The importance of expert knowledge in big data and machine learning

Hansen, Jens Ulrik; Quinon, Paula

Published in:
Synthese

DOI:
10.1007/s11229-023-04041-5

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Hansen, J. U., & Quinon, P. (2023). The importance of expert knowledge in big data and machine learning. *Synthese*, *201*(35), [35]. https://doi.org/10.1007/s11229-023-04041-5

Synthese (2023) 201:35
https://doi.org/10.1007/s11229-023-04041-5

ORIGINAL RESEARCH

The importance of expert knowledge in big data and machine learning

Jens Ulrik Hansen[1] · Paula Quinon[2]

Received: 8 August 2022 / Accepted: 1 January 2023

Abstract
According to popular belief, big data and machine learning provide a wholly novel approach to science that has the potential to revolutionise scientific progress and will ultimately lead to the 'end of theory'. Proponents of this view argue that advanced algorithms are able to mine vast amounts of data relating to a given problem without any prior knowledge and that we do not need to concern ourselves with causality, as correlation is sufficient for handling complex issues. Consequently, the human contribution to scientific progress is deemed to be non-essential and replaceable. We, however, following the position most commonly represented in the philosophy of science, argue that the need for human expertise remains. Based on an analysis of big data and machine learning methods in two case studies—skin cancer detection and protein folding—we show that expert knowledge is essential and inherent in the application of these methods. Drawing on this analysis, we establish a classification of the different kinds of expert knowledge that are involved in the application of big data and machine learning in scientific contexts. We address the ramifications of a human-driven expert knowledge approach to big data and machine learning for scientific practice and the discussion about the role of theory. Finally, we show that the ways in which big data and machine learning both influence and are influenced by scientific methodology involve continuous conceptual shifts rather than a rigid paradigm change.

Keywords Big data · Machine learning · Expert knowledge · Agnostic sciences · Inductive method · Role of theory · Paradigm shift

The authors are listed in alphabetical order.

✉ Jens Ulrik Hansen
jensuh@ruc.dk

✉ Paula Quinon
paula.quinon@pw.edu.pl

[1] Department of People and Technology, Roskilde University, Roskilde, Denmark

[2] Faculty of Administration and Social Sciences, Warsaw University of Technology, Warsaw, Poland

Published online: 20 January 2023

1 Introduction

According to popular belief, big data and machine learning provide a wholly novel approach to science that could potentially revolutionise scientific progress (Hey et al., 2009a; Kitchin, 2014). A radical expression of this belief can be found in Chris Anderson's, 2008 claim that big data and machine learning in science will lead to the 'end of theory', which was also the title of his famous editorial in WIRED Magazine where he first put forward this proposal (Anderson, 2008). According to Anderson, advanced big data and machine learning algorithms enable us to mine vast amounts of data relating to a given task without any prior knowledge of that task; as such, we do not need to possess a causal understanding of the issues at stake, because correlation is all that is required for the scientific progress to happen. Anderson argues that once you have access to big data, there is no need to 'settle for models'. He notes that in the 'Petabyte Era', "information is not a matter of simple three- and four-dimensional taxonomy and order, but dimensionally agnostic statistics" (Anderson, 2008). The article is full of buzzwords; in addition to his references to the 'Petabyte Era' and an 'agnostic' method, Anderson prophesies what he calls the 'end of theory', i.e., the demise of classical scientific explanation 'built around testable hypotheses', in which 'coherent models' and 'unified theories', confirmed by experiments, serve as the basis for the scientific understanding of facts and the prediction of the future. Anderson's claims have become so influential that it is almost impossible to find a scholarly analysis of big data and machine learning published in the past 15 years that does not cite him.[1]

Another highly influential voice in the debate around the epistemological status of big data and machine learning has been James (Jim) Grey, a computer scientist and entrepreneur respected in both science and business. Grey claims that a 'Fourth Paradigm' for science's role as an academic discipline has been established through the development of 'eScience', a term that Grey uses to denote the application of data science methods to scientific data collected by instruments or generated by simulations. The results of such applications are not necessarily processed by humans to be turned into knowledge, but are instead stored on computers as additional data. Grey has expressed the hope that future computer scientists and information technology developers will design and build 'generic tools for scientists' in the form of 'cheap "data bricks"'. These 'data bricks' could be used by both simulation and data analysis tools, which would then be applicable to various sets of data regardless of their potential or known significance (Hey et al., 2009b, p. xx).[2]

[1] We can say without much hesitation that every author cited in our bibliography who is involved in the discussion of the new method mentions Anderson. The same can be said about Grey, whose views are presented in the next paragraph.

[2] Grey defines the first three paradigms in order as experimental science (which describes natural phenomena), theoretical science (which uses models and generalisations), and the use of computer simulations. The simulations used in this third paradigm, Grey explains, "are generating a whole lot of data, along with a huge increase in data from the experimental sciences"; consequently, a huge amount of data must be stored and then processed on computers before scientists can make any use of it. Data-intensive sciences and their associated 'techniques and technologies' form Grey's 'Fourth Paradigm'—an expression that has gained popularity and is widely used, although it does not fit the traditional Kuhnian sense. The term is discussed further in Sect. 6 of this paper.

We take Anderson and Grey's views as prime examples of views that discard any need for expert knowledge in applying big data and machine learning methods to science. These views raise two key questions: where did these ideas originate, and to what extent, if at all, are they justified? To address the first question, we must consider the hype surrounding big data and machine learning in industry. Conversations about new innovations and disruptions (which are commonplace in industry) easily transform into conversations about paradigm shifts in science; such exaggerated claims should be carefully monitored and frequently questioned. Furthermore, the current application of big data and machine learning methods to many areas of industry, as well as the significant amount of research that large tech companies have conducted on big data and machine learning, have resulted in a growing pressure to bridge the gap between business and science, which in turn makes it difficult to distinguish valuable insights and ideas from superficial chatter. Thus, it is important to clearly trace the flow of knowledge between business and science. In this paper, however, we shall focus solely on the second question, examining the rationale for these views in relation to the current backdrop of big data and machine learning.

It should first be noted that philosophers of science reject the radical claim that big data and machine learning provide a completely new and human-independent approach to answering scientific questions. However, the indispensability of human contribution is often tacitly assumed. Researchers on this topic focus instead on explaining big data and machine learning methods in relation to other well-established and widely studied approaches. For instance, Wolfgang Pietsch (2021) has recently examined the extent to which big data and machine learning represent a resurgence of inductive methods. Other investigators have considered the extent to which it is possible to gain knowledge of a phenomenon without first understanding it; Domenico Napoletani et al. (2011, 2014, 2021), for example, have formulated the concept of 'agnostic science'. While it may sound as if the views could be in line with Anderson and Grey, this is not the case. On the contrary, Napoletani et al.'s notion of 'agnostic science' differs substantially from that of Anderson and Pietsch is explicit about the need for a theoretical background to proper application of big data and machine learning methods (we will elaborate on this in Sect. 2).

