Philosophy & Technology (2023) 36:63
https://doi.org/10.1007/s13347-023-00663-2

COMMENTARY

Commentary on David Watson, "On the Philosophy of Unsupervised Learning," *Philosophy & Technology*

Tom F. Sterkenburg[1]

Received: 6 September 2023 / Accepted: 9 September 2023 / Published online: 19 September 2023

The domain of unsupervised learning, unlike supervised learning and reinforcement learning, has largely been neglected in philosophical work on machine learning. This is the premise of Watson's paper (2023), and he forges a number of interesting connections between unsupervised learning approaches and philosophical debates.

Have we unduly neglected unsupervised learning? The complaint might already seem odd in light of the current wave of philosophical attention for large language models, which are also trained with techniques based on unsupervised learning components. My own impression is rather that most philosophical theorizing about machine learning is, for better or worse, simply not so concerned with the actual kind of underlying learning mechanism. In the debate around interpretability, for instance, the main concern is with making sense of trained output models, irrespective of the actual learning procedure yielding this model (and whether it was purely supervised or had significant unsupervised—or semi-supervised or self-supervised—components). The debate around fairness, likewise primarily concerned with output models, was indeed prompted by prediction software that did not use machine learning at all (Rudin, 2019, p. 209). At the same time, those (formal) philosophers that do engage with the algorithmic details of machine learning mostly if not exclusively restrict attention to supervised procedures; and explicit philosophical study of distinctive aspects of unsupervised learning methods is hard to find indeed. Watson's very well-informed paper is therefore a welcome contribution.

As for the reason *why* unsupervised learning has been relatively neglected, I think Watson is also right to observe that "philosophers are generally partial to well-defined concepts and clear success criteria, neither of which are immediately forthcoming in this subfield" (p. 3). Watson cites Hastie et al. (2009), who further write that "in the context of unsupervised learning, there is no […] direct measure of success. It is difficult to ascertain the validity of inferences drawn from the output of most unsupervised learning algorithms. One must resort to

✉ Tom F. Sterkenburg
tom.sterkenburg@lmu.de

[1] Munich Center for Mathematical Philosophy, LMU Munich, Geschwister-Scholl-Platz 1, 80539 Munich, Germany

heuristic arguments not only for motivating the algorithms, as is often the case in supervised learning as well, but also for judgments as to the quality of the results. This uncomfortable situation has led to heavy proliferation of proposed methods, since effectiveness is a matter of opinion and cannot be verified directly" (p. 487; also see Shalev-Shwartz & Ben-David, 2014, ch. 22). I want to suggest that this also constitutes a challenge for philosophical claims about the capacities of unsupervised learning.

I will take a look at Watson's discussion of clustering, and in particular his epistemic claim (p. 6):

> "EC: We learn to identify natural kinds via clustering algorithms, or something very much like them."

Obviously it is impossible to properly evaluate such a claim without engaging more fully with the epistemology (and indeed metaphysics) of natural kinds (Bird & Tobin, 2023). Even the meaning of EC appears susceptible to a range of interpretations, from an assertion about human cognition to a claim about the dynamics of science. The motivation Watson gives for EC remains at the correspondingly broad level of an intuition that clustering algorithms give "natural" classifications, underwritten by analogies like that between hierarchical clustering and Linnaeus's sorting of biological species by recursive partitioning (p. 6). Of course, this further leaves open how close such analogies are supposed to be. If the idea is merely that both hierarchical clustering and Linnaeus's methodology arrive at something that can be visualized as a dendogram, this is indeed unobjectionable, but also not particularly interesting. If the idea is that Linnaeus was actually executing something very much like a modern hierarchical clustering algorithm, this is not so immediately plausible. (Leaving aside that the status of biological species as natural kinds is itself controversial, see Bird & Tobin, 2023, sect. 2.1.)

I would like to put the focus here on the clustering algorithms, and on what is arguably a component of EC, namely that

> EC*: Clustering algorithms identify natural kinds.

This is a claim about the capacity of clustering algorithms to identify natural kinds. What are the prospects for an interesting and plausible version of this claim?

An immediate worry is that with these algorithms, we really get out what we put in. Clustering algorithms first ask for the specification of a similarity or distance metric between the data instances (Shalev-Shwartz & Ben-David, 2014, p. 266), which is obviously a hefty choice for the problem of identifying natural kinds (to Quine, for instance, *kind* and *similarity* were "substantially one notion," see Bird & Tobin, 2023, sect. 1.2). Indeed, when it comes to identifying natural kinds, already the specific way in which the instances in the dataset are "carved out" from nature is hardly a neutral step. The worry is that the real work must already be done by these input choices, which trivializes EC*: with the right

"natural" choice of input components, clustering algorithms will (or can) identify natural kinds; with the wrong choice, they will not.

