Forthcoming in *Philosophy of Science*

The Interplay of Data, Models, and Theories in Machine Learning

Maria Federica Norelli, Ioannis Votsis and Jon Williamson

Abstract

This paper discusses the role of data within scientific reasoning and as evidence for theoretical claims, arguing for the idea that data can yield theoretically grounded models and be inferred, predicted, or explained from/by such models. Contrary to Bogen and Woodward's skepticism regarding the feasibility and epistemic relevance of data-to-theory and theory-to-data inferences, we draw upon scientific artificial intelligence literature to advocate that: a) many models are routinely inferred and predicted from the data and routinely used to infer and predict data: b) such models can, at least in some contexts, play the role of theoretical device.

1. Introduction

Almost four decades ago, Jim Bogen and James Woodward, hereafter B&W, proposed to sharply distinguish between data, phenomena and theories. Data are thought of as typically observable, idiosyncratic to a particular context of investigation and methodology, and relatively easy to collect and analyse. Phenomena are conceived of as for the most part unobservable and investigator-independent. In a typical scenario, scientists gather data through direct observation or measurement, and use it as evidence for the existence of a phenomenon of interest. Subsequently, scientists develop theories—abstract and general frameworks aiming to make, predict, and/or explain claims about phenomena, not data. By making this distinction and positing an indirect relationship between data and theories, B&W challenge the long-standing tradition rooted in empiricism of taking data/observations as the basis for developing and evaluating scientific theories. While their account does not entirely dismiss the role of data within scientific practice, as it still figures in the process of phenomena detection, it regards any data-theory inference as misguided and epistemically irrelevant.

This paper acknowledges the significance of the data-phenomena distinction, while challenging B&W's epistemological claims regarding the role of data in theory construction and evaluation. We argue that modern-day scientific applications of machine learning methods as well as causal discovery methods provide compelling examples of how data are routinely inferred and predicted from theoretical models and routinely used to infer and predict such models. The paper is structured as follows: Section 2 presents a fairly detailed account of B&W's tripartite distinction, as well as their associated epistemological claims. Section 3 discusses two cases from machine learning (ML) that directly challenge B&W epistemological claims. Section 4 discusses a further case from the causal Bayes nets (CBN) literature that also calls into question those claims. Section 5 helps ground the importance of those cases by motivating the theoretical nature of ML and CBN models. Section 6 offers a brief summary.

2. B&W's Account

In this section, we discuss B&W's epistemological account of the relationship between data, phenomena, and theories. Among other things, we provide examples and review some of the most poignant reactions against it.

B&W developed their account in a series of papers that stretch across several decades (1988; 1989; 1992; 2003; 2011). Data, or observational reports, are "public records produced by measurement and experiment that serve as evidence for the existence or features of phenomena" (Woodward 2011: 166). Examples include "bubble chamber photographs, patterns of discharge in electronic particle detectors, and records of reaction times and error rates in various psychological experiments" (Bogen and Woodward 1988: 306). By contrast, phenomena are "relatively stable and general features of the world which are potential objects of explanation and prediction by general theory" (1989: 393).[1] Examples include "weak neutral currents, […], proton decay, capacity limitations and recency effects in short-term memory" (ibid.).[2] Scientific theories are abstract and general frameworks that "generally seek to predict and systematically explain" phenomena (Bogen and Woodward 1992: 592). Examples include "classical mechanics, general relativity, and the electroweak theory that unifies electromagnetic and weak nuclear forces" (Woodward 2011: 166).

Crucially, on B&W's account, phenomena are explained by theories and can provide evidence for them, but data only serve as evidence for phenomena and can neither be explained or inferred from theories nor explain or infer theories. When proposed, this account offered a radical new philosophical vision of the relationship between data and theories. In the past, philosophers took phenomena to be the kind of things that we directly perceive or observe.[3] As such, and according always to B&W, philosophers were misidentifying the concept of phenomena with that of data. Their 'folly' has roots in the etymological meaning of the term 'phenomenon'. As Hacking notes, "In Greek it denotes a thing, event, or process that can be seen, and derives from the verb that means, 'to appear' " (1983: 221). B&W instead assert that it is the data, not the phenomena, that constitute the actual objects of observation. While data are dependent on the context and methods involved in their production and collection,

