance, it could be found in the collected data that 90% of the customers that have already purchased items in the site for a total amount of more than 2,000$ actually revisit the site. On the basis of such an observation, a simple decision tree model might be learned by a machine learning algorithm that contains a rule which states that if a customer has purchased more than 2,000$ on the site, then she is predicted to revisit. Here the amount 2,000$ is an adjustable parameter which is learned from the data.

However, there are many possible ways to define simplicity in the machine learning context (Domingos 1999, p. 409; Kelly 2007, p. 270). Next, we review four of them.

2.1 The Bayesian concept

The standard view in the Bayesian school understands probabilities as degrees of belief in the truth of a statement (Sober 2015, p. 64; Forster 2001, p. 88). Bayesian model selection chooses the model which has the highest probability given the data (Wasserman 2000, p. 93). In order to compare two models, their Bayes factor can be computed, which is the evidence of one of the models versus the other (Wasserman 2000, p. 98). This calculation requires the assumption of prior probability distributions for the models and their parameters.

The Bayesian Information Criterion (BIC) is the criterion of choice according to Bayesianism (Sober 2015, p. 135). It is an approximation to the log Bayes factor (Wasserman 2000, p. 100) that has the advantage that no prior information must be supplied since only the maximum likelihood instantiation of each model is necessary (Claeskens and Hjort 2008, p. 81). The BIC is an unbiased estimator of the probability of observing the data given the model (Sober 2015, p. 139). After that, the model associated with the highest posterior probability of observing the data is chosen (Claeskens and Hjort 2008, p. 79). The maximization of this posterior probability automatically leads to a balance between the goodness of fit of the model to the data and the complexity of the model, which is known as the Bayesian Ockam's razor (Huang and Beck 2018, pp. 712-713; Murphy 2012, p. 157). The Bayesian framework favors models with fewer adjustable parameters because they concentrate the prior probability mass in a smaller range of options (Sober 2015, p. 125). In other words, simpler models have a smaller number of possible instantiations, which means

that more prior probability mass is allocated to each instantiation (Henderson et al. 2010, pp. 186-187).

It is sometimes interpreted that the Bayes factors or the BIC select the model which is believed to be true with the highest probability (Grünwald 2007, p. 540). However, this interpretation fails when a model is compared to a restriction of it (a submodel). From the axioms of probability, it follows that the submodel cannot have a higher probability than the model, but Bayesian model selection sometimes chooses the submodel (Forster 2001, p. 95).

Bayesian model selection might be employed to predict whether a customer will revisit our example e-commerce site, given her past activity on the site. Typically the competing models have several learnable parameters that are adjusted from the data collected by the web site software about many customers over time. Given the history of a customer, each model will output the probability that the customer revisits the site. Then the acquired data about the customers who actually revisited the site could be used, so that the BIC applies the Bayesian Ockam's razor to choose the model which attains the best balance between the model accuracy, i.e. how the predicted revisit probabilities match the actual revisit data, and the model complexity, i.e. the amount of parameters to be adjusted from the customer activity data. This selection is based on the maximization of the posterior probability of observing the revisit data.

2.2 The Frequentist concept

The Frequentist approach holds that probability only has a meaning when it refers to a repeatable experiment (de Rooij and Grünwald 2011, p. 895). Therefore, the Bayesian evaluation of a model, based on the probability of the model given the observed data, does not make sense under this interpretation. Frequentism conducts model assessment without considering subjectively assigned prior probabilities, which enhances the perceived objectivity of its conclusions, as compared to Bayesianism (Dawid 2017, pp. 378-381). The AIC is the best known Frequentist criterion (Sober 2015, pp. 128-135). The AIC does not assign probabilities to models since it is based on an unbiased estimation of the predictive accuracy of each model.

The AIC favors models with fewer adjustable parameters because it is based on an unbiased estimate of the predictive accuracy of a model which contains the number of adjustable parameters with a negative sign. This means that the lower the number of adjustable parameters, the higher (better) the AIC. This mathematical fact reflects the experience of scientists when they try to employ a model which fits the training data very well but performs poorly on new test data (Sober 2015, pp. 130-131). In machine learning terms, it is said that a model with too many adjustable parameters overfits the data.

While both AIC and BIC favor models with fewer adjustable parameters, they diverge in their interpretation of Ockam's razor. The main difference between Frequentist and Bayesian model selection is that the Bayesian approach intends to maximize the probability of the model given the data so that it assigns probabilities to models. In the Bayesian context, more adjustable parameters mean lower model probability, and this is provided as the Bayesian foundation of Ockam's razor (Sober

2015, p. 141). The Frequentist approach refrains from model probability assignments, which means that models with more adjustable parameters are penalized on the grounds that they are estimated to have lower predictive accuracy, irrespective of their probability. Therefore, the Frequentist justification of Ockam's razor refers to the estimated accuracy of the models. As mentioned in (Sober 2015, p. 149), model parsimony is used as a surrogate for accuracy. Following our previous account of Bayesian model selection for e-commerce customer revisit prediction, the AIC differs from the BIC in that the AIC chooses the model that is estimated to have the highest predictive accuracy for future customer revisits, irrespective of any model probabilities.

Accuracy is what really matters, and this explains why the AIC and the BIC are less used than direct estimations of predictive accuracy obtained by cross validation. The AIC is a parsimony based, indirect estimator of predictive accuracy (Bandyopadhyay and Forster 2011, p. 3). The BIC does not even aim to maximize predictive accuracy, but it evaluates the evidence for a model given the observed data (Wasserman 2000, pp. 99-100). In contrast to these criteria, cross validation actually measures the predictive accuracy over validation data sets. The predictive accuracy measured by cross validation over validation data sets often turns out to be a better estimator of the performance of a model on a test set than BIC or AIC (Hastie et al. 2009, p. 254). Moreover, cross validation can be used for non probabilistic models (Murphy 2012, p. 370). All of this assumes that predictive accuracy is the primary goal. Therefore, machine learning practitioners mostly adhere to an instrumentalist conception of science (Sober 2015, pp. 143-144). Parsimony is seen as a secondary goal so that if the predictive accuracy is similar among several candidate models, the candidate with the smallest number of adjustable parameters is chosen (James et al. 2014, p. 214).

2.3 The information theoretic concept

Information coding is at the root of the Minimum Description Length (MDL) concept of simplicity, see (Montanez 2017, pp. 73-75; Domingos 1999, pp. 412-413). Under the MDL framework, probability distributions are equivalent to codes, so that the truth or the belief in the truth of a model is not relevant (de Rooij and Grünwald 2011, p. 895). MDL aims to minimize the sum of the number of bits that are required to represent the data plus those required to represent the model (Domingos 1999, p. 412; Pothos and Wolff 2006, p. 213; Grünwald 2007, p. 132). Hence MDL can be decomposed into a model fit term plus a model simplicity term (Montanez 2017, p. 73). The MDL principle is based on the observation that any regularities on the data enable efficient compression of such data. Models with fewer adjustable parameters require fewer bits to be coded. This means that more parsimonious models are preferred, provided that the models adequately capture the underlying patterns in the data so that the data are efficiently compressed.

Both MDL and BIC have been proved to behave suboptimally when the true model does not belong to the set of models to choose from (Grünwald 2007, p. 530), a situation that occurs very frequently in practice (Grünwald and Langford 2007, p. 139).

The main drawback of MDL is that, while models that are good at prediction implicitly compress the training data by identifying data patterns (Grünwald 2007, p. 595), there are models which compress the training data adequately but do not perform well at predicting new test data. In other words, implicit training data compression is necessary, but not sufficient, to attain good predictive performance. For the customer revisit example, this is a significant impediment to employ MDL since the true model does not belong to the set of models under comparison. This means that MDL could choose a model which summarizes the observed customer behavior data very efficiently, but predicts future revisits poorly.