Thus, in this paper, we will not argue against Pietsch and Napoletani et al., but against views that discard the need for expert knowledge in the application of big data and machine learning methods in science. In other words, we are adding to the 'informed' debate a direct focus on the question of whether methods using big data and machine learning eliminate the need for domain-specific expert knowledge altogether, and could thus lead to some form of automated science. Our approach originates from an analysis of two cases of scientific practice that use big data and machine learning, which led us to develop a taxonomy of expert knowledge involved in such types of research. Thus, our approach aligns well with other studies in the philosophy of science that focus on analysing scientific practice (Ankeny et al., 2011; Leonelli, 2016; Northcott, 2020); Leonelli (2016) and Northcott (2020) in particular seem to anticipate our claim that expert knowledge is significantly involved in research driven by big data and machine learning. The main contribution of this paper is an analysis of case studies that support this claim, as well as a taxonomy that explicates this dependence on expert knowledge.

We begin by discussing what we refer to as the ‘informed’ (or ‘critical’) view in contrast to Anderson’s popular view (Sect. 2) and explain what we mean by big data and machine learning methods (Sect. 3). We then elaborate on two paradigmatic uses of big data and machine learning methods in science: skin cancer detection and protein folding (Sect. 4). On this basis, we develop a taxonomy of expert knowledge that can be applied to big data and machine learning-driven research (Sect. 5). Finally, we argue that this taxonomy provides a fresh perspective on several debates surrounding the role of big data and machine learning in science: whether they concern inductive methods, the removal of the need for theory, or the constitution of a new scientific paradigm (Sect. 6).

2 The informed view of big data and machine learning in science

In a series of papers published between 2011 and 2021, Napoletani et al. put forward the thesis that there is a particular kind of methodology that differs from those underlying classical scientific methods[3] and makes it possible to find significant correlations across huge datasets. According to these authors, certain scientific methods, when applied to phenomena that are not even tentatively understood, can represent instances of what they call ‘agnostic science’. One paradigmatic example of such a method that *might be* agnostic in this sense, is machine learning. A comparison of two usages of machine learning, the PageRank algorithm and the microarray method, illustrates how it can be applied to both, phenomena that are understood and those that are not. Napoletani and his co-authors argue that the PageRank algorithm, which provides a hierarchical classification of websites, represents a way of applying a machine learning algorithm to a well-understood problem that possesses a foreseeable solution-structure; because the result of this application is an increased understanding of the problem and its solution, this usage does not count as an example of agnostic science. Conversely, the microarray method, which classifies messenger ribonucleic acid (mRNA) molecules according to their function as co-occurrent with specific diseases, involves the application of machine learning to a problem that is known to exist, but for which we lack any insight regarding its structure and for which predicting an outcome is infeasible; as Napoletani et al. (2021, p. 45) argue, “[the] mechanism that leads from a certain distribution of mRNA molecules to the manifestation of a certain disease is […] rarely understood. In addition, it is also unclear which specific mRNA molecules are relevant in particular diseases”. Consequently, although it provides us with meaningful correlations, the microarray method does not increase our understanding of the problem and its solution-structure is not transparent. That is, ‘agnostic science’ according to Napoletani et al. refers to whether we gain any understanding of a phenomenon when applying big data and machine learning methods in science, while Anderson’s notion of ‘agnostic science’ is that these methods can be applied without

[3] Classical scientific methods (i) use models and/or theories and experiments to test hypotheses (hence are also called hypothetico-deductive methods), (ii) strongly rely on correlations, (iii) see the predictive power in models and/or theories, (iv) relate to human understanding and human cognition. This is how we see it, but also how the authors we quote understand what classical scientific methods are. Hepburn and Andersen (2021) provide further insights.

any prior expert knowledge or theory. Thus, the importance of expert knowledge in applying these methods will have different ramifications for the different notions of 'agnostic science'. We will return to this discussion in Sect. 6.1.

In his book on big data, Wolfgang Pietsch analyses the extent to which a method based on the use of big data and machine learning is similar to inductive methods—particularly to variational induction in its experimental form, which is known as 'exploratory experimentation'. He concludes that, despite the differences between the two approaches, an underlying epistemological similarity between them provides a means by which we can more clearly understand big data and machine learning: "big data approaches allow [us] to analyze a wide range of phenomena that are not accessible to conventional exploratory experimentation. This makes a huge difference to scientific practice in data-rich special sciences like medicine or the social sciences" (Pietsch, 2021, p. 69). Pietsch further argues that—as with the inductive and experimental methods—some traditional theoretical background must underlie any successful application of big data and machine learning, which clearly contradicts the view of Anderson. We will return to the discussion of the role of theory in regards to big data and machine learning methods in Sect. 6.2.

3 Big data and machine learning methods

Given the centrality of 'big data and machine learning methods' to our discussion, it is essential to clarify our precise terms of reference. Machine learning is the science of making machines (specifically computers) 'learn' from data by enabling them to discover certain patterns in that data. Because learning is an essential part of intelligence, machine learning is often regarded as a subfield or offshoot of artificial intelligence. Traditionally, machine learning is divided into three types: supervised, unsupervised, and reinforcement learning. Supervised learning refers to situations where the training data provide labelled examples of the pattern the machine is trying to learn, while unsupervised learning requires a machine to look for patterns without any prior direction or labelled examples of the pattern it should learn. Finally, reinforcement learning refers to a machine learning through interaction with its environment by acquiring data from environmental responses to particular actions in particular states.

All three types of machine learning involve finding mathematical functions that map input to output. Supervised learning in particular is centred on finding functions, referred to as 'machine learning algorithms' or 'machine learning models', that can be inferred from training examples of matching input–output pairs. These machine learning algorithms come in different classes, such as linear regression, decision trees, random forest, and deep neural networks. Within each class, particular algorithms can be specified in different ways depending on particular parameters; the process of finding the optimal values for these parameters is exactly what constitutes the 'learning' step in machine learning.[4]

[4] This learning step is itself performed by an algorithm. Thus, one can regard the field of machine learning as the development of algorithms that can produce other algorithms from data (Domingos, 2015).

The approach taken in our two case studies, which we examine in Sect. 4 of this paper, is often referred to as 'deep learning', in which deep neural networks are used to carry out machine learning. Deep neural networks comprise layers of connected neurons that process data in a manner that progresses from input to output. The number and types of layers, as well as the connections between them, are referred to as the model's architecture. For any particular model architecture, there is a corresponding set of parameters whose optimal values are obtained in the step of learning (or 'fitting') the model.