[1] Similarly, "phenomena — i.e., features of the world that in principle could recur under different contexts or conditions" (Woodward 2011: 166).
[2] Woodward suggests that phenomena are considerably more difficult to detect: "Detecting a phenomenon is like looking for a needle in a haystack or […] like fiddling with a malfunctioning radio until one's favourite station finally comes through clearly" (1989: 438).
[3] See, e.g., Nagel (1940) and Carnap (1950).

phenomena can be essentially regarded as investigator-independent. This understanding aligns well with the way scientists, particularly physicists, conceptualise phenomena, which doesn't necessarily involve direct observation, and may even involve typically unobservable entities. In principle, we have no issue with the fact that B&W are employing the term 'phenomenon' in a way that departs from its historical and philosophical lineage, at least insofar as it is coherently defined (which we cautiously deem it to be).

It is worth discussing one example, namely the melting point of lead (Bogen and Woodward 1988: 308-309), to illustrate the way B&W understand the relationship between data, phenomena and theories.[4] To detect the phenomenon of the true melting point of lead, scientists collect data/observations by measuring the temperature of lead samples, using a thermometer. The 'final' dataset consists of a scattering of temperature reading points. Assuming that there is no systematic error, and a certain level of control of small causes of variation – such as that the confounding variables "operate independently, are roughly equal in magnitude, are as likely to be positive as negative, and have a cumulative effect which is additive" (ibid.) – B&W claim that the resulting distribution of measurements will be normal, and, hence, its mean can be taken to be a "good estimate of the true melting point of lead [i.e. the phenomenon at issue]" (ibid.). That phenomenon, they argue, is not directly observable. On the supposition that the *true* melting point of lead is 327.5°C, even if a temperature reading of 327.5°C were obtained, it would not amount to a direct observation, but rather to a special case of its occurrence. Notably, the very same phenomenon could have been identified, even if this precise value had never occurred. The scattered data points simply provide evidence for the phenomenon of lead's melting point, and this phenomenon provides evidence for, and is explained by, theories related to electron bonding.

According always to B&W, we cannot use such theories to infer, predict, or explain individual data points because the mean of the observed distribution is not a property of any specific data point and is unlikely to coincide with any observed value expect in exceptional cases. Data, even when collected and processed under controlled experimental conditions, are inevitably constrained by 'local' factors that do not reflect or relate the 'non-local' phenomenon under investigation. These 'local' factors can be numerous and complex and are intimately tied to the specific measurement setup and experimental contexts—ranging from how the measuring instrument is affected by ambient conditions, to procedural decisions made by the

[4] Bogen and Woodward borrow this example from Ernest Nagel's *Structure of Science*. Other examples they discuss include starlight deflection and weak neural currents (Woodward 2011: 166).

researcher, and challenges in ensuring repeatability. It is the local or idiosyncratic nature of data, i.e., being specific to the particular methods and contexts of production or collection, that B&W exploit in making the claim that data cannot be inferred, predicted or explained by theories. In their own words, "one can justifiably believe that data provide reliable evidence for some phenomenon without being in a position to explain or derive facts about the data" (Woodward 1989: 398).[5]

Critiques of B&W's account have generally fallen into three broad categories. The first category concerns those who question the very need to demarcate between data and phenomena. Glymour, for instance, belongs to this category, arguing that "Bogen and Woodward are mistaken in thinking that the distinction is necessary" (2000: 29). He restricts his argument to statistical inferences and asserts that either B&W's distinction corresponds to the distinction between sample structure and population structure, or it does not. Either way, he argues, it is unnecessary. If it does not correspond to it, the distinction is unnecessary because inferences from sample structure to population structure can be made without it. If it does correspond, then the distinction is also unnecessary because it "simply gives a new name to a distinction which is already deeply embedded in the literature on statistical inference" (34).

The second category concerns those who object to the way the distinction is drawn. Teller (2011), for example, argues that the distinction between observables and unobservables cuts across the distinction between data and phenomena.[6] Similarly, Leonelli (2009) argues that the distinction between local vs. non-local cuts across the distinction between data and phenomena. She gives as an example the case of biology, where data typically travel beyond their original experimental contexts, through the use of standardized data labels, decontextualization procedures, and classification practices for the comparison of different datasets. These practices transform data from highly localized evidence for one claim into generalized evidence for other claims in entirely different research contexts. Furthermore, she argues that claims about phenomena are not intrinsically non-local, offering examples where phenomena only play a local evidential role towards theories, e.g. the terms 'pathogen' and 'cell wall' carry distinct meanings within different model organism communities.