2.4 The expressive power concept

Perhaps the most typical task in machine learning is classification. It consists of predicting a class label given an input vector which lies in some input space comprising several features. Classification problems greatly vary in their difficulty. The most difficult problems are those where nearby vectors in the input space have different class labels. This intricacy of the distribution of the labels must be matched by the classification models which aim to solve the problem. That is, a classification model must have enough expressive power to represent the intricate boundaries among the classes in the input space. This observation has led to some measures of model simplicity which are based on the richness of the kind of class boundaries that they can represent.

The best known of such measures is the Vapnik-Chervonenkis (VC) dimension. The VC dimension of a classification model is defined as the largest number of input vectors which can be shattered by the model, where the set of points is shattered if the model can learn a boundary which perfectly separates them no matter how the class labels are assigned to the points (Hastie et al. 2009, p. 238). This way, a researcher can evaluate the relative simplicity of two classification models by comparing their VC dimensions. Models with excessive VC dimension should be avoided because they might overfit (Vapnik 2000, pp. 297-298), i.e. they could focus on irrelevant features of the input dataset.

A key shortcoming of the VC dimension is that it does not assume any form of the distribution of the input vectors and their labels. This implies that it does not consider any kind of regularity in such distribution, while real problems do exhibit strong regularities. For example, e-commerce customer activity contains significant patterns, such as daily and seasonal activity trends, which are ignored by the VC dimension theory. Therefore the estimation by the VC dimension of the capability of a model to solve a classification problem is often too conservative, which greatly limits its applicability (Bishop 2006, pp. 344-345; Shalev-Shwartz and Ben-David 2014, p. 116).

Another quantitative measure of the flexibility of a classification model to learn an input distribution is Rademacher complexity. It measures to which extent, in average terms, a model can fit the random noise that may corrupt a given input distribution (Mohri et al. 2014, pp. 34-35). A higher Rademacher complexity indicates that the model is more rich and flexible so that it can accommodate more intricate input distributions. It is defined with respect to a certain input distribution, so it takes into

account the regularities of the input, unlike the VC dimension. However, Rademacher complexity is known to be hard to compute (Oneto et al. 2018, p. 4660). In particular, it is more difficult to bound or estimate than the VC dimension, and in some cases it is NP-hard, i.e. it requires an exponential number of calculations (Mohri et al. 2014, pp. 33, 38).

3 Models, learnable parameters and learning algorithms

In this section, I will argue that in machine learning the complexity of models is inextricably linked to the complexity of the algorithms that learn them. Software is made of algorithms plus data structures (Wirth 1985):

> Programs, after all, are concrete formulations of abstract algorithms based on particular representations and structures of data.

In prediction software, the data structures are the model. In this work I refer to the computational complexity of a specific algorithm, also called algorithm efficiency (Levitin 2011, p. 42), and not to the (more abstract and elusive) computational complexity of a problem, also called intrinsic complexity of a problem (Moore and Mertens 2011, p. 23). Computational efficiency is acknowledged as a criterion to justify the suitability of software (Kelly 2007, pp. 271, 273). The mere fact that the problem of learning a certain model can be solved, which is all that the simplicity concepts considered in Section 2 care about in terms of computation, says very little about the potential of the model to be applied in practice. There are many possible algorithms to learn a model, and for each learning algorithm, there is an infinity of possible programs that implement it, each with its own computational complexity. Therefore, in machine learning, the sole specification of a learnable model is not enough to assess simplicity (Kelly 2007, p. 273). Model evaluation, model selection, and algorithm selection are seen as different aspects of the same task (Raschka 2018, p. 1). The number of parameters of the model and the speed with which those parameters can be learned are seen as two equally important factors (Lin and Tegmark 2016, p. 2). The object whose simplicity should be assessed is the overall *learning system*, which is composed (at least) of:

1. The data acquisition process.
2. The hardware where the computations are performed.
3. The learnable model.
4. The learning algorithm (software) which is executed on the hardware, with the acquired data as input, and outputs the learned instantiation of the model.
5. The prediction procedure (software), executed on the hardware, which accepts the learned instantiation of the model and a query as inputs, and outputs a prediction.

Hardware is also important to assess the simplicity of a learning system, because models which are unfeasible to learn on a certain kind of hardware, become easy to use on other kinds of hardware. The differences in the computational complexity

among implementations of the same algorithm depending on hardware considerations can be enormous (Parashar et al. 2019, p. 305). This is what happens with deep learning neural networks, which are too complex to be learned on standard hardware, but affordable on specific graphics hardware. However, in this work, I will focus on the importance of learning algorithms to evaluate simplicity.

Let us consider an example of a learning system, namely a recommender system for our example e-commerce site. The data acquisition consists of recording successive purchase transactions by the same customer so that a register of the purchase history of many customers is collected. This data acquisition process requires dedicated software, databases, and computers. Some additional hardware and software are necessary to run the machine learning methods themselves. Next, a data scientist chooses a learnable model that might be adequate to predict, given the past behavior of a customer, which items she would be interested to buy in the future. After that, the learnable model is trained on the collected purchase histories by a suitable learning algorithm. Finally, the trained model is employed to execute a prediction procedure to show a visiting customer some items which she might like to purchase.

Inspired on the Inference to the Best Explanation (IBE) principle (Cabrera 2017, p. 1248), we can define the Refinement to the Best Prediction (RBP) scheme:

1. D is a set of data.
2. Software S_1 predicts D sufficiently well.
3. No competing software S_2, ..., S_N predicts D better than S_1.
4. One is justified in using S_1 to predict D.

Please note that truth does not have an explicit role here. It is not a logical inference process, but a software refinement one. Maybe the model which exists in S_1 is closest to the underlying real process which D comes from, but this is not a goal. In other words, machine learning allows that the models inside the S_i are black boxes with little or no resemblance to the physical process which generated the data D. Once the data are generated in step 1, the underlying physical process is irrelevant to the subsequent steps of the RBP scheme. In this context, the computational efficiency of the prediction software becomes a key factor to choose a learning system over another.

Learnable models contain a certain number of learnable parameters whose values must be estimated from the data. The first three classic approaches considered in Section 2 (Bayesian, frequentist and information-theoretic) evaluate the simplicity of the learned model with all their adjusted learnable parameters with respect to a space of possible models, but they completely ignore the computational procedure by which such learnable parameters have been determined. Furthermore, these three classic approaches are designed to measure the complexity of what are known as parametric models, as opposed to nonparametric models. On one hand, parametric models are based on the assumption that the observed data can be explained and summarized by a probability distribution of a specific mathematical form which is characterized by a relatively small number of learnable parameters. On the other hand, nonparametric models do not make assumptions about the mathematical form of the probability distribution which underlies the observed data nor they try to summarize the available dataset. This means that nonparametric models grow with the size of the dataset,

while they have very few learnable parameters, often just one learnable parameter. The neglection of nonparametric models causes the failure of classic simplicity concepts to account for nonparametric model selection (Rochefort-Maranda 2016, p. 270; de Rooij and Grünwald 2011, p. 892). Moreover, these three classic conceptions of simplicity are useless to explain how scientists choose among a range of parametric and nonparametric models. That is, parametric and nonparametric models are incommensurable under these classic simplicity models. The situation only gets worse for machine learning models which do not define any probability distribution. Notably enough, this class of models includes many of the most celebrated machine learning models.[2] In these cases, the AIC and the BIC cannot even be defined.

The fourth kind of classic simplicity measures, namely the expressive power ones (Section 2.4) also has important shortcomings. Like the previous ones, they do not consider the computational procedure that is followed to learn a model. The VC dimension completely ignores the strong regularities which often exist in real datasets D, which leads to inaccurate assessments of the capability of the models to learn such real datasets. Rademacher complexity is hard to compute, which hampers its practical applicability. Furthermore, both VC dimension and Rademacher complexity are defined for supervised learning problems where the desired output is provided by a supervisor, so that the simplicity of models for many fundamental tasks in machine learning such as unsupervised clustering, dimensionality reduction, density estimation, and reinforcement learning cannot be analyzed with them. In contrast to this, computational efficiency can be employed to assess the simplicity of all kinds of machine learning models.