The term 'big data' is often employed as a generic term that simultaneously refers to the size of any individual dataset, the methods that are used to analyse such data, and the entire approach of undertaking such analysis on such data. We do not define the term precisely here, but instead simply indicate some key features of big data analyses. In contrast to traditional data analysis, where data are often collected through carefully planned, randomised controlled trials, big data utilizes data that are inherently digital in terms of collection, storage, and analysis. Each of these three aspects has required the invention of new methods and techniques in the fields of computer science and engineering. For the remainder of this paper, it is sufficiently precise to define big data and machine learning methods as a set of methods and algorithms that uses significant computational resources to discover patterns in vast datasets.

4 The role of expert knowledge

Big data and machine learning methods are not passive; a researcher does not simply feed raw data into an algorithm and then wait for it to detect correlations between certain features of a massive dataset. All of these data must be manipulated and cleaned, for instance—acts that require significant expert knowledge of scientific applications.[5] In addition to expert knowledge about the data involved, specific knowledge of machine learning is also often necessary when using these methods, as algorithms cannot be applied blindly in practice. Instead, a promising model architecture must be selected, appropriate data augmentation techniques must be applied to improve the algorithm's performance, and the algorithm itself must be tuned and adjusted.

As such, although we agree that the kind of expert knowledge used in big data and machine learning methods may be different from that required in traditional methods, we argue that this difference is not fundamental—a point to which we return in Sect. 6. Here, we concern ourselves with the extent of expert knowledge that is required for big data and machine learning methods to function efficiently. By carefully examining two scientific applications of big data and machine learning—skin cancer detection (Esteva et al., 2017) and protein folding (Jumper et al., 2021)—we assess the role played by expert knowledge in each case.

[5] This is similar to what Leonelli (2016, p. 16) calls 'data packaging'—a process that involves the selection, formatting, standardisation, and classification of data, and is typically done before publishing data in large public scientific databases. The aim in the examples presented here, however, is not to publish the data in a database, but to use the data as training data for a machine learning model.

4.1 Skin cancer detection

In a study published in 2017, Esteva et al. trained a deep neural network to classify different types of skin cancer based on images alone. The researchers fed approximately 130,000 images labelled with types of skin lesions into a pre-trained deep convolutional neural network (CNN) and then enabled it to classify previously unseen images into one of 757 lesion types. The final network classified these images of skin lesions at a level of accuracy that corresponded to that of a professional dermatologist.

Initially, this study may seem to prove that computer scientists with no specialist knowledge of dermatology can achieve in an afternoon what dermatologists train for years to be able to do. Yet, while such impressions reflect a common perception of deep learning, the truth is far more complex. To achieve their results, Esteva et al. put in a significant amount of work at many different levels, both drawing on their own experience with deep learning and relying on considerable dermatological expert knowledge; as such, their work illustrates how big data and machine learning research requires a combination of different kinds of expert knowledge and skill.

Prior to the study, the 129,450 images that would ultimately be used to train the network were labelled by dermatologists to indicate the type of skin cancer represented in the image. Thus, considerable dermatological knowledge was used to create the large dataset that was a prerequisite for training the deep neural network in this case; such use of labelled training data is the key to supervised machine learning. Some of the 2032 different types of skin lesions contained in the images occurred very rarely in the dataset; as this can cause problems when training massive machine learning models, researchers had to find a way to create classes with more examples of each type. To do this, they developed a precise and hierarchically organised taxonomy of different skin lesions—a taxonomy that itself could exist only thanks to other researchers' extensive prior work in dermatology. Based on this taxonomy, an algorithm could be used to group the images into 757 different classes, each of which had enough sample images to train a neural network effectively.

The final stages of testing were also highly dependent on prior dermatological knowledge. Because deep neural networks of this size are extremely powerful and flexible models, it is usually not difficult to make them learn efficiently from training data. However, this practice carries the risk that the neural networks will learn specific errors and noise from the training data that can significantly degrade their ability to generalise to images beyond the training data —in machine learning, we are not interested in performance on training data, but in the ability generalise to new unseen test data. The final testing of the neural network, in which the algorithm's performance was compared with the combined performance of more than twenty dermatologists, was thus crucial for gauging an appropriate estimate of the network's ability to generalise. While this testing may seem extraneous to the training or construction of a deep neural network, it is essential in confirming the network's success and is therefore a necessary part of any research based on deep learning applications. Furthermore, the criteria for the success of a deep neural network may depend on its application, and thus the appropriate design of the testing phase may require relevant expert knowledge.

Finally, we argue that a more elementary level of dermatological expertise is also important for several of the design choices taken during the research process. First, one must know that skin lesions are usually visually classifiable based on skin images. While this may seem like a trivial point, it is important to bear in mind when collecting the correct training data; one most know to collect images of skin lesions rather than X-rays, sound waves, or patients' recent meals. Basic domain knowledge of the problem in question is also necessary for the proper use of such data once collected. In this study, for example, images of skin lesions were randomly rotated and flipped vertically to create a more diverse training set—a method called 'data augmentation', commonly used when training CNNs on images. Without basic domain knowledge, however, such data augmentation could be applied in unhelpful ways; if the task is facial recognition, for example, image rotation and vertical flipping do not make any sense. Thus, although neither researchers nor the deep neural network required any specialist knowledge of the biology of cell division or the causes of skin cancer, significant domain expert knowledge was required to develop a successful scientific application of deep learning on this large dataset.

4.2 Protein folding

Another example of the novel application of machine learning in science is the prediction of the 3D structures of proteins based on their amino acid sequences. How proteins fold is an important issue in molecular biology, as how a protein folds largely determines its function. Determining how a particular protein folds is a long-standing biological problem that has traditionally been solved for individual proteins one at a time through highly expensive and time-consuming experimental methods. Thus, the development of algorithms or automatic methods that can determine how a protein folds based on its amino acid sequence is an important task. Indeed, this task has been the object of the biennial competition CASP (*Protein Structure Prediction Center*, n.d.), in which different teams of scientists compete with one another to formulate the best predictions of folded proteins' 3D structures.

Although there has been significant progress in this enterprise, the problem has not yet been fully solved. In 2020, however, a team of scientists from DeepMind entered CASP14 with their AlphaFold algorithm, which went on to beat the competitors by a considerable order of magnitude and, in the opinion of many observers, solved the problem. This scientific achievement is expected to have far-reaching consequences in the biological sciences, and the journal *Nature Methods* declared protein-folding prediction the scientific method of the year in 2021 (Method of the year 2021, 2022).

AlphaFold is a collaborative effort by a group of scientists (the paper—Jumper et al. (2021)—describing the work lists 34 authors) and it draws on expert knowledge in physics and biology, as well as ingenuity in deep learning. According to the authors, "[u]nderpinning the latest version of AlphaFold is a novel machine learning approach that incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments into the design of the deep learning algorithm" (Jumper et al., 2021, p. 583). The work is comprehensive in scope and incorporates expert knowledge in numerous phases; for the sake of the present discussion, we highlight a few key points from the authors' description of the work.