[5] Here are some other relevant quotations making pretty much the same points: "there is often no obvious scientific rationale or motivation for attempting to provide detailed systematic explanations of data" (Woodward 2011: 167), "there will be little scientific point […] in deriving data […] even if this were possible" Bogen and Woodward (1988: 323).

[6] Having said this, Woodward (2011: 171) revises B&W's earlier stance, contending that the assertions that phenomena are typically unobservable, and data are typically observable, are both unnecessary for their argument.

The third category concerns those who are inclined to accept the data-phenomena distinction but not its epistemological implications. Schindler (2007), for example, argues that unobservable phenomena cannot be inferred from observable data without the involvement of theory. Votsis (2011) argues that inferences from data to theories and back to data are feasible, and even useful, once suitable auxiliaries are introduced.[7] Similarly, Lusk (2021) raises doubts about the inability of theories to explain data via derivation. He discusses the case of spectral data in atmospheric science. Such data cannot be employed to independently identify gas concentration profiles, since different profiles can yield very similar absorption spectra (a problem known in physics as the *inverse problem*). So, scientists rely on theoretical radiative transfer models to derive and explain predicted spectra from possible gas concentration profiles. Contrary to B&W's claim, then, theoretical explanations via derivation that link data to phenomena appear to be helpful in the scientific enterprise.

3. Machine Learning Models and Data

In this section, we explore two cases from the application of ML techniques to science. We argue that both involve the inferential construction of models from data but also the inference of (something like) data from those models. Moreover, we argue that such models function as theoretical devices. As such, they challenge B&W's assertion that data-theory inferences are either unfeasible or, at best, irrelevant for the construction/evaluation of scientific knowledge.

In recent years, there has been an exponential growth surge in data production. This increasing influx of data is enabling new possibilities for *data-driven* research and has had a profound impact on scientific research. One notable domain is medical diagnosis. Data from imaging, genetics, and clinical records, coupled with the advancements in data science techniques, particularly ML, have yielded remarkable results, rivalling the capabilities of human physicians. Kadir and Gleeson (2018) provide a comparative study of various ML approaches for predicting lung cancer from CT scan images of nodules. Their study encompasses risk models, radiomics, and convolutional neural networks (CNNs) and revealed that CNNs achieve the best performances for predictive tasks, with an AUC score – a common metric of model performance of diagnostic tests – approaching 0.9, where an AUC score of 1 indicates a perfect predictor.

One powerful aspect of CNNs is their ability to independently learn and extract features from input data, eliminating the need for manual feature engineering. This stands in contrast

[7] Woodward (2011) offers replies to both Schindler (2007) and Votsis (2011).

with, e.g., traditional radiomics approaches, where predefined sets of engineered features are manually selected by radiologists. In fact, multi-layer CNNs adopt a hierarchical approach to feature extraction. In image processing, this means progressing from basic edges to more complex patterns. CNNs' ability to learn without relying on manually extracted characteristics is directly linked to its successful achievement of high classification rates. Indeed, it has been showed that inter-radiologist variability in measuring nodule features can significantly impact predictions. The winning model of the 2017 Kaggle Data Science Bowl, provides a useful example of the relevance of CNNs in lung cancer detection.

The model was trained on a combination of multiple datasets consisting of CT scan images of lung nodules labelled as malignant or benign based on histology/follow-up information taken from the LIDC-IDRI public clinical dataset. Throughout the training, the model learned to identify patterns and features within the images that are indicative of whether a nodule is malignant or benign, and optimized its learning parameters through iterative adjustments to minimize the disparity between its predictions and the actual labels in the training data. Once the training was completed, the model transitioned to the prediction phase. In this stage, the model was applied to new, unseen CT scan images, and its predictions were expressed as likelihood scores indicating the probability of a patient's nodule being malignant or benign. Notably, the winning team not only achieved classification rates comparable with those achieved with traditional techniques but helped to recognize the impact of nodule sizes on algorithm performance and, more generally, its importance for an early cancer diagnosis. Indeed, when confronted with datasets where nodule sizes were deliberately matched between malignant and benign cases or encountered variations in nodule sizes between datasets, the model kept a strong ability to discriminate between benign and malignant nodules, indicating that it was picking up on morphological features beyond just size. Furthermore, the authors show that the model actually achieved an approximately 0.2 higher AUC score on the size-matched dataset. This challenges the conventional wisdom that nodule size is the primary indicator of malignancy and prompts further investigations in the use of multiple nodule features such as, for instance, texture and shape for malignancy risk assessment.