4 Computational simplicity

As seen in Section 3, classic conceptions of simplicity are incomplete because they do not address the computational burden of learning the adjustable parameters of a model. Here we propose an alternative concept of simplicity which works for parametric, nonparametric, non probabilistic, unsupervised and reinforcement learning models. It is founded on a new version of Ockham's razor, that we may call the Big Data razor:

Definition 1 *Big Data razor.* Computations should not be multiplied beyond necessity.

Computational simplicity has already been recognized as one of several alternative concepts of simplicity (Rochefort-Maranda 2016, pp. 271-272). It has also been pointed out that, while favoring simplicity does not necessarily lead to the true model,

[2]The class of non probabilistic machine learning models includes: Support Vector Machines, k-means, decision trees, random forests, artificial neural networks (including deep learning neural networks) and many others. Most of them can be adapted to output probabilities, but they are usually employed without such adaptation, i.e. the model selection is carried out without any reference to probabilities.

it speeds up the search process (Kelly 2011, p. 1000), which highlights the importance of the computational load associated with model selection. Here we aim to argue that computational simplicity is a critical criterion for current machine learning.

Computational limitations have always played a role in machine learning. The novelty is that before Big Data, the scalability of algorithms as the number of samples grows was not so important because most datasets were small. This implies that the relative importance of computational complexity was moderate, as compared to classic measures of model complexity. Nowadays sample sizes N of 10^9 to 10^{11} are common (Cahsai et al. 2017, p. 1419, 1426; Hestness et al. 2017, pp. 5-10), and they are becoming even more frequent. For example, more than 10^9 items are shipped by the same e-commerce site per year (Carman 2018). This means that machine learning models whose computational requirements scale badly to those sample sizes are simply discarded, no matter how simple their structures may be, because only computationally cheaper models can harness the incoming data deluge.

An example of the Big Data razor is the success of two deep learning neural networks to detect objects in images: FasterRCNN (Ren et al. 2017) is more accurate than Yolo (Redmon and Farhadi 2018), but Yolo is faster (Dhiraj 2019, p. 118). FasterRCNN follows a two step approach (Ren et al. 2017, pp. 1138-1139). First, it generates a relatively large number of region proposals, i.e. rectangles (also called bounding boxes) which might enclose an object. After that, it analyzes the region proposals to estimate how likely they are to actually contain an object. In contrast to this, Yolo performs the object detection in a single step, since it directly generates regions which are very likely to enclose an object. Furthermore, Yolo comes in several versions that accept images of different resolutions. The higher the resolution, the more accurate the object detection, but the larger the network and its associated computational load. That is, Yolo versions are in a relation of computational load versus accuracy tradeoff (Redmon and Farhadi 2018, p. 4). Due to its two step architecture, FasterRCNN is slower than Yolo, but FasterRCNN is more accurate because its analysis of the input image is more detailed. Depending on the time requirements of the application, one of them is chosen: FasterRCNN or some of the Yolo versions. Even for the same network, some implementations are more or less accurate depending on the object detection accuracy that you wish to attain (Kang et al. 2018; Ma et al. 2018). In general terms, it can be said that the more accuracy, the more calculations are required (Huang et al. 2017, p. 7315). A linear relationship has been experimentally found between the accuracy and the number of images that can be processed per unit time (Canziani et al. 2016, p. 6). A similar tradeoff is found in classification. The more calculations, the more accuracy which can be attained (Kpotufe and Verma 2017, pp. 1, 16; Jose et al. 2013, pp. 1, 8).

The above examples illustrate the fact that in machine learning often predictive accuracy and computational simplicity are competing goals which stand in a relation of tradeoff, while classic concepts of simplicity (Bayesian, Frequentist, information theoretic and expressive power) are not considered.

There is a practical reason for this quest for computational simplicity. Computation has an economic cost in terms of hardware, software, and energy. In our example recommender system for an e-commerce site, there is a need for machine learning algorithms that can be executed with a small computational effort, so that cheaper

web servers are required and less electrical power is employed to supply them. All of this reduces the monetary cost of running the web site, i.e. the profits are increased. Therefore less computation means the cheaper application of machine learning to an ever growing range of tasks (Agrawal et al. 2018, ch. 3). In turn, this economic rationale directs researchers and practitioners towards computationally simple models. This justification of simplicity departs from the epistemic justifications of the classic concepts of simplicity. These classic justifications are problematic and suggest that parsimony is a surrogate goal (Sober 2015, p. 149).

Since predictive accuracy comes at a computational (and hence economic) cost, for the Big Data era the models which are obtained by a learning algorithm with a computational complexity which is higher than linear are not practical (Hong et al. 2019, p. 1; Burkov 2019, ch. 8). It is not only that the space of possible models is infinite (Korb 2004, p. 437), which was already known before the advent of Big Data. It is also that many of the possible models cannot be learned (adjusted) within practical computer resource limits.

The old (Bayesian, statistical) balance was between overfitting and underfitting, i.e. between the simplicity of the model and its predictive accuracy. While the old balance is still valid for parametric models, there is a new balance between the predictive accuracy of the learned model and the computational complexity of the associated learning algorithm, measured in terms of computation time and memory usage (Jiang et al. 2019, p. 201). The new balance is not restricted to parametric models since it applies to nonparametric and non probabilistic models too. The computational complexity of the test phase, i.e. when the learned model is employed to generate predictions, is also very important. This is because the test phase might be more computationally demanding than the training phase if the learned model is to be maintained for some time to yield many predictions. In other words, the relative importance of the computational complexity of the training and test phases depends on the expendability of the learned model.

The computational simplicity concept, considered in the context of the learning system depicted in Section 3, can help to explain why Big Data calls for models whose learning algorithms have linear complexity (or lower). To a first approximation, it can be assumed that the energy (and monetary cost) of acquiring N samples of data is directly proportional to N, i.e. it is linear with N. If the learning algorithm has linear complexity, this implies that the ratio between the energy required to learn the model and the energy required to acquire the data is a constant independent of N. However, if the learning algorithm has a complexity which is higher than linear, then the ratio tends to infinity as N grows. Hence the learning phase absorbs an unsustainable fraction of the energy devoted to run the entire learning system, as the number of available data samples N increases. For our example recommender system, this means that the energy employed to predict the future behavior of the customers grows much larger than the energy devoted to processing their actual purchase orders. This situation is characteristic of the Big Data era since the rate of growth of the amount of computation required to learn state of the art machine learning models has dramatically accelerated since 2012 (Amodei et al. 2019).

In the light of the above considerations, it can be inferred that a triple balance must be attained among three variables: the cost of acquiring the data, the cost of

learning the model, and the cost of using the model to make accurate predictions. The computational simplicity concept can provide a unified framework to understand the two last variables, while the first one (the data acquisition cost) depends on the scientific or technological field that the models are applied to.

Parsimony in terms of the number of adjustable parameters is not a goal in itself, although it affects the memory usage and indirectly the computation time. In other words, each adjustable parameter occupies some memory space, so models with more parameters have more memory usage. Besides, each adjustable parameter requires some computational effort to learn it, so adding more parameters to a given model usually results in higher computation time. Truth is not a goal either since it is assumed that none of the competing black box models reflects the actual structure of the problem. As mentioned before, the exact behavior of the customers of an e-commerce site is too complex to be captured by a machine learning model. There was a probabilistic turn (Sober 2015, p. 152) in the 20th century which meant that we did not assume anymore that nature is simple or that it is probably simple. The Big Data turn of the 21st century means that machine learning practitioners assume that models whose computational complexity is small enough to manage large and ever increasing volumes of data have the best chances to generate good predictions. Computationally simple algorithms which performed poorly with small amounts of data yield excellent results when supplied with large datasets, and it is believed that some problems can be essentially solved as soon as enough data is provided (Pereira et al. 2009, p. 9; Sun et al. 2017, p. 8). It is acknowledged that these simple models are not true. In other words, after the Big Data turn model selection is not driven by purely theoretical motivations, because the cost of running the overall learning system has become a fundamental factor to choose one model over another. It must be noted that before the advent of massive data processing by machine learning methods, the computational cost of adjusting the parameters of a model was not a pressing concern for machine learning practitioners. Nowadays, a standard strategy to cope with a given dataset is to try to learn a set of computationally cheap models, and then see which one yields the best predictive accuracy. The cheaper the models, the more tries that can be attempted for a fixed computational budget.