Although the problem of protein folding centres on determining a 3D structure from an amino acid sequence, the AlphaFold algorithm takes a more elaborate data structure as its input, combining multiple sequence alignments (MSA)[6] and pairwise features to generate a specially designed representation of the 3D structure as its output. The data that are used to train the model contain historical information about known 3D structures of proteins taken from the Protein Data Bank Archive (PDB), a data archive amassed by structural biologists over more than 50 years (wwPDB consortium, 2018). Thanks to their efforts, a substantial body of expert knowledge is available to provide the requisite training data for AlphaFold.[7] Furthermore, the collection of MSA data used in the AlphaFold algorithm involved multiple design decisions and was constructed on the basis of searches in multiple protein databases. The compilation of the final training data for the algorithm was a more involved process than is outlined here; we limit our discussion to these examples, which show that the data used for AlphaFold were not a simple inductive, automatically generated dataset blindly fed into an algorithm, but rather the product of many decades of expert knowledge accumulated by biological researchers.

Pre-processing the complex data produces both an MSA representation and a pair representation, both of which are then fed into a deep neural network. These data representations are specific to the problem of protein folding and would not be appropriate for other problems (such as the skin cancer image detection problem discussed above). Moreover, this data representation influences the structure of the AlphaFold algorithm through a specifically designed module, dubbed 'Evoformer' by Jumper et al., which uses attention mechanisms from other state-of-the-art deep learning networks to exchange and simultaneously update the MSA and pair representations through 48 layers of computation. The intuition behind the pair representation is that it will correspond to distances between elements of the amino acid sequence in a 3D space and will thus satisfy the triangular inequality of distances, which is captured by a tailor-made updating mechanism in the Evoformer module. After the Evoformer module follows a structural module, which contains a geometry-aware Invariant Point Attention mechanism designed specifically for this problem that translates the MSA and pair representations into a 3D structure. Thus, the deep learning algorithm used in AlphaFold is not a standard algorithm or architecture, but one carefully and deliberately designed by experts in protein folding and machine learning, incorporating knowledge from biology and physics.

Evaluation of the AlphaFold algorithm's performance was a complex process, conducted in accordance with a twofold set of criteria (as is the case in most machine learning projects). First, an evaluative component was involved in the training and fine-tuning of the algorithm, in the sense that a particular loss-function was designed

[6] MSA data refer to sequence data of proteins that are evolutionarily similar to a given protein and have been aligned to the given protein's sequence. Thus, such data again belong to the body of expert knowledge within biology and bioinformatics.

[7] Leonelli (2016) elaborates in great details how such large databases in biology are created through data packaging (see the footnote 5). Moreover, she further unfolds the importance of this work and its reliance on decisions made by people and institutions.

(which, in turn, was composed of six different components— see Jumper et al., 2021, Supplementary Information). Second, the final algorithm was evaluated to establish its expected future performance (as well as to compare it with other CASP participants). Here too, criteria specific to the protein folding problem were used, such as measurements taken to ascertain the root-mean-square deviation on the carbon atom backbone of the protein and the local Distance Difference Test (Jumper et al., 2021). Finally, the paper contains a detailed ablation study that demonstrates that the success of the AlphaFold system relied on the collection of a variety of different mechanisms, all of which contributed in some way to its overall accuracy (Jumper et al., 2021). This once again demonstrates how the success of AlphaFold is the result of an incredibly detailed and finely tuned piece of expert-specific work.

5 A taxonomy of expert knowledge in research driven by big data and machine learning

In the previous sections, we discussed two case studies (skin cancer detection and protein folding) and highlighted numerous instances in which expert domain-specific knowledge and expert knowledge of machine learning were involved in research in reciprocal ways. Based on these observations, we now suggest a classification of the various kinds of expert knowledge involved in the application of big data and machine learning methods, categorising these methods' use in four main contexts: (i) defining the problem, (ii) creating the dataset, (iii) pre-processing the data, and (iv) testing and validating the trained model. One could also consider the use, or deployment, of the trained model as another context, but one could also argue that this is beyond the scope of research and is an aspect of product development and operations instead. As the extent to which expert knowledge is involved in the use of a trained model does not affect the other contexts listed, nor the main argument of the paper, we refrain from discussing it further here.[8] In addition to these contexts, the task of choosing a proper model architecture or machine learning algorithm often also requires elaborate knowledge of big data and machine learning, and in some cases domain-specific knowledge as well. Before we elaborate on this, however, we discuss each of the four contexts mentioned above.

5.1 Defining the problem

The cases discussed show that domain-specific expert knowledge and machine learning expert knowledge are both frequently needed when selecting the problems to address with big data and machine learning methods and determining how to formulate these

[8] It is important that trained machine learning models are used on data that resemble the data they are trained on, both in quality and in distribution, as otherwise the model might not perform as well as tests initially show. Moreover, changes in real-world conditions may lead to data drift (change in data distribution) or concept drift (change in the meaning of the response variable); in fact, the issue of deploying, monitoring, retraining, and using machine learning models in production has given rise to an entire new practice called machine learning operations, or MLOps (Mäkinen et al., 2021; Sculley et al., 2015). The extent to which these tasks involve expert knowledge has yet to be determined but is beyond the scope of this paper.

problems precisely. When selecting a scientific problem to resolve with big data and machine learning methods, one's chances of success often rely on certain assumptions about the domain of expertise in question. In the case of skin cancer detection, for instance, it was assumed that a sensible categorisation of skin lesions could be made only on the basis of visually detectable differences at a certain image resolution. Conversely, in the case of protein folding, researchers assumed that the 3D structure of a protein could be determined from information about which amino acid chains constitute the protein. Furthermore, both cases required that particular output structures for the designed deep learning algorithm be defined, such as the particular taxonomy of skin lesions or the particular means of representation for the 3D structure of a protein. Finally, the success of big data and machine learning methods depends on the stability of the phenomenon being modelled; the data generated by the phenomenon and used to train the model must be stationary, i.e., they remain a good representation of the phenomenon (Northcott, 2020; Pietsch, 2015).