Watson explicitly addresses the "extreme takeaway" that "these methods are basically vacuous, since they do little more than spell out the statistical consequences of our own choices" (p. 18). But while I agree with the replies he gives to the general charge of vacuity, I do not think they resolve the worry about EC*.

One of the replies that Watson offers is that the concern of "anything goes" simply does not arise in practice, and in particular the theoretical possibility of "baroque" input choices is practically not relevant (pp. 18f). I agree, but that not only suggests an (unclear) restriction of the scope of EC* to clustering algorithms supplied with "natural" or "non-baroque" inputs, it also leaves untouched (even highlights) the worry that the real work is done by the right choice of input. Something similar holds for the next reply, that these algorithms "do not generally work in isolation," but "are used in conjunction with a range of other methods to build evidence for or against particular conclusions" (p. 19). Again I fully agree: these algorithms will normally operate within a wider context of inquiry. Relatedly, I also have no objections to Watson's earlier stated "broadly pragmatic view of natural kinds," where any grouping in kinds is "undertaken within some context and for some purpose" (p. 8). However, all of this just puts more weight on a constellation of further (informal) choices and (formal) methods in any discovery or determination of natural kinds, and puts further pressure on the question what work is actually done by the clustering algorithm itself. In particular, if clustering algorithms are primarily methods for exploratory data analysis (Shalev-Shwartz & Ben-David, 2014, p. 264; James et al., 2021, p. 498), must the heavy lifting not rather be done elsewhere in the relevant "discovery pipeline"? Namely, on the one side again, in a fortuitous choice or careful design of the right data and similarity metric; or on the other, in a highly nontrivial venture of "testing" the candidate natural kinds, specifically whether they satisfy typical desiderata such as permitting inductive inferences and participating in natural laws (Watson, p. 6; Bird & Tobin, 2023, sect. 1.1.1).

To rephrase the present concern, EC* has little content if we cannot say more about what it is about the clustering algorithms *themselves* that makes them suited for identifying natural kinds. The remaining reply Watson gives to the worry of vacuity concerns the algorithms themselves: "the speed and scalability of unsupervised learning algorithms are sufficiently great that […] they represent a step change in our analytic capacity, enabling large-scale data mining procedures that can reveal unexpected patterns and generate novel hypotheses" (p. 18). Again I do not object. One might similarly worry that the results of *supervised* algorithms do not strictly go beyond the given data and (implicit or explicit) inductive assumptions: but even if this were so, these algorithms are still clearly of value because (especially in the case of much data and complex or ill-understood inductive assumptions) we cannot immediately oversee these consequences ourselves. But there is a more fundamental reply still to the vacuity objection for many types of learning algorithms. Namely, in many learning approaches, what an algorithm returns does not already follow logically from what it is given as input. In the case of supervised statistical learning, the given data and the (implicitly or explicity) specified hypothesis class do not decide the prediction: though some choices of

predictions (hence, possible learning algorithms) are better than others, as supported by theoretical learning guarantees (Shalev-Shwartz & Ben-David, 2014). Likewise, unsupervised classifications do not just follow logically from a given metric: certain further choices must be involved. And the question what further choices clustering algorithms make is obviously relevant to the question what makes them suitable for identifying natural kinds.

However, here the aforementioned characteristic of unsupervised learning becomes pertinent. Unlike supervised algorithms, clustering algorithms do not come with clear success criteria and accompanying theoretical success guarantees. It is actually a difficult question "[w]hat distinguishes a *clustering* algorithm from any arbitrary function that takes an input space and outputs a partition of that space" (Shalev-Shwartz & Ben-David, 2014, p. 274). Discussing the attempt to characterize clustering by a number of plausible axioms, and the impossibility result due to Kleinberg (2002) that no algorithm can satisfy these axioms simultaneously, Shalev-Shwartz and Ben-David conclude that "there is no 'ideal' clustering function. Every clustering function will have some 'undesirable' properties" (p. 276). All of this at least suggests that we have no principled, theoretical basis for the intuition that clustering algorithms somehow carve out "natural" classifications, even from given "natural" distance metrics; that is to say, for the intuition that underlies EC* and indeed EC.

Acknowledgements Thanks to David Watson for feedback on a first version of this commentary.

Author Contributions Single author.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported by the German Research Foundation (DFG)—Projektnummer 511917847.

Data Availability Not applicable.

Declarations

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication Not applicable.

Competing Interests The author has no relevant financial or non-financial interests to disclose.

References

Bird, A., & Tobin, E. (2023). Natural kinds. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of philosophy.* Metaphysics Research Lab, Stanford University.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer Series in Statistics. Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer.

Kleinberg, J. M. (2002). An impossibility theorem for clustering. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15: Proceedings of the Annual Conference (NIPS 2002).* MIT Press.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, 206–215.

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

Watson, D. (2023). On the philosophy of unsupervised learning. *Philosophy & Technology, 36*(28), 1–28.