In this new context, the key questions are:

1. *How accurate can we get within our current computational limits?* Prediction is seen as a process of progressive refinement, which advances at the pace of the improvements in the computational resources. The discovery of the true model is not a concern because predictive accuracy is all that matters for most application fields of machine learning.
2. *Is there a limit to the accuracy increase of these simple models as the number of data samples N grows?* As seen above, the true model is not searched for, while it is not clear how close we can get to the truth by approximate models. This question can not be answered by theoretical reasoning. It can only be ascertained by experimentation on ever growing datasets provided by Big Data techniques. This strategy is driven by the observation that larger datasets lead to better results, although it is also observed that the marginal performance decreases as the size

> of the dataset increases (Sun et al. 2017, p. 850). There is a subjective perception that for a given model, there is a maximum possible accuracy which can not be surpassed no matter how big the training set is (Fernández-Delgado et al. 2014, pp. 3134-3135; Hestness et al. 2017, p. 11). For the e-commerce site example, this means that customer behavior is not fully predictable by any particular model, even if an unlimited amount of data is available. This leaves the question of whether radically new models could outperform the current best models to generate accurate predictions for a particular problem (Huang et al. 2017, p. 7315).

We may call this the *computational turn*. The state of affairs in machine learning is that researchers no longer aim to obtain the true model, but an approximate one which can be learned with a small computational load, while providing high accuracy.

The main reason behind this computational turn is the recent increase by several orders of magnitude of the number of available data samples N, which has dramatically emphasized the relevance of computational simplicity with respect to classic measures of model simplicity. Often machine learning algorithms whose computational complexity is higher than linear cannot be applied to very large datasets (Witten et al. 2017, p. 507), which did not happen before the Big Data deluge.

Let us consider the example of comparing a machine learning model *Quad* that is associated to a quadratic complexity learning algorithm[3] with a model *Lin* that is associated to a linear complexity learning algorithm.[4] The average execution times of both algorithms can be written as follows:

$$T_{Quad} = N^2 K_{Quad} F_{Quad} \tag{1}$$

$$T_{Lin} = N K_{Lin} F_{Lin} \tag{2}$$

where N is the number of training samples; F_{Quad} and F_{Lin} are the numbers of free parameters of *Quad* and *Lin*, respectively; and K_{Quad} and K_{Lin} are constants that depend on the models, the software implementations of the training algorithms, and the hardware where the algorithms are executed on. It must be pointed out that for parametric models, F_{Quad} and F_{Lin} must not depend on N, i.e. the size of the model must not depend on the number of available training samples (Russell and Norvig 2016, p. 737). Moreover, K_{Quad} and K_{Lin} must not depend on N either since we focus on a particular setup of software and hardware.

Now let us define the execution time ratio between *Quad* and *Lin*:

$$R = \frac{T_{Quad}}{T_{Lin}} = \frac{N^2 K_{Quad} F_{Quad}}{N K_{Lin} F_{Lin}} = \frac{N K_{Quad} F_{Quad}}{K_{Lin} F_{Lin}} \tag{3}$$

which is the number of times that *Lin* is faster than *Quad*.

[3]This means that the execution time of the learning algorithm is proportional to N^2. For example, kernelized Support Vector Machines (Bishop 2006, p. 349), and decision tree induction by the C4.5 algorithm with numeric attributes (Witten et al. 2017, pp. 219-220, 508).

[4]This means that the execution time of the learning algorithm is proportional to N. For example, Naive Bayes classifiers (Bishop 2006, p. 380), and logistic regression (Hastie et al. 2009, p. 120).

If the dataset has a moderate size N, then it makes sense to compare *Quad* and *Lin* by means of classic model simplicity measures, which focus on the numbers of free parameters F_{Quad} and F_{Lin}. For example, it could be the case that $F_{Quad} < F_{Lin}$ so that *Quad* is judged to be simpler than *Lin*, provided that the execution time ratio R is not too large. In other words, the classic simplicity measures and the execution time are complementary simplicity measures which can be collectively assessed by the machine learning practitioner. However, the advent of Big Data means that N grows by several orders of magnitude. Before Big Data, N was in the hundreds or thousands of samples, while nowadays N can be in the millions or billions. This implies that the execution time ratio R also grows by several orders of magnitude, as seen in Eq. 3. As N grows to infinity, the execution time ratio R also tends to infinity. Here is where the Big Data razor shaves off the *Quad* model since it does not matter how simple *Quad* might be in terms of the classic simplicity measures, because the difference in the execution times as measured by R is just too large. The situation becomes even more dramatic for learning algorithms whose execution time is proportional to N^3, i.e. cubic complexity.[5] That is, the relevance of model simplicity criteria that do not consider the execution time decays as N grows in the Big Data era.

5 The interplay among the classic and computational notions of simplicity

Here the relations among the classic and computational accounts of simplicity are explored. The classic ones have not been abandoned in current machine learning practices, although they play a secondary role:

Claim In the Big Data era, the classic notions of simplicity are surrogate goals of computational simplicity.

That is, the classic notions which deal with the structural simplicity of the models are considered by machine learning practitioners because they are indirect indicators of the computational load required to train the model, and not by their intrinsic value.

Next, the above claim is justified. The classic notions outlined in Section 2 are mainly devoted to assessing the simplicity of the machine learning models. As seen in Section 3, a machine learning model is a mathematical structure with one or more learnable parameters. The computational complexity of a learning algorithm is related to the structural complexity of the model that the algorithm is applied to. For example, we may take (1) and (2), and put the dependence with respect to the number of training samples N into a function $\mathcal{F}(N)$, so that the average execution time of a generic learning algorithm reads as follows:

$$T_{Generic} = \mathcal{F}(N)\, K_{Generic} F_{Generic} \tag{4}$$

[5]This means that the execution time of the learning algorithm is proportional to N^3. For example, kernel ridge regression (Witten et al. 2017, p. 508).

where $F_{Generic}$ is the number of free parameters of the model, and $K_{Generic}$ is a constant that depends on the model, the software implementation of the training algorithm, and the hardware where the algorithm is executed on. Equation 4 means that the average execution time is directly proportional to the number of free parameters, which is associated with the structural complexity of the model. Now, depending on the number of free parameters $F_{Generic}$ of the chosen model, the average execution time $T_{Generic}$ varies. Consequently, structural complexity is often correlated with computational complexity, so the former is an indirect indicator of the latter.

The preference for simple models of classic notions can be interpreted as a search strategy in the space of the possible models (de Rooij and Grünwald 2011, p. 893). However, a model must be instantiated in order to yield predictions, which are the ultimate goal of machine learning activity. Therefore, it is mandatory to learn the parameters of the model in order to instantiate it, prior to the extraction of predictions and their evaluation to measure predictive performance. This means that the relevant search space for machine learning is the space of possible model instantiations, which comprises all instantiations of all the considered models. In order to carry out a search in such space, classic notions of model complexity can play an auxiliary role. Classic notions can direct the search to models whose structure is smaller. But the key factor is the computational load of instantiating (training) the models, as seen in Section 4. This is managed by the computational notion of complexity that speeds up the process of instantiation of the models by choosing learning algorithms with a small computational complexity. In other words, computational simplicity governs the overall search process, while classic notions can help in the prioritization of some search directions over others in the space of all model instantiations.

This framework where classic and computational notions of simplicity work together is associated with scenarios where there are many possible models and many learning algorithms to choose from. The same algorithm can be employed to train several different models. This is the case of the Expectation Maximization (EM), which can be employed to train probabilistic mixtures of Gaussian (Bishop 2006, p. 435) or Bernoulli distributions (Bishop 2006, p. 444), or Bayesian linear regression models (Bishop 2006, p. 448). Conversely, a model can be trained by several different algorithms. For example, Bayesian networks can be learned by a wide range of algorithms (Acid et al. 2004, p. 219).