Similarly, Nemlander et al. (2022) shows how ML can advance our risk assessments of lung cancer. The assessment involves different models, all based on patients' smoking history and self-reported symptoms. A challenge in the early detection of lung cancer is that symptoms are often non-specific and common. To address this challenge, Nemlander et al. employ a stochastic gradient boosting (SGB) algorithm to train and test different risk models on the same dataset. The SGB algorithm is a type of ensemble learning, a class of methods to derive and

combine multiple base models into a single final model to enhance robustness and accuracy. The two major ensemble techniques are *bagging,* where each model is independently fitted using a randomly selected subset of the original dataset, and *boosting*, where fitting is sequential, building on the results of the previous model. All models are then aggregated by summing them up, and the final classification for each observation is determined based on the most frequently occurring classification among all the models. Nemlander et al. successfully identified chest pain as a crucial symptom for risk assessment and early diagnosis of lung cancer, despite its undervaluation in more traditional models.

What do the above cases tell us B&W's claim that there are no direct relations between data and theories? Let us first consider the data-to-theory direction. The cases presented above challenge this part of B&W's claim, as models are directly and inferentially constructed from data. By 'directly', we do not mean purely. ML models are not constructed from data alone, but the process of constructing them is largely data-driven.[8] In the CNN case, the training data are CT scan images of lung nodules that are labelled as malignant or benign. In the SGB case, the training data are smoking history and self-reported symptoms. Let us next consider the theory-to-data direction. The cases presented above also challenge this part of B&W's claim, as the trained models are unleashed onto the world to make inferences/predictions that directly bear on the data. While these predictions are not observations or data per se, they can be conceived of as simulated data. At any rate, the predictions can be directly checked against new or previously unseen data. If the predictions come out true, the models are further confirmed. If they come out false, the models must be retrained, or, in the worst case, abandoned.

To sum up this section, ML practices underscore both the feasibility and the importance of bidirectional data-theory reasoning. Reasoning from data to theory can be instrumental in identifying previously unanticipated phenomena or links between existing phenomena. Simultaneously, employing theories to make inferences/predictions that have a direct bearing on the data plays a crucial role in assessing the validity and accuracy of those theories.

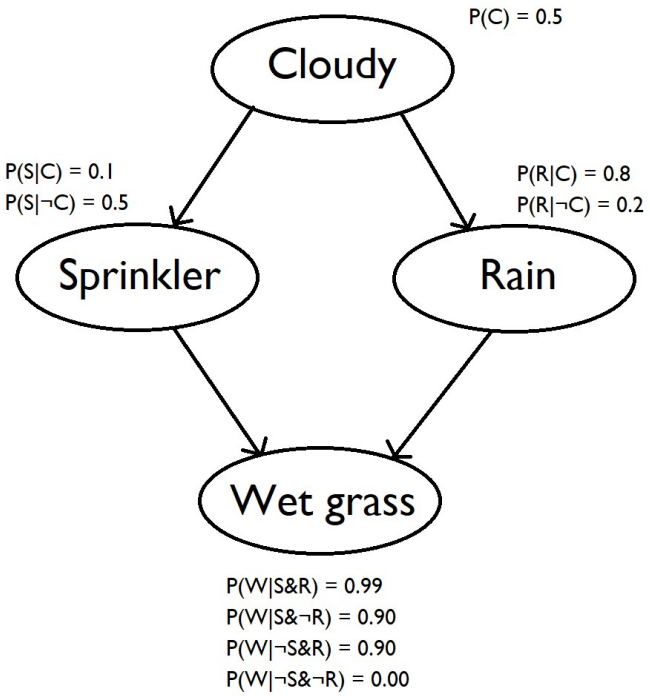
4. Causal Bayesian Nets and Data

[8] Other considerations that play a role in the construction of ML models include various hyperparameters, e.g. the learning rate, gamma and regularization.

A causal Bayesian network (CBN) is a kind of model that can encapsulate theoretical claims and yet can also directly predict and explain observable data.