5.2 Creating the dataset

The process of skin cancer prediction described above is a prime example of a supervised machine learning problem, as it requires training data that contain explicit positive and negative examples to guide the algorithm's prediction. In this example, every image was labelled as either 'no cancer' or the type of cancer depicted. Such labelled training data cannot be collected automatically but require labelling by dermatologists, as well as theoretical reasoning to establish a taxonomy of skin cancer. In the case of protein folding, researchers relied on the Protein Data Bank Archive, a vast body of biological domain-specific knowledge carefully collected over a 50-year period. Moreover, a particular theory pertaining to sequence alignment was used to create the MSA data representation. Thus, novel applications of big data and machine learning methods often rely on carefully created data that require varied and extensive expert knowledge.[9]

In addition to its use in labelling training data, expert knowledge may also be required to determine what training data to retrieve. In work on skin cancer, for instance, should the images be black and white or colour? Should the lighting, resolution, or angle be set in certain ways? Are there examples of skin cancer that should not be included, or examples of skin lesions that are not cancerous but should nevertheless be included as negative examples? Does a patient's skin type or age matter for these predictions? The answers to these and other similar questions could all influence the data that ought to be retrieved. Crucially, these questions predominantly involve expert knowledge and are also connected to the task of defining the problem, which was examined in the previous sub-section. Finally, the improper labelling of data can have severe consequences beyond the determination of scientific truth; numerous studies have shown how biases in labelled data can transfer to become biases in algorithms,

[9] This is true for scientific applications of machine learning but even more so for industrial applications, which have now created an entire new labour market of AI (Crawford, 2021). Of course, researchers are attempting to develop forms of machine learning and AI that could label training data automatically, but it is inconceivable that such techniques would ever be able to eliminate the need for expert knowledge in creating data at the frontiers of science.

creating unfair or unethical predictive machine learning applications (Barocas et al., 2019).

We have phrased the issue here as that of creating the data required rather than on that of collecting it because ‘collecting’ has a more passive connotation than ‘creating’. As we have demonstrated, extensive expert knowledge was involved in the delivery of the required data in both cases. The process of creating public datasets is referred to as ‘data packaging’ by Leonelli (2016) who carefully elaborate on the journey data often need take from collection to being exposed in public databases. In this journey data are often formatted, standardised, and categorised by what she calls ‘data curators’ to fit a format in a database that is deemed most useful to the research community. As our case studies also show, however, data in such databases often require additional processing before they can be used to train machine learning algorithms. Moreover, the fact that data creation entails an active work on the data means additionally that there is not always a clear-cut distinction between data creation and data pre-processing.

5.3 Pre-processing the data

Data pre-processing is the preparation of data for analysis or use in training a machine learning algorithm; other terms for the process are ‘data wrangling’, ‘data transformation’, ‘data cleaning’, and ‘tidying’. It is often stated that this stage takes 70–80% of the time in a data analysis or machine learning project (Wickham, 2014). No standard approach encompasses all of the data pre-processing that might be required; as Hadley Wickham states, “[l]ike families, tidy datasets are all alike but every messy dataset is messy in its own way” (Wickham, 2014, p. 2). Common subtasks of data pre-processing include reshaping the data, re-formatting the data, removing or imputing missing values in data, dealing with outliers, augmenting the data, and feature engineering. Let us consider a dataset containing information about people’s features, such as their age or height: should the age of a person be a number or an age interval? If a person’s height is missing in the data, should the person be excluded from the dataset or should her height be imputed from the mean height of people in the dataset? If a person is reported to be 210 years old, what should be done with this obvious outlier? Should more features, such as body mass index, be added to the data? Such questions are just a few simple examples of how such pre-processing could be conducted. While some steps in pre-processing data can be performed without any domain-specific expert knowledge, such knowledge is often required in dealing with (and discovering) missing values and outliers. For instance, different contexts may require data to be imputed in different ways; a missing temperature reading on a particular day could often be replaced with the mean temperature of the surrounding days, but a missing sales number reported on a particular day might indicate a closed shop, which would lead the missing value to be replaced with a zero rather than a mean from the surrounding days. Finally, it should be noted that training a machine learning algorithm is an iterative process that often involves researchers going back to perform additional data pre-processing.

Reports on scientific findings generally give little attention to the pre-processing of data, even though this might have constituted substantial work in the various iterations of the research. Several examples of pre-processing are reported in our two cases, however. In the skin cancer example, as explained previously, data augmentation was used to create additional training data. In addition, data transformation or reformatting was performed in the creation of the final objective (often referred to as the 'response variable')—in this case, the labelling of images according to the taxonomy. In the protein folding example, the elaborate creation of the MSA and pair representations both constitute a step in data pre-processing.

Data pre-processing is rarely discussed at length, but it is essential for successful research. Although various software vendors sell tools that purport to make data pre-processing easy or superfluous, and a significant body of research on the creation of automatic machine learning pipelines (AutoML) has developed, it is highly unlikely that the need for data pre-processing will ever be removed from big data and machine learning applications, especially in the case of novel scientific applications.

5.4 Testing and validating the model

Generalisability is an important virtue in science. Within classical statistics used in randomised controlled experiments, generalisability is achieved through the careful sampling of participants. Researchers in machine learning, however, have sought another approach to ensure generalisability. Here, data have been divided into training and test sets; training data are used to train the algorithm, while test data are used only in the final stage to provide an unbiased estimate of the machine learning algorithm's overall performance. Assuming that the test dataset is a representative sample of the target population, the algorithm's claim to generalisability would resemble the procedure followed in classical statistics. However, the test dataset might not always be a good representation of the problem in mind; as such, other test datasets can be used, as seen in the example of skin cancer detection, in which the algorithm's performance was evaluated through the addition of new images that had also been labelled by twenty dermatologists. In the protein folding example, the different algorithms in the CASP competition were tested on proteins whose 3D structures had recently been discovered experimentally but had not yet been published.

Furthermore, it is one thing to determine what data should be used to test a machine learning algorithm, but another to determine how to measure this algorithm's performance. If the algorithm generates a particular label as its output, as in the skin cancer example, a natural goal would be to measure the number of mistakes made by the algorithm. Classification algorithms such as this often generate a probability or degree of confidence that a particular example belongs to a particular lass, however; as a result, the quality of the algorithm's prediction can be measured in other ways. In the case of protein folding, for example, we saw that 3D structures of proteins could be compared in multiple ways and that several metrics were used in both the training and the final evaluation of the model.

Finally, machine learning algorithms produce different types of errors, and because some of these will be more important than others, a simple tally of errors is an insufficient measure of success in many cases. For instance, if an algorithm is used in an initial screening for skin cancer, it might be preferable to mistakenly flag a potential skin cancer that can later be dismissed by further tests rather than to mistakenly overlook a case of skin cancer and conduct no further tests.

The proper evaluation of a big data or machine learning algorithm is essential for ensuring the generalisability of its results and guaranteeing that the work constitutes a scientific contribution. It is also essential for judging the feasibility of future applications of the algorithm. This testing and validation require domain-specific and machine learning expert knowledge to ensure that the right data and measures are used and to interpret the results within the wider scientific context.