In order to avoid overfitting in current applications of machine learning to Big Data, computational techniques such as cross validation, bootstrap, regularization and early stopping are commonly employed (Hastie et al. 2009, p. 241, 253, 398; Russell and Norvig 2016, p. 708, 713; Witten et al. 2017, p. 162, 169, 393, 419, 431). This tendency is due to their lack of assumptions about the datasets, which widens their applicability.

6 Non epistemic justifications of computational simplicity

In this section, we investigate how the previously explained computational simplicity concept (Section 4) relates to the energy consumption and economic cost of the application of machine learning to real problems.

Some operations within the model learning stage of a learning system (Section 3) spend a disproportionately high amount of energy. This is the case of the optimization of the neural architecture for deep learning artificial neural networks since each step in this optimization often requires training of a full neural network. Sometimes the performance increment is small, so it must be evaluated whether it is worth the huge amounts of extra computation and the associated energy cost. The manager of our example e-commerce site may discover that it is not profitable to employ the most accurate customer behavior prediction model if the monetary value of the additional customer purchases is smaller than the extra cost of running the prediction software. Energy efficiency is acknowledged as a design criterion for the proposal of new deep neural network architectures, where a balance must be attained between energy consumption and predictive accuracy (Yang et al. 2017, pp. 1916, 1920; Li et al. 2016, p. 477), since the energy cost of deep learning is notoriously high (Ganguly et al. 2019, p. 335). This is directly connected to the computational simplicity criterion since energy consumption is to be measured in terms of the number of computational operations required to complete a learning task (Strubell et al. 2019). An optimization with at least two independent goals is established, where one of the goals is the minimization of the economic cost and the other goal is the maximization of the predictive accuracy.

In the particular case of deep learning neural networks, substantial energy savings can be attained by using the fine tuning technique. First, a neural network is trained with a standard set of training samples. This is a computationally heavy task since the entire network must be trained from scratch. Then the neural network is tuned to accomplish a specific task, which involves retraining a small portion of the network. This way, most of the initial training effort is reused, thereby drastically reducing the overall energy expenditure. This technique departs from classic machine learning approaches, which usually involved retraining the models for each new application. In the case of deep learning, the computational requirements are so huge that the classic naive retraining approach is simply not feasible for many organizations because they can not afford the costs. That is, typically our example e-commerce site software is based on some readily available pre-trained neural network, which is then tuned to learn the behavior of the customers of this specific web site so that the computational effort and the associated energy budget devoted to learning the predictive model are kept as low as possible.

There are reasons for the past neglection of computational simplicity and other energy consumption criteria. Machine learning is seen as a rapidly expanding branch of science with great potential to enable the automation of many tasks in the near future. Therefore, advanced machine learning systems are not subject to the criticisms that older technologies may have, because automation is recognized as a possible way to save human time and effort. While it is true that many tasks might be more efficiently executed by learning machines, optimal efficiency can only be attained by constraining the energy consumption of these machines. Otherwise, it would still be more energy efficient to employ humans for the task than replacing them with machines. In other words, machine learning systems should not be granted an unlimited amount of computational resources. In order to make informed economic

decisions about this matter, the benefits and costs of different strategies to implement machine learning methods must be elicited.

7 Conclusion

Classic concepts of simplicity aim to capture the parsimony of the machine learning model. The number of adjustable parameters is the most straightforward way to measure model complexity, and it is considered both by the Bayesian and Frequentist approaches. These two approaches differ mainly in their interpretation of the concept of probability. Bayesian model selection intends to choose the model that is estimated to have the highest probability of having generated the data, with the help of suitable choices for the prior probabilities, while the Frequentist field does not assign probabilities to models and focuses on estimating the predictive accuracy of the models. Minimum Description Length measures model complexity as the number of bits that are required to encode the model, under a suitable coding system. On their part, expressive power approaches measure the expressive power of a model, i.e. its ability to fit the intrinsic structure of the problem at hand. As seen, the four classic approaches to simplicity considered here only analyze the structure of the model, so that they ignore other aspects that are essential to employ machine learning models in practice. These additional aspects are covered by the learning system concept, which integrates all the relevant conceptual and physical structures required to successfully apply machine learning methods. The computational load necessary to train and test a model is a reliable, end to end measure of the effort which is devoted to accomplishing a machine learning task. Moreover, it is more encompassing than any of the classic concepts, since computational simplicity can directly be employed to compare parametric, non parametric, non probabilistic, unsupervised and reinforcement machine learning models.

Current practices in machine learning suggest that computational simplicity is what practitioners really seek to minimize since the simplicity of the structure of the model or its expressivity are not by themselves direct indicators of the real effort required to learn and apply the model. Therefore, structural simplicity can be regarded as a surrogate goal of computational simplicity. Computational resources are always limited, and this implies that a tradeoff must be found between computational effort and predictive accuracy. This tradeoff can be seen as an economic decision since computational load translates directly into energy and hardware costs. Choices about computational effort are associated with various energy consumption patterns. In most cases, there is a technological limit for the predictive accuracy that machine learning can attain at the current state of the art so that the tradeoff between effort and accuracy is bounded within the computational resource limits and the predictive accuracy limits. In other words, in many cases, it is not optimal to raise the energy consumption levels without limits until the maximum technologically possible predictive accuracy is achieved. Therefore, purely theoretical criteria about the structure of the models play a less important role for model selection in the Big Data era, since the size of the datasets has grown by several orders of magnitude.

These aspects of machine learning activity are no longer overlooked by scientists and practitioners.

Acknowledgments The author is grateful to David Teira (Universidad Nacional de Educación a Distancia, Madrid, Spain) and Emanuele Ratti (University of Notre Dame, Notre Dame, Indiana, USA) for their valuable comments.

References

Acid, S., de Campos, L.M., Fernández-Luna, J.M., Rodríguez, S., Rodríguez, J.M., Salcedo, J.L. (2004). A comparison of learning algorithms for Bayesian networks: a case study based on data from an emergency medical service. *Artificial Intelligence in Medicine*, *30*(3), 215–232.

Agrawal, A., Gans, J., Goldfarb, A. (2018). Prediction machines: the simple economics of artificial intelligence. Harvard Business Review Press.

Amodei, D., Hernandez, D., Sastry, G., Clark, J., Brockman, G., Sutskever, I. (2019). AI and compute. https://openai.com/blog/ai-and-compute/.

Bandyopadhyay, P.S., & Forster, M.R. (2011). Philosophy of statistics: an introduction. North Holland, pp 1–52.

Bishop, C.M. (2006). *Pattern recognition and machine learning*. Berlin: Springer.

Burkov, A. (2019). The hundred-page machine learning book. Andriy Burkov.

Cabrera, F. (2017). Can there be a Bayesian explanationism? On the prospects of a productive partnership. *Synthese*, *194*(4), 1245–1272.

Cahsai, A., Ntarmos, N., Anagnostopoulos, C., Triantafillou, P. (2017). Scaling k-nearest neighbours queries (the right way). In *2017 IEEE 37th international conference on distributed computing systems (ICDCS)* (pp. 1419–1430).

Canziani, A., Paszke, A., Culurciello, E. (2016). An analysis of deep neural network models for practical applications. CoRR arXiv:1605.07678.

Carman, A. (2018). Amazon shipped over 5 billion items worldwide through prime in 2017. https://www.theverge.com/2018/1/2/16841786/amazon-prime-2017-users-ship-five-billion.

Claeskens, G., & Hjort, N.L. (2008). *Model selection and model averaging*. Cambridge: Cambridge University Press.

Dawid, R. (2017). Bayesian perspectives on the discovery of the Higgs particle. *Synthese*, *194*(2), 377–394.

de Rooij, S., & Grünwald, P.D. (2011). Luckiness and regret in minimum description length inference. North Holland, pp 865–900.

Dhiraj, J.D.K. (2019). An evaluation of deep learning based object detection strategies for threat object detection in baggage security imagery. *Pattern Recognition Letters*, *120*, 112–119.

Domingos, P. (1999). The role of Occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery*, *3*(4), 409–425.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, *15*, 3133–3181.

