Forthcoming, *Observational Studies*.

TECHNICAL REPORT
R-503
March 2021

Causally Colored Reflections on Leo Breiman's "Statistical Modeling: The Two Cultures" (2001)

Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles, CA 90024 USA
judea@cs.ucla.edu

March 21, 2021

Abstract

This note provides a re-assessment of Breiman's contributions to the art of statistical modeling, in light of recent advances in machine learning and causal inference. It highlights the crisp separation between the data-fitting and data-interpretation components of statistical modeling.

Keywords: Causality, Prediction, data-science, data fusion, missing data, counterfactuals

1 What Cultures Dominate Statistics?

When Breiman's paper first appeared, in 2001, I had the impression that, although the word "cause" did not appear explicitly, he was trying to distinguish data-descriptive models from models of the data-generation process, also called "causal," "substantive," "subject-matter," or "structural" models.[1] Unhappy with his preference for model-blind predictions, and his reluctance to discuss questions concerning causal effects, I was glad nevertheless that a statistician of Breiman's standing had recognized lingering cultural tensions in the field, and was calling for making distinctions crisp.

Upon re-reading the paper in 2020 I have realized that the two cultures contrasted by Breiman are not descriptive vs. causal but, rather, two styles of descriptive modeling, one interpretable, the other uninterpretable. The former is exemplified by predictive regression models, which seductively tempt researchers to assign them process interpretations, and the latter by modern big-data algorithms such as

[1] Breiman called the latter "data models" instead of "data-generation models," which might lead to some confusion with modern terminology as in (Pearl and Mackenzie, 2018).

deep-learning, CART, trees and forests. The former carries the potential of being interpreted as causal, the latter leaves no room for such interpretation; it describes the prediction process chosen by the analyst, not the data-generation process chosen by nature. Breiman's main point is: If you want prediction, do prediction for its own sake and forget about the illusion of representing nature.

Breiman's paper deserves its reputation as a forerunner of modern machine learning techniques, but falls short of telling us what we should do if we want the model to do more than just prediction, say, to extract some information about how nature works, or to guide policies and interventions, not to mention finding explanations of observed effects. For him, accurate prediction is the ultimate measure of merit for statistical models, an objective shared by present day machine learning enterprise, which also accounts for many of its limitations (Pearl, 2019).

In their comments on Breiman's paper, David Cox and Bradley Efron noticed this deficiency and wrote:

> "... fit, which is broadly related to predictive success, is not the primary basis for model choice and formal methods of model choice that take no account of the broader objectives are suspect. [The broader objectives are:] to establish data descriptions that are potentially causal." (Cox, 2001)

And Efron concurs:

> "Prediction by itself is only occasionally sufficient. ...Most statistical surveys have the identification of causal factors as their ultimate goal." (Efron, 2001)

These comments by Cox and Efron are revealing, since interest in "causal factors" has not been central to statistics research. Causal questions have in fact been off limit in mainstream statistics, informally motivating many studies but badly lacking formal representations, and rigorous inference tools.[2]

2 Two Cultures or Collaborative Symbiosis?

What Cox and Efron probably tried to convey in their comments was that implicit causal considerations, though lacking formal representation, are behind most statisticians' goals and thoughts. This profound observation is supported indeed by several recent explorations. Problems areas such as "meta-analysis" (also known as "data fusion") and "missing data," which were thought to be purely statistical in nature, turned out to be causal problems in disguise (Bareinboim and Pearl, 2016; Mohan and Pearl, 2021).

Moreover, the taxonomy of statistical and causal problems is no longer at the mercy of intuitive speculations, but has received a crisp formal definition through

[2] Statistical textbooks, for example, are still void of causal vocabulary or causal modeling. Stigler's historical account, as another example, *The Seven Pillars of Statistical Wisdom* (2016) barely makes a passing remark to two (hardly known) publications in causal analysis.

the Ladder of Causation (see (Pearl and Mackenzie, 2018)): 1. Association, 2. Intervention, 3. Counterfactuals. By examining the syntax of a model one can tell whether it can answer a given research question, and where the information supporting the answer should come from, be it observational studies, experimental data, or theoretical assumptions. In particular, one cannot answer questions at level i unless one has information of type i or higher. For example, there is no way to answer policy related questions unless one has experimental data or assumptions about such data. As another example, there is no way of answering questions at the individual level from population data unless we invoke counterfactual models (Li and Pearl, 2019; Pearl, 2015).

As we read Breiman's paper today, armed with what we know about the proper symbiosis of machine learning and causal modeling, we may say that his advocacy of algorithmic prediction was justified. Guided by a formal causal model for identification and bias reduction, the predictive component in the analysis can safely be trusted to non-interpretable algorithms. The interpretation can be accomplished separately by the causal component of the analysis, as demonstrated, for example, in (Pearl, 2019).

Separating data-fitting from interpretation, an idea that was rather innovative in 2001, has withstood the test of time.

References

Bareinboim, E. and Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences* **113** 7345–7352.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science* **16** 199–231.

Cox, D. (2001). Comment: Statistical modeling: The two cultures. *Statistical Science* **16** 216–218.

Efron, B. (2001). Comment: Statistical modeling: The two cultures. *Statistical Science* **16** 218–219.

Li, A. and Pearl, J. (2019). Unit selection based on counterfactual logic. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization.

Mohan, K. and Pearl, J. (2021). Graphical models for processing missing data. Forthcoming, *Journal of the American Statistical Association* **0** 1–16.
URL `https://ftp.cs.ucla.edu/pub/stat_ser/r473-reprint.pdf`

Pearl, J. (2015). Causes of effects and effects of causes. *Journal of Sociological Methods and Research* **44** 149–164.

PEARL, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM* **62** 54–60.
URL https://ftp.cs.ucla.edu/pub/stat_ser/r481-reprint.pdf

PEARL, J. and MACKENZIE, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York.

STIGLER, S. M. (2016). *The Seven Pillars of Statistical Wisdom*. Harvard University Press, Cambridge, MA.