5.5 Selecting model architecture

Although the skin cancer example used a model architecture that would be considered standard for such a task today, its use still necessitated certain design choices, such as the size of the convolution layers of the CNN architecture; moreover, the protein folding example demonstrates that a standard architecture is sometimes entirely insufficient for a task, in which case new custom-made architecture such as the Evoformer and the Invariant Point Attention mechanism must be developed. The creation of these new architectures involved significant research by skilled machine learning researchers at Google DeepMind; they were not developed in isolation by a team of researchers solely concerned with machine learning, however, but by an interdisciplinary team of researchers working on the AlphaFold algorithm, as Jumper et al. (2021, p. 583) state in their analysis of the algorithm: “Underpinning the latest version of AlphaFold is a novel machine learning approach that incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments, into the design of the deep learning algorithm”. Thus, the creation of model architectures for big data and machine learning methods requires experts not only within machine learning but also within a variety of other scientific fields.

6 Big data and machine learning in scientific practice

Our claim that expert knowledge is required at multiple stages in the application of big data and machine learning methods carries ramifications for philosophical questions about science. In this section, we address some of these questions, including: to what extent can these methods be seen as agnostic? How much do they resemble inductive procedures? What role, if any, does theory play in big data and machine learning-driven research? To what extent does the application of big data and machine learning methods represent a paradigm shift in science? Our goal in this final section is merely to indicate areas in which this paper’s analysis of the role of expert knowledge may provide researchers with additional ideas or suggest new perspectives.

6.1 To what extent are big data and machine learning methods agnostic?

We earlier cited two distinct uses of the term ‘agnostic’ in the literature, one arising from the popular view and one arising from the informed view on big data and machine learning in the sciences. Thus far, we have considered, with reference to specific examples, the extent to which expert knowledge is involved in research that applies big data and machine learning methods and have shown that such knowledge is involved at multiple stages, including the creation of the proper training data, the pre-processing of the data, and the development of testing and measurements of objectives. If the popular view of ‘agnostic’ science is taken to mean that big data and machine learning methods can be used without prior expert knowledge or theory, or that these methods easily generalise to other fields, no compelling evidence supports the existence of such agnostic science.[10]

In the conception of ‘agnostic’ science proposed by Napoletani et al. (2011, 2014, 2021), however, this ‘agnosticism’ seems to pertain to the understanding of phenomena that is gained by the application of the big data and machine learning methods; for this reason, these authors might not disagree with our claims made in the last paragraph. In their discussion of mRNA and diseases, Napoletani et al. (2021) argue that the use of big data and machine learning methods in this case might qualify as ‘agnostic’ science because we do not gain any understanding of the phenomenon under investigation. The examples discussed above corroborate this position; in the case of skin cancer detection, we do not gain any further understanding of skin cancer by being able to detect its different types with a deep learning algorithm, nor does the AlphaFold algorithm teach us anything about why proteins fold in the way that they do.[11]

Thus, the necessity of expert knowledge is consistent with that big data and machine learning methods are not agnostic in the popular view, but may be agnostic in the informed view.

6.2 The role of theory in applying big data and machine learning methods to science

The role of expert knowledge in the scientific application of big data and machine learning methods is also related to the debate about the role of theory in science—a controversy sparked in part by the contentious and influential critique made by Chris Anderson (2008), outlined above, in which Anderson “envisioned a future of atheoretical, automated science” (Napoletani et al., 2014, p. 3) and maintained that the use of big data and machine learning were leading to the ‘end of theory’. This controversial

[10] Furthermore, in terms of applicability to a wide range of fields, the same can be said for the use of classical statistical methods (such as the t-test).

[11] Because no understanding of a phenomenon is obtained through the application of big data and machine learning methods, one might question whether these examples constitute science at all, or whether they should instead be viewed as cases of engineering. A discussion of the definition of ‘science’ or its relation to engineering is outside the scope of this paper; here, we note only that the history of science contains many examples of scientific research that do not concern themselves solely with understanding phenomena. For instance, David Baird (2004) has highlighted the importance of the development of scientific instruments and the scientific knowledge they constitute.

but attractive claim prompted a lively debate on the question of whether there can be any understanding without theory in science. Many intellectuals, including writers, scientists, and philosophers, took part in this debate and marked out their positions.

Mieke Boon (2012) provides a broader context, contending that the debate surrounding the role of theory in big data and machine learning can be regarded as a natural extension of historical ideas about the objectivity of scientific methods and the efficacy of various scientific explanations (cf. Duhem, 1914/1954; Hempel, 1962, 1966; van Fraassen, 1980). In this debate, concepts such as 'the objectivity of scientific methods', 'mathematical models', and 'logic-based theories' are criticised as "arbitrary intellectual instruments to fit the data" (Boon, 2012, p. 50). Viewed in this way, data science offers a method that is "not confined by the kinds of idealizations and simplifications humans need to make to fit data into comprehensive mathematical formalisms" (p. 51). After all, regularities, patterns, and correlations discovered by humans may turn out to describe the world no more adequately than the regularities, patterns, and correlations detected by machines.

Our careful mapping of the role of expert knowledge in the discussion above adds further arguments to this debate. We have seen how experts' theoretical background is involved in data generation, problem formulation, and algorithm evaluation. This observation implies that even if arbitrariness can be imputed to human-induced factors (at least at the current state of the art in big data and machine learning methods), pre- and post-analytic human involvement cannot be avoided in practice. As Rob Kitchin (2014) suggests, arguments that emphasise the need for theoretical underpinnings when formulating a scientific problem and selecting an algorithm are usually based on the observation that no methodologically sound scientific inquiry can be based solely on 'raw data', because raw data never occur in a 'scientific vacuum' but are always "discursively framed by previous findings, theories, and training; by speculation that is grounded in experience and knowledge" (p. 5).

Wolfgang Pietsch (2015) takes enquiry into the inherent theoretical commitment even further when he distinguishes between two ways in which 'data-intensive science', understood as a form of inductive science, including big data and machine learning methods,[12] is theory-dependent: data-intensive science is theory-laden in the *external* sense as it is dependent on theoretical assumptions held by the researcher, but it is not theory-laden in the *internal* sense because it does not depend on or participate in the formation of theory. Both we, emphasising that the initial and post-analytic involvement of a certain theory underlying any expert intervention is unavoidable, and Kitchin, arguing that no 'raw data' resides in a 'scientific vacuum', qualify data-intensive science as theory-laden only in the external sense. None of us make claims about the internal sense.[13]

[12] Wolfgang Pietsch (2021) has defended the thesis that the big data and machine learning method is a form of a specific type of induction known as variational induction. Although we agree with Pietsch that the training of machine learning algorithms resembles the process of variational induction, we argue that the mere training of a machine learning algorithm should not be regarded as a scientific method on its own, because the stages of data creation, data cleaning, and the evaluation and testing of the machine learning algorithm are all necessary steps if the process is to be considered scientific.