Forster, M.R. (2001). *The new science of simplicity*, (pp. 83–119). Cambridge: Cambridge University Press.

Ganguly, A., Muralidhar, R., Singh, V. (2019). Towards energy efficient non-von Neumann architectures for deep learning. In *20th international symposium on quality electronic design (ISQED)* (pp. 335–342).

Grünwald, P.D. (2007). The minimum description length principle. The MIT Press.

Grünwald, P., & Langford, J. (2007). Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning*, *66*(2), 119–149.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The elements of statistical learning* Vol. 2. Berlin: Springer.

Henderson, L., Goodman, N., Tenenbaum, J., Woodward, J. (2010). The structure and dynamics of scientific theories: a hierarchical Bayesian perspective. *Philosophy of Science*, *77*(2), 172–200.

Hestness, J., Narang, S., Ardalani, N., Diamos, G.F., et al. (2017). Deep learning scaling is predictable, empirically. CoRR arXiv:1712.00409.

Hong, J., Wang, Z., Niu, W. (2019). A simple approximation algorithm for the diameter of a set of points in an Euclidean plane. *PLOS ONE*, *14*(2), 1–13.

Huang, Y., & Beck, J.L. (2018). Full Gibbs sampling procedure for Bayesian system identification incorporating sparse Bayesian learning with automatic relevance determination. *Computer-Aided Civil and Infrastructure Engineering*, *33*(9), 712–730.

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., Murphy, K. (2017). Speed/accuracy trade-offs for modern convolutional object detectors. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3296–3297).

James, G., Witten, D., Hastie, T., Tibshirani, R. (2014). *An introduction to statistical learning with applications in R*. Berlin: Springer.

Jiang, L., Zhang, L., Li, C., Wu, J. (2019). A correlation-based feature weighting filter for Naive Bayes. *IEEE Transactions on Knowledge and Data Engineering*, *31*(2), 201–213.

Jose, C., Goyal, P., Aggrwal, P., Varma, M. (2013). Local deep kernel learning for efficient non-linear SVM prediction. In Dasgupta, S., & McAllester, D. (Eds.) *Proceedings of the 30th international conference on machine learning, PMLR, Atlanta, Georgia, USA, Proceedings of Machine Learning Research*, (Vol. 28 pp. 486–494).

Kang, D., Kang, D., Kang, J., Yoo, S., Ha, S. (2018). Joint optimization of speed, accuracy, and energy for embedded image recognition systems. In *Proceedings of the 2018 design, automation and test in Europe conference and exhibition*. DATE 2018, vol 2018-January, pp. 715–720.

Kelly, K.T. (2007). Ockham's razor, empirical complexity, and truth-finding efficiency. *Theoretical Computer Science*, *383*(2), 270–289.

Kelly, K.T. (2011). Simplicity, truth and probability. North Holland, pp 983–1026.

Korb, K.B. (2004). Introduction: machine learning as philosophy of science. *Minds and Machines*, *14*(4), 433–440.

Kpotufe, S., & Verma, N. (2017). Time-accuracy tradeoffs in kernel prediction: controlling prediction quality. *Journal of Machine Learning Research*, *18*(44), 1–29.

Levitin, A. (2011). In 3 (Ed.) *The design and analysis of algorithms*. London: Pearson Education.

Li, D., Chen, X., Becchi, M., Zong, Z. (2016). Evaluating the energy efficiency of deep convolutional neural networks on CPUs and GPUs. In *2016 IEEE international conferences on big data and cloud computing (BDCloud), social computing and networking (SocialCom), sustainable computing and communications (SustainCom) (BDCloud-SocialCom-SustainCom),* pp. 477–484.

Lin, H.W., & Tegmark, M. (2016). Why does deep and cheap learning work so well? arXiv:1608.08225.

Ma, J., Chen, L., Gao, Z. (2018). Hardware implementation and optimization of tiny-YOLO network. *Communications in Computer and Information Science*, *815*, 224–234.

Mohri, M., Rostamizadeh, A., Talwalkar, A. (2014). *Foundations of machine learning*. Cambridge: The MIT Press.

Montanez, G.D. (2017). Why machine learning works. https://www.cs.cmu.edu/gmontane/montanez_dissertation.pdf.

Moore, C., & Mertens, S. (2011). *The nature of computation*. Oxford: Oxford University Press.

Murphy, K.P. (2012). *Machine learning: a probabilistic perspective*. Cambridge: The MIT Press.

Oneto, L., Navarin, N., Donini, M., Ridella, S., Sperduti, A., Aiolli, F., Anguita, D. (2018). Learning with kernels: a local Rademacher complexity-based analysis with application to graph kernels. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(10), 4660–4671.

Parashar, A., Raina, P., Shao, Y.S., Chen, Y., Ying, V.A., Mukkara, A., Venkatesan, R., Khailany, B., Keckler, S.W., Emer, J. (2019). Timeloop: a systematic approach to DNN accelerator evaluation. In *2019 IEEE international symposium on performance analysis of systems and software (ISPASS)* (pp. 304–315).

Pereira, F., Norvig, P., Halevy, A. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, *24*(2), 8–12.

Pothos, E.M., & Wolff, J.G. (2006). The simplicity and power model for inductive inference. *Artificial Intelligence Review*, *26*(3), 211–225.

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. CoRR arXiv:1811.12808.

Redmon, J., & Farhadi, A. (2018). YOLOv3: an incremental improvement. CoRR arXiv:1804.02767.

Ren, S., He, K., Girshick, R., Sun, J. (2017). Faster r-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149.

Rochefort-Maranda, G. (2016). Simplicity and model selection. *European Journal for Philosophy of Science* (6): 261–279.

Russell, S.J., & Norvig, P. (2016). Artificial intelligence: a modern approach. Pearson Education Limited, Harlow, Essex, England.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: from theory to algorithms*. Cambridge: Cambridge University Press.

Sober, E. (2015). *Ockham's razor: a user manual*. Cambridge: Cambridge University Press.

Strubell, E., Ganesh, A., McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th annual meeting of the association for computational linguistics*.

Sun, C., Shrivastava, A., Singh, S., Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *The IEEE international conference on computer vision (ICCV)* (pp. 843–852).

Vapnik, V.N. (2000). *The nature of statistical learning theory*. New York: Springer.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*(1), 92–107.

Wirth, N. (1985). In 2 (Ed.) *Algorithms + data structures = programs*. Englewood Cliffs: Prentice-Hall.

Witten, I.H., Frank, E., Hall, M.A., Pal, C.J. (2017). *Data mining: practical machine learning tools and techniques*, 4th edn. Cambridge: Morgan Kaufmann.

Yang, T., Chen, Y., Emer, J., Sze, V. (2017). A method to estimate the energy consumption of deep neural networks. In *2017 51st asilomar conference on signals, systems, and computers* (pp. 1916–1920).

2 Conclusions

The debate about the role of machine learning in scientific research has mainly been sterilized by a lack of understanding among philosophers of science, data scientists, and the scientists who employ data science techniques. Many data scientists focus on the novelty of the possibilities that the data revolution has brought, while they overlook the fact that the scientific process can not be fully automated, no matter how much effort is put on the computational branch of scientific disciplines. This attitude raises all kinds of suspicions from the philosophical point of view, so that data scientists are often seen as naive computer enthusiasts who do not understand the principles of the scientific method. On their part, scientists of the other disciplines struggle to get their job done with the help of algorithmic models while they try to justify that their practices still stick to the scientific method.

One of the most controversial claims of data scientists is that machine learning algorithms generate theory free models. This apparently contradicts the conventional wisdom that experiments and the data which comes from them are theory laden. In my opinion, this puzzle can be solved by considering two parts within scientific theories, a distinction which has been done earlier by Wolfgang Pietsch, Emanuele Ratti, and Laura Franklin. The conceptual part is a prerequisite for any data collection, no matter if experimental or observational, so that the mechanistic part is built on top of the conceptual part. For the mechanistic part, automatically learned models are alternatives to traditional ones, and the choice should be guided by the characteristics of the problem to be solved.

