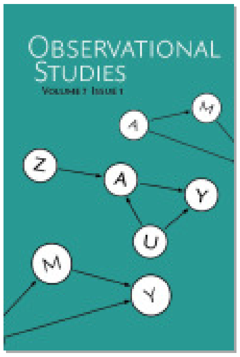
PROJECT MUSE®

Discussion of Breiman's "Two Cultures": From Two Cultures to One

Anna Neufeld, Daniela Witten

Observational Studies, Volume 7, Issue 1, 2021, pp. 171-174 (Article)

Published by University of Pennsylvania Press
DOI: *https://doi.org/10.1353/obs.2021.0004*



For additional information about this article
https://muse.jhu.edu/article/799729

[3.142.174.55] Project MUSE (2024-04-20 01:28 GMT)

Observational Studies Vol. 7.1 (2021): Published July 2021

Discussion of Breiman's "Two Cultures": *From Two Cultures to One*

Anna Neufeld aneufeld@uw.edu
Department of Statistics
University of Washington
Seattle, Washington 98195, USA

Daniela Witten dwitten@uw.edu
Departments of Statistics and Biostatistics
University of Washington
Seattle, Washington 98195, USA

Abstract

We argue that algorithmic models, though powerful and appropriate in some circumstances, rely on just as many tenuous assumptions as parametric probabilistic models; these assumptions, their violations, and the ethical consequences of these violations are simply obscured within a black box. We advocate for a future in which statisticians play a central role in bridging the gap between Breiman's two cultures.

Keywords: Algorithmic ethics, black box models, assumptions, prediction, two cultures

Breiman was ahead of his time: when computer-intensive data analysis was still in its infancy, he correctly predicted the growing importance of "algorithmic" models in both industry and academia. But while he was correct about some things, he was mistaken about others.

Breiman divides the world into two camps: the data modelers, who assume a (likely incorrect) probabilistic model, and the algorithmic modelers, who focus only on predictive accuracy. He points out that data modelers draw conclusions about an assumed model; when the assumption is wrong, the conclusions are irrelevant. He implies that by not assuming a probabilistic model, algorithmic modelers can avoid such spurious conclusions. But Breiman's suggested workflow — to fit a model on a training set and evaluate it using predictive accuracy on a test set — relies on three key assumptions:

Breiman's Assumption 1: The ultimate goal is to imitate "nature's black box."

Breiman's Assumption 2: The training and test sets are representative of the broader population of interest.

Breiman's Assumption 3: Aggregate accuracy on the test set is paramount, regardless of how the errors are distributed among subgroups of individuals.

©2021 Anna Neufeld and Daniela M. Witten.

[3.142.174.55] Project MUSE (2024-04-20 01:28 GMT)

In recent years, we have seen the danger of blithe adherence to the culture of algorithmic modeling when any of these three assumptions are violated. Here we focus on three examples.

> *Example 1: Hiring.* In an effort to automate hiring, Amazon trained a machine learning model to use a job candidate's resumé to predict whether they should be hired, using a dataset labeled with Amazon's past hiring decisions (Goodman, 2018). The model had high predictive accuracy on the test set. However, rather than learning to imitate the relationship between resumé characteristics and whether or not a candidate would be a "good employee", the algorithm learned to imitate the relationship between resumé characteristics and whether or not a human Amazon recruiter would have hired this candidate in the past. Furthermore, this modeling task implicitly assumed that the collection of (mostly male) individuals hired in the past was representative of the set of people Amazon wanted to hire in the future. Amazon discontinued the model in 2018 after realizing that it downgraded all resumés with references to women's organizations or women's colleges.
>
> *Example 2: Recidivism.* Recidivism algorithms have recently been adopted in the criminal justice system to predict the probability that a defendant will commit a future crime, with the idea that defendants with low risk scores can more safely be released or assigned a lower bail (Angwin et al., 2016). However, these algorithms are trained and tested on data that connects personal characteristics to future arrests, *not* future criminal activity. In learning to imitate the relationships in the dataset, the algorithms perpetuate bias towards highly policed populations. In fact, one particular recidivism algorithm (COMPAS) was shown to falsely label Black defendants as future criminals at twice the rate of white defendants (Angwin et al., 2016).
>
> *Example 3: Facial recognition.* Commercial facial recognition software (such as those marketed by Microsoft, Google, and Face++) boast high reported predictive accuracy (up to 97.5%) on benchmark testing sets such as the LFW database of celebrity faces, which is approximately 77.5% male and 83.5% white (Buolamwini and Gebru, 2018). The pursuit of high aggregate predictive accuracy on such a benchmark dataset favors models that are excellent at detecting white male faces, but that may achieve much lower accuracy when detecting the faces of women or people of color. In one high-profile case in 2020, an incorrect facial recognition match led to the arrest of a Black man for a crime he did not commit, demonstrating the ethical concerns that arise when the advertised accuracy of an algorithm holds in the aggregate but not for specific subgroups (Hill, 2020).