CBNs model causal relationships amongst a set of variables of interest, as well as the joint probability distribution of those variables (Williamson 2005). A CBN consists of (i) a directed acyclic graph whose nodes represent the variables of interest and whose arrows represent causal relationships amongst those variables, and (ii) the probability distribution of each variable conditional on its parents in the graph. The joint distribution can be captured by the CBN if we assume the *Causal Markov Condition*: that each variable is probabilistically independent of its non-effects, conditional on its direct causes in the graph.

Consider the following simple example (Murphy 1998):



This CBN says that cloudy conditions affect both sprinkler use and rain, which in turn are both causes of grass being wet. Given the causal Markov condition, the above conditional probability tables then suffice to determine the joint distribution over these four variables.

CBNs are widely used in ML and AI, because there are many algorithms for constructing a CBN directly from data, and many algorithms for using the CBN to draw predictions, including predictions about the effects of interventions. For example, one can use the above CBN to predict how deciding not to use a sprinkler will influence the probability of wet grass.

To see how CBNs can work in practice, consider the work of Xue et al (2019) on discovering disease mechanisms that differentiate cancer subtypes. Xue et al. used ML

algorithms to learn causal relationships between somatic genome alterations and differentially expressed genes within tumours from the Cancer Genome Atlas dataset. This allowed them to divide breast cancers into five mechanistically distinct subgroups and glioblastomas into six subgroups. These different mechanisms lead to different clinical outcomes in patients, and the authors conclude that their approach can identify clinically relevant disease subtypes.

This example illustrates the connection between CBNs and data. In one direction, data can be used to directly construct a CBN. Thus, there is a direct link from data to the model. For example, CBNs for breast cancer can be learned directly from the Cancer Genome Atlas dataset. In the other direction, CBNs predict the distribution of data: for example, they can be used to predict disease outcomes of patients with a particular kind of breast cancer. New data can then be collected to test these predictions, and thereby confirm or disconfirm the CBN model.

In addition, there is a clear sense in which the CBNs of Xue et al. (2019) encapsulate theoretical claims—claims about the differential disease mechanisms of breast and brain cancers. Thus CBNs can act as scientific theories, can be constructed directly from data, and can be confirmed or disconfirmed according to whether their predictions about data are borne out. These theoretical claims can also be used to explain differences in outcomes amongst cancer patients, i.e., differences in data.

These facts about CBNs do not sit well with the claims of Bogen and Woodward, who say that 'data typically cannot be predicted or systematically explained by theory' (Bogen & Woodward 1988, pp. 305-6). In particular, in the CBN case, there are no phenomena that mediate the connection between data and theory.

5. Theoretically Vested Models

A reader may be sympathetic with the claim that there is a bidirectional relation between, on the one hand, ML and CBN models, and, on the other hand, data, without agreeing that these models possess theoretical credentials. In this section, we attempt to put this objection to bed by motivating why the aforesaid models deserve those credentials.

Sizeable disagreement exists concerning the right conceptualisation of theories and models. Some claim that theories are families of models, while others claim that theories are sets of sentences. Some claim that models play a subordinate role to theories, while others claim that models have greater, and, even, complete independence from theories. Setting aside these fascinating debates, it is generally agreed that theories typically have greater scope and are more abstract than models. Still, both theories and models are thought of as capable of

representing, generalising over and providing explanations of phenomena (and in our view also data). That models can be explanatory is made clear by various participants in the aforementioned debates. Bokulich (2011), for example, reports that there is "widespread use of models to explain phenomena in science" (33). Craver (2006) insists that "[m]odels are explanatory when they describe mechanisms" (367). Even Woodward, in his most celebrated work on the manipulability account of causation, takes causal models to explain phenomena.

We take the ability of models to represent, generalise over and provide explanations of phenomena to be a sufficient condition for the attribution of theoretical credentials. After all, despite metaphysical disputes about the nature of theories, nearly all parties seem to agree that theories play those roles.[9] Indeed, some philosophers, make this connection more explicit: "models embody different theoretical virtues… [s]cope, precision, specificity, accuracy, generality, completeness, simplicity,… providing understanding, being explanatory, and being predictively successful" (Frigg 2022: 431). Having said this, we concede that most models are narrower in scope than theories in their generalising and explanatory reach.[10] As such, they may be considered as low-level theories. To give an example, the Bohr model of the hydrogen atom can be regarded as a low-level theory (when compared to a theory as high-level as quantum mechanics), precisely because it seeks to represent, generalise and explain various phenomena, namely aspects of electron behaviour such as quantized electron orbits. Given their obvious ability to represent, generalise and explain phenomena, we hereafter say that such models are 'theoretically vested'.