I have proposed a new kind of model complexity that is suited for machine learning, namely computational complexity, as detailed in (López-Rubio, 2020). The amount of computations that are required to train a model is what really matters to present day machine learning practitioners since their goal is to attain the best possible predictive performance at the lowest possible computational cost. Older simplicity measures based on the structure of the model are less popular since they only give indirect indications of predictive performance.

In this thesis, I have argued that the role of machine learning critically depends on the complexity of the phenomenon at hand, as seen in (López-Rubio & Ratti, 2019). The kind of complexity that is relevant here is the difficulty of producing an understandable description in the mathematical language of the mechanism which relates the variables so that the description generalizes to other situations and can be used to infer testable predictions. The bias-variance tradeoff means that for complex underlying relations among variables, explanation and prediction become two inde-

pendent goals so that the best models according to one of them are suboptimal for the other. This draws a distinction between the hard sciences, where understandable descriptions are common, and the soft sciences, where they are scarce or inexistent. The epistemic status of data science techniques essentially varies depending on this. If those descriptions are not likely to be found, then explanation and prediction must be considered as two separate dimensions of science. In this case, machine learning models with appropriate predictive performance fulfill a role that hypothesis based ones can not. On the other hand, the existence of suitable mathematical accounts of the causal structure of a phenomenon reduces the task of data science to find interesting patterns that must be explained. Under these conditions, explanation and prediction can be achieved at the same time. Nevertheless, even for hard sciences, data mining techniques are fundamentally different from other tools. Scientists are no longer able to study the data directly, but they are forced by their volume to see them through the lens of a particular set of machine learning algorithms. This choice profoundly influences what can be seen, so there is a fundamental interplay between data science models and the construction of models in the other disciplines.

The clash between the hypothesis-driven and the data-driven approaches to research is most evident in life sciences. According to my complexity argument, this is because biological systems lie between physical systems, at one extreme of the complexity spectrum, and social systems, at the other extreme. This means that algorithmic models are more useful for those biological systems which can not be adequately studied by splitting them into simpler independent subsystems. Under these circumstances, mechanistic explanations can only be obtained for small subsystems. Predictive machine learning models are better suited to cope with the complexity of the overall systems, although they are not amenable to extract explanations from them.

Long term predictions are risky in the social sciences, so it is hard to predict the future importance of machine learning for scientists from other disciplines. However, the above arguments suggest that, as science advances towards the study of more complex systems based on an ever increasing volume of data, clear cut explanations will be scarcer, and approximate prediction will become more acceptable.

The revolution of deep learning has completely changed the ways that machine learning is applied to computer vision and natural language processing in just a few years. I have investigated the possible similarities among deep learning artificial neural networks and the human brain in (López-Rubio, 2018). Evidence collected so far suggests that versions of computational functionalism that are based on symbolic approaches to computation are outdated. In view of this, I have proposed neural computational functionalism, which establishes deep learning artificial neural networks as the computational reference for the internal operation of the human brain. Future research in neuroscience will determine to what extent this association is valid.

3 Publication report

The publications which support this thesis are the following:

1. The article *Computational Functionalism for the Deep Learning Era* (López-Rubio, 2018), by Ezequiel López-Rubio, published in 2018 in the international journal *Minds and Machines*, ISSN 0924-6495. The journal is in the first quartile (Q1) of the *Philosophy* category of Scopus for 2018, with a CiteScore of 1.24. It is also in the first quartile (Q1) of the *Philosophy* category of SCImago for 2018, with impact factor 0.474. Finally, it is in the third quartile (Q3) of the *Computer Science, Artificial Intelligence* category of the Journal Citation Reports, Science Citation Index (JCR-SCI), for 2018, with impact factor 1.400.

2. The article *Data science and molecular biology: prediction and mechanistic explanation* (López-Rubio & Ratti, 2019), by Ezequiel López-Rubio and Emanuele Ratti, published in 2019 in the international journal *Synthese*, ISSN 0039-7857. The journal is in the first quartile (Q1) of the *Philosophy* category of Scopus for 2018, with a CiteScore of 1.17. It is also in the first quartile (Q1) of the *Philosophy* category of SCImago for 2018, with impact factor 0.757. Finally, it is in the second quartile (Q2) of the *History & Philosophy of Science* category of the Journal Citation Reports, Social Science Citation Index (JCR-SSCI), for 2018, with impact factor 1.262.

3. The article *The Big Data razor* (López-Rubio, 2020), by Ezequiel López-Rubio, published in 2020 in the international journal *European Journal for Philosophy of Science*, ISSN 1879-4912. The journal is in the first quartile (Q1) of the *History and Philosophy of Science* category of Scopus for 2018, with a CiteScore of 1.01. It is also in the first quartile (Q1) of the *History and Philosophy of Science* category of SCImago for 2018, with impact factor 0.458. Finally, it is in the second quartile (Q2) of the *History & Philosophy of Science* category of the Journal Citation Reports, Science Citation Index (JCR-SCI), for 2018, with impact factor 0.627.

References

Agarwal, R., & Dhar, V. (2014). Big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research*, *25*(3), 443–448.

Anderson, D. R., Burnham, K. P., Gould, W. R., & Cherry, S. (2001). Concerns about finding effects that are actually spurious. *Wildlife Society Bulletin*, *29*(1), 311-316.

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management*, *64*(4), 912-923.

Anderson, J., Bothell, D., Byrne, M., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036-1060. Retrieved from http://dx.doi.org/10.1037/0033-295X.111.4.1036

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Braun, E., & Marom, S. (2015). Universality, complexity and the praxis of biology: Two case studies. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *53*, 68-72.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199-231.

Buchanan, B. G. (1972). *Review of hubert Dreyfus' What computers can't do: a critique of artificial reason* (Tech. Rep. No. STAN-CS-72-325). Stanford University. (http://i.stanford.edu/pub/cstr/reports/cs/tr/72/325/CS-TR-72-325.pdf)

Burian, R. (2007). On microRNA and the need for exploratory experimentation in post-genomic molecular biology. *History and Philosophy of the Life Sciences*, *29*(3), 285-312.

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

Cass, S. (2011). Unthinking machines. *MIT Technology Review*. Retrieved from http://www.technologyreview.com/news/423917/unthinking-machines/

Chiang, R., Goes, P., & Stohr, E. (2012). Business intelligence and analytics education, and program development: A unique opportunity for the information

systems discipline. *ACM Transactions on Management Information Systems*, *3*(3), 1-13.

Cleveland, W. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, *69*(1), 21–26.

Cox, D. R. (2001). Comment to 'Statistical modeling: The two cultures'. *Statistical Science*, *16*(3), 216-218.

Cristianini, N. (2010). Are we there yet? *Neural Networks*, *23*(4), 466-470. Retrieved from http://dx.doi.org/10.1016/j.neunet.2010.01.006

Dalton, L. A., & Dougherty, E. R. (2013a). Optimal classifiers with minimum expected error within a Bayesian framework - Part I: Discrete and Gaussian models. *Pattern Recognition*, *46*(5), 1301 - 1314. doi: http://dx.doi.org/10.1016/j.patcog.2012.10.018

Dalton, L. A., & Dougherty, E. R. (2013b). Optimal classifiers with minimum expected error within a Bayesian framework - Part II: Properties and performance analysis. *Pattern Recognition*, *46*(5), 1288 - 1300. doi: http://dx.doi.org/10.1016/j.patcog.2012.10.019

Dalton, L. A., & Yousefi, M. R. (2015). On optimal Bayesian classification and risk estimation under multiple classes. *EURASIP Journal on Bioinformatics and Systems Biology*, *2015*(1), 1–23.

Davenport, D. (2013). The two (computational) faces of AI. In V. C. Müller (Ed.), *Philosophy and theory of artificial intelligence* (Vol. 5, p. 43-58). Springer Berlin Heidelberg. (http://dx.doi.org/10.1007/978-3-642-31674-6_4)

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, *56*(12), 64–73.

Dolinski, K., & Troyanskaya, O. (2015). Implications of big data for cell biology. *Molecular Biology of the Cell*, *26*(14), 2575-2578.