At first glance, it seems that Breiman foretold the innovative use of algorithmic models in new and creative contexts such as these. But in each example, something went wrong. In the hiring and recidivism examples, Breiman's Assumption 1 is violated: "nature's black box" carries with it societal injustices, and thus imitating it is not (or rather, should not be!) the goal[1]. These two examples also violate Assumption 2, as the training and testing

[1] While Breiman used the terminology "nature's black box", in these examples the term "society's black box" seems more appropriate.

data are not representative of the population. The facial recognition example shows that violations of Assumptions 2 and 3 can, together, lead to situations in which algorithms perform poorly on specific subgroups. Critically, these problems do not stem from the use of a neural net versus a generalized linear model; nothing about one culture versus the other can solve the problems of societal bias in "nature's black box" or unrepresentative data[2]. Breiman warned that data modelers who wrongly assume simple parametric relationships are elegantly solving the wrong problem. But in fact, *any* modeler — whether in the data camp or the algorithmic camp — who falls prey to any of Breiman's faulty assumptions risks solving the wrong problem. Interpretable models are not immune to these issues, but these issues are more easily identified when a model is interpretable. By contrast, when a model is trained on a huge number of variables and its inner workings are impenetrable, ethical issues are far more likely to slip through the cracks.

Luckily, the limitations of pure prediction algorithms are well-understood by many of today's researchers. To see this, one need look no further than the 2020 programs of NeurIPS and ICML, two of the flagship conferences for the algorithmic culture. While predictive accuracy remains a major focus, the topics of interpretability, inference, causality, fairness, and ethics are of growing interest. The European Union's General Data Protection Regulation now guarantees a "right to explanation" for all algorithmic decisions, essentially outlawing black boxes and providing major incentives for research into topics such as explainable deep learning (Xie et al., 2020). Efron (2020) highlights "two hopeful trends", which involve (i) adding mechanisms to pure prediction algorithms to achieve interpretability and inference; and (ii) translating insights from the pure prediction culture (such as strategies for the $p \gg n$ setting) to traditional probabilistic models. In our own ongoing work, we are developing p-values and confidence intervals associated with the terminal nodes of Breiman's CART regression trees (Breiman et al., 1984), in recognition of the importance of quantifying the uncertainty in the tree structure caused by sampling variation.

Critically, while non-statisticians played a key role in developing the algorithmic culture, statisticians are taking center stage in improving our understanding of those algorithms, and — in effect — merging the modeling and algorithmic cultures.

Looking forward, we believe that there will be only one culture: tomorrow's successful data analyst will be well-versed in the classical data modeling toolkit, as well as in algorithmic modeling approaches. Someone who is skilled only in the former will be left behind in a fast-paced world of terabyte-sized data, whereas someone who exclusively focuses on the latter will perpetually fall prey to unreasonable assumptions that are baked into impenetrable black boxes. A data analyst will be a jack-of-all-trades, who can understand the scientific contexts, the statistical models, the algorithmic approaches, and the ethical implications — or else who is aware of their own limitations, and can collaborate with others to fill in the gaps.

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *Pro Publica*, 2016. `https://www.propublica.org/article/machine-bias-risk-assessments`

[2] In fact, COMPAS, a widely-used recidivism algorithm, is a black box because it is proprietary, not because it is a particularly complicated model.

-in-criminal-sentencing.

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and Regression Trees.* CRC press, 1984.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

Bradley Efron. Prediction, estimation, and attribution. *Journal of the American Statistical Association*, 115(530):636–655, 2020.

Rachel Goodman. Why Amazon's automated hiring tool discriminated against women. *ACLU*, 2018. https://www.aclu.org/blog/womens-rights/ womens-rights-workplace/why-amazons-automated-hiring-tool-discriminated-against.

Kashmir Hill. Wrongfully accused by an algorithm. *New York Times*, 2020. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html.

Ning Xie, Gabrielle Ras, Marcel van Gerven, and Derek Doran. Explainable deep learning: A field guide for the uninitiated. *arXiv preprint arXiv:2004.14545*, 2020.