What remains to be seen is whether the kinds of models we paraded in sections 3 and 4 are indeed theoretically vested, with particular emphasis to whether they are explanatory. In the case of CBN models, this is fairly straightforward to establish. Such models attempt to represent causal relations (a type of generalisation) and support explanations via counterfactuals. For example, a CBN model may assert that only changes in the value of a variable X systematically bring about changes in the value of a variable Y. From this, one can infer and explain that if no changes occur in the value of variable X, then no changes occur in the value of variable Y. Coincidentally, this is also the opinion of Woodward, who has the following to say about the relationship between counterfactuals and explanations:

[9] Exceptions include French (2020) who is an eliminativist about theories. Far from standing in our way, French's insistence that the locus of scientific activity is the model, which, among other functions, can be explanatory, provides support for the kind of view advocated here.
[10] Even so, there is no reason why a model cannot be constructed that is as wide in scope as any theory.

> The notion of information that is relevant to manipulation thus needs to be understood modally or counterfactually: the information that is relevant to causally explaining an outcome involves the identification of factors and relationships such that if (perhaps contrary to fact) manipulation of these factors were possible, this would be a way of manipulating or altering the phenomenon in question (2003: 10).

The case of ML models being explanatory is, by contrast, harder to establish. That's because of issues relating to the *black-box* nature of many such models. Complex ML models produced via deep learning, are notoriously difficult to interpret and thus cannot be relied on to be explanatory. The SGB model presented earlier is not impacted by this problem, as it does not employ neural nets, but rather symbolic representations. Moreover, the representations it employs are arguably explanatory in that they track and reveal the significance of symptoms, e.g. chest pain, for risk assessment and early diagnosis in a robust way that presumably supports counterfactual probabilities. So, for example, if chest pain were not present in patient α, then the lung cancer risk assessment for that patient is aptly revised to be more optimistic. Alas, the same cannot be said about the CNN model presented earlier, as it clearly falls under the deep neural net approach to doing science. In what follows, we attempt to motivate the claim that this model does indeed have explanatory value.

The CNN model challenges the simplistic assertion that nodule size is the primary indicator of malignancy. Indeed, as already noted, it outperforms other methods of detection, even when the sizes between benign and malignant nodules are matched. The problem with this model is that it is not immediately clear which aspects of nodules, e.g. texture or shape, play a discriminatory role. Even so, we would like to argue that the model is explanatory, at least in a minimal sense of the term. That's because it clearly tells us about the relative non-significance of nodule size in lung cancer medical diagnoses. Doing so, provides the hard-earned insight that other geometrical features (since these are the only ones encoded in the processed images) must be relevant for diagnoses.

Stronger forms of explanatory value may be possible for such models. We would like to point out that there are ways to improve the interpretability, and hence potentially the explainability, of neural net models. For example, there are techniques available to identify what features are more exploited by the model, offering a partial glimpse into the underlying representation. One strategy employs *integrated gradients* (IG).[11] Broadly speaking, IG is a

[11] For an exhaustive overview of what are integrated gradients and how they work see: Sundararajan et al. 2017.

technique for attributing an importance or *attribution* value to each input feature of the model based on the gradients of the output with respect to the input, i.e., based on how the model's predictions are affected by changes in each feature. This involves systematically varying the input features along a path from a predefined reference point to the actual input, calculating gradients at each step, and integrating these gradients to understand the cumulative impact of each feature on the model's output.

Another approach to improve interpretability, and potentially explainability, is via pruning. This involves the selective removal of parameters or nodes from models, which are then trained/retrained using the original datasets (Le Cun, Denker and Solla 1989), to produce models that are "less than 10-20% of the [original model's] size" and that enjoy comparable accuracies (Frankle and Carbin 2018: 1). More impressively, evidence is surfacing that pruning can result in models that exceed the accuracy of their pre-pruned rivals. Other methods of simplifying complex models involve weight clustering (i.e. reducing the number of unique values that weights can take) and quantisation (i.e. reducing the number of bits that are needed to express the parameters) – see Freire et al. (2023) for an overview. All these methods offer hope that future iterations of neural net models may be significantly simpler, more interpretable and more explanatory.