[13] Proponents of the internal sense of theory-ladenness claim that, even if big data and machine learning do not currently have the ability to create or contribute to theory, there is no reason to believe that this will not

6.3 Big data and machine learning as a new scientific paradigm

In his famous talk to the Computer Science and Telecommunication Board on 11 January 2007, Jim Grey predicted the following:

> [A]lmost everything about science is changing because of the impact of information technology. Experimental, theoretical, and computational science are all being affected by the data deluge, and a fourth, 'data-intensive' science paradigm is emerging. The goal is to have a world in which all of the science literature is online, all of the science data is online, and they interoperate with each other. Lots of new tools are needed to make this happen. (Hey et al., 2009b, p. xxx)

While we find Grey's enthusiastic presentation to be a useful visionary endeavour, we argue that it lacks the scientific grounding to establish a serious argument for a fourth scientific revolution and paradigm shift.

The possibility for such a paradigm shift involving big data and machine learning methods has been widely explored. Napoletani et al. (2011) have argued that big data and machine learning methods have the potential to solve problems that the current form of understanding-based science cannot solve. They emphasise that a new methodological paradigm, dependent on a new conception of science, is appearing:

> Instead of attempting to understand and model a phenomenon, this paradigm suggests that a scientist needs to approach a phenomenon with a limited set of assumptions, and needs to look for specific techniques capable to solve some of the problems it presents, without attempting any sort of structural understanding of the phenomenon itself. (p. 6)

In contrast, Robert Northcott (2020) has reached the opposite conclusion; based on four case studies of the predictive power of big data methods, he finds that in all cases this predictive power is limited, concluding that "they suggest caution about whether prediction, and thus scientific method generally, will really be revolutionized by big data" (p. 103). Pietsch (2016, p. 138) observes that reference to a 'change of paradigm' is misleading from a philosophical perspective because it diverges from the classical Kuhnian sense, although he agrees that one can adequately speak of an emergence of a novel scientific methodology.

We also suggest that the relative frequency of use of traditional methods and big data and machine learning methods will vary depending on the field, so that the extent to which each science incorporates 'agnostic' elements will differ, as will the extent to which each domain manages to rule out any reference to theory. Consequently, we argue that big data and machine learning methods will become incorporated into scientific methodology through continuous small conceptual shifts rather than a rigid paradigm shift in the Kuhnian sense.

Yet another interesting line of thought might involve considering to what degree the application of these methods will lead to scientific progress; while we leave a deeper discussion of this for future research, it should be mentioned that the problems that

Footnote 13 continued
change in the future; Donoho (2000), for example, speculates that "[t]he present approach will eventually be replaced by another, more traditional approach, which relies on new, yet undiscovered, theories".

these methods are used to solve may fit well with a functional approach to scientific progress in comparison to the more widespread epistemic and semantic approaches (Bird, 2007; Shan, 2019).[14]

6.4 Limitations of big data and machine learning methods

By mapping the role of expert knowledge in the scientific applications of big data and machine learning methods, we have shown that a complete departure from expert knowledge and theory is unlikely, and thus that these methods have a limited potential to replace traditional scientific methods and lead to an entirely automated science. Here, we briefly review other researchers' similar theoretical arguments on the limitations of big data and machine learning methods, categorising one as formal and one as philosophical.

Formal arguments centre on the impossibility of the emergence of an entirely automated science. Calude and Longo (2017), for instance, provide mathematical evidence to demonstrate the impossibility of consistently determining those correlations that are relevant in big data analyses, claiming that "the more data, the more arbitrary, meaningless, and useless (for future action) correlations will be found in them" (p. 600). Their argument is based mainly on Ramsey's theory, which essentially states that there are regular patterns in any sufficiently large set of mathematical objects; this theory—combined with the observation that no regularity can lead to predictability in dynamical systems and with the theorem that algorithmic information is always randomly distributed—proves that most correlations are insignificant and do not allow for any scientific generalisations. We do not replicate the technical details of this argument here, but the basic point—that big data and machine learning analyses depend on human-driven causal understanding—should stand.

Philosophical arguments emphasise the limitations of big data and machine learning methods based on analyses of the epistemological assumptions of scientific methods. Boon (2012) critiques the optimism surrounding the effectiveness of data science and machine learning methods and data models by making a distinction between 'useful' theories (theories that provide the basis for developing applications) and 'true' theories (theories that, assuming scientific realism, adequately model reality). As she states:

> Even if it were possible to obtain the data-models from the machine, they would be useless for epistemic uses by humans as these data-models do not meet relevant pragmatic criteria to enable such uses. The other way around, in order to be useful for humans in performing epistemic tasks, scientific knowledge must also meet pragmatic criteria. (p. 59)

Boon's argument follows the same trajectory as ours. She continues:

[14] According to the epistemic approach, scientific progress should be defined in terms of knowledge, such that progress occurs as more knowledge is accumulated, while the semantic approach defines scientific progress in terms of truth, such that progress occurs when science converges closer to the truth (Bird, 2007). In the functional approach, scientific progress instead occurs when a certain function, such as problem-solving, is fulfilled or, alternatively, when a piece of science is useful (Shan, 2019).

> Every tiny step in these intricate research processes involves epistemic tasks—e.g., to explain, interpret, invent, idealize, simplify, hypothesize, model, mathematize, design, and calculate—for which all kinds of practical and scientific knowledge are crucial and needs to be developed in the research process. Therefore, scientific knowledge needs to be *comprehensible* to the extent that it allows for these epistemic tasks. (pp. 61–62)

Consequently, Boon writes, "it is inconceivable that machine-learning technologies will make science and scientists superfluous" (p. 62). Like Boon, we point out the necessity for expert knowledge at multiple levels of data processing. Boon argues for this necessity on the grounds that any dataset should meet pragmatic criteria and be useful in human applications; our arguments, conversely, are based on observations drawn from scientific practice.

Another philosophical argument on the limitations of big data and machine learning methodologies is based on the observation that no matter how capacious a dataset is, there may always be new cases that humans will know belong in the dataset, but that algorithms will not be able to recognise as such. In the example of the skin cancer dataset presented in Sect. 4, for instance, it can be anticipated with great certainty that currently unknown types of skin cancer will be found; these will have to be manually added to the dataset, and machine learning algorithms will have to be trained on them anew.[15]

7 Conclusion

This paper argues against the view that science based on theory, models, and hypotheses can be replaced by atheoretical and automated algorithmic methods. As our research indicates, any data analysis or machine learning project begins with a well-formulated scientific problem, systematic data collection, data pre-processing, model training, and model evaluation; in other words, the entire data-mining process is carried out in an expert- and theory-driven manner. Moreover, the evaluation of machine learning algorithms on new test datasets and their comparison with other methods (such as human performance) is an important component of the research procedure that ascribes seriousness and credibility to the results. Although we hesitate to postulate that it will never be possible to exclude human input from big data and machine learning methods, our two case studies and our discussion in the previous section show that this possibility is highly unlikely.