Dreyfus, H. (1972). *What computers can't do.* New York: MIT Press.

Dreyfus, H. (1992). *What computers still can't do.* New York: MIT Press.

Efron, B. (2001). Comment to 'Statistical modeling: The two cultures'. *Statistical Science*, *16*(3), 218-219.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37-54.

Franklin, L. R. (2005). Exploratory experiments. *Philosophy of Science*, *72*(5), 888-899.

Frické, M. (2015). Big data and its epistemology. *Journal of the Association for Information Science and Technology*, *66*(4), 651-661.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1-58.

Goertzel, B. (2012). Perception processing for general intelligence: Bridging the symbolic/subsymbolic gap. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7716 LNAI*, 79-88. Retrieved from http://dx.doi.org/10.1007/978-3-642-35506-6_9

Golub, T. (2010). Counterpoint: Data first. *Nature*, *464*, 679.

Green, S. (2015). Can biological complexity be reverse engineered? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *53*, 73-83.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). New York: Springer.

Haufe, C. (2013). Why do funding agencies favor hypothesis testing? *Studies in History and Philosophy of Science Part A*, *44*(3), 363-374.

Hawking, S., Russell, S., Tegmark, M., & Wilczek, F. (2014, May). Transcendence looks at the implications of artificial intelligence - but are we taking AI seriously enough? *The Independent*. Retrieved from http://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence--but-are-we-taking-ai-seriously-enough-9313474.html

Hedges, L. (1987). How hard is hard science, how soft is soft science?: The empirical cumulativeness of research. *American Psychologist*, *42*(5), 443-455.

Hong, F. T. (2013). The role of pattern recognition in creative problem solving: A case study in search of new mathematics for biology. *Progress in Biophysics and Molecular Biology*, *113*(1), 181 - 215.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*(2), 251-257.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359-366.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (G. Casella, S. Fienberg, & I. Olkin, Eds.). New York: Springer.

Jordan, M., & Mitchell, T. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Kaelbling, L., Littman, M., & Moore, A. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237-285. Retrieved from http://dx.doi.org/10.1613/jair.301

Karaca, K. (2013). The strong and weak senses of theory-ladenness of experimentation: Theory-driven versus exploratory experiments in the history of high-energy particle physics. *Science in Context*, *26*(1), 93-136.

Kosko, B. (1994). Fuzzy systems as universal approximators. *IEEE Transactions on Computers*, *43*(11), 1329-1333.

Krohs, U. (2012). Convenience experimentation. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 52 - 57.

Krohs, U. (2015). Can functionality in evolving networks be explained reductively? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *53*, 94-101.

Leonelli, S. (2012). Introduction: Making sense of data-driven research in the biological and biomedical sciences. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical*

Sciences, *43*(1), 1 - 3.

Leonelli, S. (2015). What counts as scientific data? a relational framework. *Philosophy of Science*, *82*(5), 810-821.

Levin, N. (2014). Multivariate statistics and the enactment of metabolic complexity. *Social Studies of Science*, *44*(4), 555-578.

López-Rubio, E. (2018). Computational functionalism for the deep learning era. *Minds and Machines*, *28*, 667-688. Retrieved from https://doi.org/10.1007/s11023-018-9480-7

López-Rubio, E. (2020). The Big Data razor. *European Journal for Philosophy of Science*, *10*, 22. Retrieved from https://doi.org/10.1007/s13194-020-00288-8

López-Rubio, E., & Ratti, E. (2019). Data science and molecular biology: prediction and mechanistic explanation. *Synthese*. Retrieved from https://doi.org/10.1007/s11229-019-02271-0

Mryglod, O., Kenna, R., Holovatch, Y., & Berche, B. (2013). Comparison of a citation-based indicator and peer review for absolute and specific measures of research-group excellence. *Scientometrics*, *97*(3), 767-777.

Müller-Wille, S., & Charmantier, I. (2012). Natural history and information overload: The case of linnaeus. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 4 - 15.

National Science Board. (2005, September). *Long-lived digital data collections: Enabling research and education in the 21st century* (Tech. Rep.). National Science Foundation. Retrieved from http://www.nsf.gov/pubs/2005/nsb0540/

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, *19*(3), 113-126. Retrieved from http://doi.acm.org/10.1145/360018.360022

Nilsson, N. J. (2010). *The quest for artificial intelligence*. Cambridge University Press. (http://ai.stanford.edu/~nilsson/QAI/qai.pdf)

Norvig, P. (2012). *On Chomsky and the two cultures of statistical learning*. http://norvig.com/chomsky.html.

Peters, D., Havstad, K., Cushing, J., Tweedie, C., Fuentes, O., & Villanueva-Rosales, N. (2014). Harnessing the power of big data: Infusing the scientific method with machine learning to transform ecology. *Ecosphere*, *5*(6), 1-15.

Pietsch, W. (2015). Aspects of theory-ladenness in data-intensive science. *Philosophy of Science*, *82*(5), 905-916.

Press, G. (2013, May 28). A very short history of data science. *Forbes*. (http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/)

Ratti, E. (2015). Big data biology: Between eliminative inferences and exploratory experiments. *Philosophy of Science*, *82*(2), 198-218.

Russell, S., & Norvig, P. (2009). *Artificial intelligence: a modern approach* (3rd ed.). Pearson. Retrieved from http://aima.cs.berkeley.edu/

Schapire, R. (1990). The strength of weak learnability. *Machine Learning*, *5*(2), 197-227. Retrieved from http://dx.doi.org/10.1007/BF00116037

Schickore, J. (2016). "Exploratory experimentation" as a probe into the relation between historiography and philosophy of science. *Studies in History and Philosophy of Science Part A*, *55*, 20-26.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417-457. Retrieved from http://dx.doi.org/10.1017/S0140525X00005756

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.

Shmueli, G., & Koppius, O. (2011). Predictive analytics in information systems research. *MIS Quarterly: Management Information Systems*, *35*(3), 553-572.

Smith, L., Best, L., Stubbs, D., Johnston, J., & Archibald, A. (2000). Scientific graphs and the hierarchy of the sciences: A Latourian survey of inscription practices. *Social Studies of Science*, *30*(1), 73-94.

Smolensky, P. (1988, 3). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*, 1–23. Retrieved from http://dx.doi.org/10.1017/S0140525X00052432

Stephens, Z., Lee, S., Faghri, F., Campbell, R., Zhai, C., Efron, M., ... Robinson, G. (2015). Big data: Astronomical or genomical? *PLoS Biology*, *13*(7), e1002195.

Strasser, B. (2012). Data-driven sciences: From wonder cabinets to electronic databases. *Studies in history and philosophy of biological and biomedical sciences*, *43*(1), 85-87.

Sundararajan, A., Provost, F., Oestreicher-Singer, G., & Aral, S. (2013). Information in digital, economic, and social networks. *Information Systems Research*, *24*(4), 883–905.

Thompson, S., & Higgins, J. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*, *21*(11), 1559-1573.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, Massachussetts: Addison-Wesley.

van Eck, N., Waltman, L., den Berg, J., & Kaymak, U. (2006). Visualizing the computational intelligence field. *IEEE Computational Intelligence Magazine*, *1*(4), 6-10. Retrieved from http://dx.doi.org/10.1109/MCI.2006.329702

Van Horn, J., & Toga, A. (2014). Human neuroimaging as a "big data" science. *Brain Imaging and Behavior*, *8*(2), 323–331.

Waller, M., & Fawcett, S. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics*, *34*(2), 77–84.

Wang, L.-X. (1992). Fuzzy systems are universal approximators. In *Ieee international conference on fuzzy systems* (p. 1163-1170).

Weinberg, R. (2010). Point: Hypotheses first. *Nature*, *464*, 678.

White, P. T. (1970). Behold the computer revolution. *National Geographic*, *138*(5).

Xu, R., & Wunsch II, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645-678. Retrieved from http://dx.doi.org/

10.1109/TNN.2005.845141

Zhu, Y., & Xiong, Y. (2015). Defining data science. *CoRR*, *abs/1501.05039*. Retrieved from http://arxiv.org/abs/1501.05039