6. Conclusion

In this paper, we sought to reevaluate B&W's data-phenomena-theory distinction and its epistemological implications. While accepting the usefulness of differentiating data, phenomena and theories, we challenged their view that data can neither be used to make inferences about theories, nor serve as evidence for evaluating them. We did so with the help of two cases from ML practice (in the domains of lung cancer medical diagnosis and risk assessment) as well as one case from CBN practice (in the domain of cancer subtype differentiation). These cases, we argued, clearly demonstrate the feasibility and promise of bidirectional reasoning between data and models. To counter the objection that models are not theories, we motivated the claim that models can be theoretically vested and illustrated how the specific models discussed exhibit this property. In more detail, we argued that those models exhibit explanatory prowess at various levels. We concluded with the thought that even those models that currently exhibit a modicum of explanatory prowess, namely deep neural net models, there is hope that one day this will be expanded.

References

Bogen, J. & Woodward, J. (1988). Saving the phenomena. *Philosophical Review* 97 (3):303-352.

Bokulich, A. (2011). How Scientific Models Can Explain. *Synthese*, 180(1): 33–45.

Carnap, R. (1950). Empiricism, semantics, and ontology. *Review Internationale De Philosophie, 4*(11), 20–40.

Craver, C. F. (2006). When mechanistic models explain. *Synthese* 153 (3):355-376.

Frankle, J., & Carbin, M. (2018). The Lottery Ticket Hypothesis: Training Pruned Neural Networks. *ArXiv*.

Freire, P., & Napoli, A., & Arguello R. D., & Spinnler, B., & Anderson, M., & Schairer, W., & Bex, T., & Costa, N., & Turitsyn, S., & Prilepsky, J. (2023). Reducing Computational Complexity of Neural Networks in Optical Channel Equalization: From Concepts to Implementation. *Journal of Lightwave Technology*, 1-26.

French, S. (2020). *There Are No Such Things as Theories.* New York, NY, United States of America: Oxford University Press.

Frigg, R. (2022). *Models and Theories: A Philosophical inquiry.* Routledge Taylor and Francis Group, New York.

Glymour, B. (2000). Data and phenomena: A distinction reconsidered. *Erkenntnis* 52 (1):29-37.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge University Press.

Kadir, T., & Gleeson, F. (2018). Lung cancer prediction using machine learning and advanced imaging techniques. *Translational lung cancer research*, 7(3), 304.

LeCun, Y., Denker, J.S., & Solla, S.A. (1989). Optimal Brain Damage. *Neural Information Processing Systems*.

Leonelli, S. (2009). On the Locality of Data and Claims about Phenomena. *Philosophy of Science*, *76*(5), 737–749.

Lusk, G. (2021). Saving the Data. *The British Journal for the Philosophy of Science,* 72:1, 277-298.

Murphy, K. (1998). A Brief Introduction to Graphical Models and Bayesian Networks.

Nagel, E. (1940). Charles S. Peirce: A Pioneer of modern empirism. *Philosophy of Science,* 7(1), 69–80.

Nemlander, E., Rosenblad, A., Abedi, E., Ekman, S., Hasselström, J., Eriksson, L. E., & Carlsson, A. C. (2022). Lung cancer prediction using machine learning on data from a symptom e- questionnaire for never smokers, former smokers and current smokers. *Plos one*, *17*(10).

Schindler S. (2007) Rehabilitating theory. The refusal of the bottom-up construction of Scientific Phenomena. *Studies in the History and Philosophy of Science* 38(1): 160–184.

Sundararajan, M.; Taly, A.; Yan, Q. (2017) Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 3319–3328.

Teller, P. (2010) "Saving the Phenomena" Today'. *Philosophy of Science* 77, no. 5: 815–26.

Votsis, I. (2011). Data meet theory: up close and inferentially personal. *Synthese 182* (1):89-100.

Williamson, J. (2005). *Bayesian nets and causality: philosophical and computational foundations*, Oxford: Oxford University Press.

Woodward, J. (1989) "Data and Phenomena". *Synthese* 79, no. 3: 393–472.

Woodward, J. (2003) *Making Things Happen*. Oxford: Oxford University Press.

Woodward, J. (2021) "Data and Phenomena: A Restatement and Defense". *Synthese* 182, no. 1: 165–79.

Xue, Y., Cooper, G., Cai, C. et al. (2019). Tumour-specific Causal Inference Discovers Distinct Disease Mechanisms Underlying Cancer Subtypes. *Sci Rep* **9**, 13225.