PROJECT MUSE®

Leo Breiman's Challenge: A Retrospective

David Banks

Observational Studies, Volume 7, Issue 1, 2021, pp. 13-15 (Article)

Published by University of Pennsylvania Press

DOI: <https://doi.org/10.1353/obs.2021.0017>

➔ *For additional information about this article*

<https://muse.jhu.edu/article/799742>

Leo Breiman's Challenge: A Retrospective

David Banks

dlbanks@duke.edu

Department of Statistical Science

Duke University

Durham, NC 27708, USA

Abstract

Breiman led the way in thinking differently about statistics. Many of his iconoclastic ideas have become standard in the data science sphere. This discussion argues for some rebalancing, while gratefully acknowledging his achievements.

Keywords: Causal Inference, Data Engineering, Interpretability

Discussion

The revolution that Leo Breiman led was both necessary and successful. But like Hector before the gates of Troy, I shall defend a cause that in many ways has already been lost.

Black box prediction has become the industry standard. New software, new attitudes, and a new hunger for data-driven decision making in IT companies have created a context in which the default approach is deep learning, or some Word2Vec application, or Leo's own Random Forest technique. Such analyses have become easy, effective and fast.

It is slower, and takes more thought, to do a study that permits interpretation. But my clever colleague, Cynthia Rudin, has persuaded me that often there is an interpretable analysis whose predictions are as good as the black box forecasts (cf. Rudin and Radin, 2019). Persi Diaconis once told me that "Honest work is never wasted," and I think that precept applies to interpretable modeling.

This journal is named *Observational Studies*, and it focuses upon causal inference. That emphasis cries out for interpretability. Yes, one can use black box techniques to create propensity scores and such, but causal understanding requires insight into the mechanism that relates conditions to consequences. Absent that insight, one has learned nothing.

In my experience, many of the datasets that seem to demand the "algorithmic modeling" culture actually arise from a mixture of simpler data generation processes (Banks et al., 2009). For example, if one is studying click-through data in computational advertising, there are several distinct reasons why a person might click, and we only get to observe the results of the superposition of all those processes. Specifically, some people click on an ad for an expensive car because they are actively shopping. Other people click because they are scouting the market, trying assess resale value. And some people cannot afford the car, but like to fantasize.

Putatively, one could regard such data as arising from a mixture model in which each of the components is interpretable. And that understanding could drive useful decisions. For example, the advertiser would try to distinguish the serious shoppers, and focus on them. The technology for identifying these people is not arcane; e.g., a fantasist would have

a multi-year record of clicking on expensive cars ads, and the scout probably would not search as widely.

Michael Jordan distinguishes data science from data engineering. One is doing data science when one looks for statistical evidence of the Higgs boson or examines small wobbles in distant stars to infer the presence of an extrasolar planet. But one is doing data engineering when one churns through massive data sets to refine Amazon’s recommender system for books or to tweak Uber’s ride sharing algorithms.

In our context, data science requires interpretability, or the data modeling culture that Leo described. If one cannot explain what one has discovered, it is difficult to call it science. But data engineering agrees happily with algorithmic modeling. The engineer’s goal is to improve performance, and although interpretability might help with that, given our new software tools, it is no longer essential.

There are other reasons to cleave to the academic heritage of modeling. For one, modeling is usually fun and helpful. Sitting down and carefully thinking about how the data are produced is salutary, especially insofar as it detects the dominant factors and flags those that are peripheral. Everyone who does statistical consulting needs that skill, even if that person ends up doing an algorithmic analysis in the end.

Also, having a model helps one to figure out the best question to ask of the data, and to structure the analysis in ways that make sense. It enables us to have realistic expectations of the data. For example, the signal-to-noise ratio is a construct derived from the modeling culture, but it usefully informs algorithmic analyses.

Buja and Lee (2001), in an undeservedly unremarked upon paper, examined the impact of a small change in the CART algorithm. Instead of making the split that most reduces the sum of the sum-of-squares error from both of the daughter nodes, they made the split that most reduced the sum-of-squares error in just one of the nodes. (There were some minor additional constraints to avoid such cheats as peeling off one case at a time.) Then the reanalyzed the Boston Housing Data.

The original tree from the CART analysis of the Boston Housing Data was, as my colleague Mike West would say, “a bit of a pig’s breakfast” (see Breiman et al., 1984, Fig. 8.1, p. 219). But Buja and Lee’s re-analysis produced a highly unbalanced but very interpretable tree. The first split peeled off high-crime neighborhoods, and the second split off a predominantly African-American community. Next were six splits, in monotonic order, on the percentage of lower-status residents, followed by five monotonic splits on the number of rooms. Buja and Lee then rewrote the model with linear functions of the monotone splits replacing branches, obtaining insights far more helpful to Boston realtors.

In 1985, I was fortunate to audit a graduate class that Leo taught. He clearly enjoyed the classroom—he was remarkably comfortable presenting his speculative ideas and discussing, perhaps even debating, with students. His personality made a powerful impression—he was stubborn, brilliant, and, as I perceived him, a bit gruff. But most of all, he was one of the most completely independent thinkers I have known.

In the context of his willingness to discuss ideas, Leo’s “Statistical Modeling: The Two Cultures” generated wonderful commentary 20 years ago, and I am grateful that *Observational Studies* is continuing that conversation. I just wish Leo were here to offer his rebuttal.

Acknowledgments

The author acknowledges support from the NSF grant DMS 1638521.

References

- David L Banks, Leanna House, and Kevin Killourhy. Cherry-picking for complex data: robust structure discovery. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4339–4359, 2009.
- Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.
- Andreas Buja and Yung-Seop Lee. Data mining criteria for tree-based regression and classification. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 27–36, 2001.
- Cynthia Rudin and Joanna Radin. Why are we using black box models in ai when we don’t need to? a lesson from an explainable ai competition. *Harvard Data Science Review*, 1(2), 2019.