It is not our intention here to dismiss new scientific methods or new lines of research. Rather, we seek to emphasise the research expertise and domain knowledge required by these new methods; in so doing, we aim to show in detail what is truly new and what is only 'business as usual'. Big data and machine learning methods may be used in a more 'agnostic' way, but they do not lead to completely agnostic science. We do not believe that these methods will lead to a radical change or revolution in science, but rather that they represent a (considerable) expansion of science's methodological

[15] Inspiration for this argument comes from the oral presentation given by Marija Slavkovik at the workshop 'Philosophy of Computing', held at the Warsaw University of Technology in September 2021.

toolkit. Even if big data and machine learning approaches do not revolutionise all of science, they will still lead to changes in subfields and stimulate the emergence of new fields or endeavours, such as the digital humanities or computational social sciences.

Acknowledgements Prior to publication, we presented the work at several seminars, conferences and workshops, most notably at HaPoC-6 in Zurich. We are grateful to those who discussed with us the ideas we presented, in particular we are indebted to Marco Panza, who challenged us fiercely. We thank our anonymous reviewers, whose comments allowed us to clarify ambiguities, add missing references and significantly improve the paper. Paula Quinon received an internal grant from the Faculty of Administration and Social Sciences of the Warsaw University of Technology that enabled us to improve the standard of our work, we are grateful for having been given this opportunity.

Declarations

Conflict of interest We have no conflict of interest to report.

References

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *WIRED, 16*(7). Retrieved from https://www.wired.com/2008/06/pb-theory/

Ankeny, R., Chang, H., Boumans, M., & Boon, M. (2011). Introduction: Philosophy of science in practice. *European Journal for Philosophy of Science, 1*, 303–307. https://doi.org/10.1007/s13194-011-0036-4

Baird, D. (2004). *Thing knowledge: A philosophy of scientific instruments*. University of California Press.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning*. fairmlbook.org. Retrieved from https://fairmlbook.org/

Bird, A. (2007). What is scientific progress? *Noûs, 41*(1), 64–89.

Boon, M. (2012). Scientific concepts in the engineering sciences: Epistemic tools for creating and intervening with phenomena. In U. Feest and F. Steinle (Eds.), *Scientific concepts and investigative practice* (pp. 219–244). De Gruyter. https://doi.org/10.1515/9783110253610.219

Calude, C. S., & Longo, G. (2017). The deluge of spurious correlations in big data. *Foundations of Science, 22*(3), 595–612. https://doi.org/10.1007/s10699-016-9489-4

Crawford, K. (2021). *Atlas of AI*. Yale University Press.

Domingos, P. (2015). *The master algorithm*. Penguin Books.

Donoho, D. L. (2000). *High-dimensional data analysis: The curses and blessings of dimensionality* [Lecture]. Mathematical Challenges of the 21st Century, Los Angeles.

Duhem, P. (1914/1954). *The aim and structure of physical theory*. Princeton University Press.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115–118. https://doi.org/10.1038/nature21056

Hempel, C. G. (1962). Explanation in science and philosophy. In R. G. Colodny (Ed.), *Frontiers of science and philosophy* (pp. 9–19). University of Pittsburgh Press.

Hempel, C. G. (1966). *Philosophy of natural science*. Prentice-Hall.

Hepburn, B., & Andersen, H. (2021). Scientific method. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2021 edition). The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University. Retrieved from https://plato.stanford.edu/archives/sum2021/entries/scientific-method/

Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009a). *The fourth paradigm: Data-intensive scientific discovery*. Microsoft Research.

Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009b). Jim Gray on eScience: A transformed scientific method. Based on the transcript of a talk given by Jim Gray to the NRC-CSTB in Mountain View, CA, on January 11, 2007. In T. Hey at al. (Eds.) *The fourth paradigm: Data-intensive scientific discovery* (pp. xvii–xxxi) Microsoft Research.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., & Bridgland, A. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*. https://doi.org/10.1177/2053951714528481

Leonelli, S. (2016). *Data-centric biology: A philosophical study*. University of Chicago Press.

Mäkinen, S., Skogström, H., Laaksonen, E., & Mikkonen, T. Who needs MLOps: What data scientists seek to accomplish and how can MLOps help? In *2021 IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI (WAIN)* (pp. 109–112). https://doi.org/10.1109/WAIN52551.2021.00024

Method of the year 2021: Protein structure prediction. (2022). *Nature Methods*, *19*(1), 1. https://doi.org/10.1038/s41592-021-01380-4

Napoletani, D., Panza, M., & Struppa, D. C. (2011). Agnostic science: Towards a philosophy of data analysis. *Foundations of Science, 16*(1), 1–20. https://doi.org/10.1007/s10699-010-9186-7

Napoletani, D., Panza, M., & Struppa, D. C. (2014). Is big data enough? A reflection on the changing role of mathematics in applications. *Notices of the AMS, 61*(5), 485–490. https://doi.org/10.1090/noti1102

Napoletani, D., Panza, M., & Struppa, D. (2021). Agnostic structure of data science methods. *Lato Sensu: Revue de la Société de Philosophie des Sciences*, *8*(2), 44–57. https://doi.org/10.20416/LSRSPS.V8I2.5

Northcott, R. (2020). Big data and prediction: Four case studies. *Studies in History and Philosophy of Science Part A, 81*, 96–104. https://doi.org/10.1016/j.shpsa.2019.09.002

Pietsch, W. (2015). Aspects of theory-ladenness in data-intensive science. *Philosophy of Science, 82*(5), 905–916. https://doi.org/10.1086/683328

Pietsch, W. (2016). The causal nature of modeling with big data. *Philosophy & Technology, 29*, 137–171. https://doi.org/10.1007/s13347-015-0202-2

Pietsch, W. (2021). *Big data*. Cambridge University Press.

Protein Structure Prediction Center. (n.d.). Protein Structure Prediction Center. Retrieved June 8, 2022, from https://predictioncenter.org/index.cgi

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J. F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems, 28*, 2503–2511.

Shan, Y. (2019). A new functional approach to scientific progress. *Philosophy of Science, 86*(4), 739–758. https://doi.org/10.1086/704980

van Fraassen, B. C. (1980). *The scientific image*. Clarendon Press.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1–23. https://doi.org/10.18637/jss.v059.i10

wwPDB consortium. (2018). Protein Data Bank: The single global archive for 3D macromolecular structure data. *Nucleic Acids Research, 47*(D1), D520–D528. https://doi.org/10.1093/nar/gky949