PROJECT MUSE®

Comment on Breiman's "Two Cultures" (2002): From Two Cultures to Multicultural

Galit Shmueli

Observational Studies, Volume 7, Issue 1, 2021, pp. 197-201 (Article)

Published by University of Pennsylvania Press

DOI: <https://doi.org/10.1353/obs.2021.0010>

➔ *For additional information about this article*

<https://muse.jhu.edu/article/799735>

Comment on Breiman's "Two Cultures" (2002): From Two Cultures to Multicultural

Galit Shmueli

galit.shmueli@iss.nthu.edu.tw

Institute of Service Science

National Tsing Hua University

Hsinchu 30013, Taiwan

Abstract

Since Breiman's "Two Cultures" paper's appearance in 2002, the term *prediction* has gained incredible significance in research, practice, society, and humanity. "Two Cultures" led to many useful advancements and surprising discoveries. Experiencing first hand the different cultures in the statistics and machine learning communities that Breiman expressed so early and clearly, I've then encountered even more differences. I describe additional modeling distinctions and further modeling "cultures". Recognizing these cultures, understanding their reasoning, and comparing and contrasting them, opens our eyes to new ways of viewing the world and creates opportunities for innovation and collaboration.

Keywords: prediction, causal explanation, description

First Encounter

When Breiman's "Two Cultures" paper came out, I was Visiting Assistant Professor at the Statistics department at Carnegie Mellon University (CMU). I recall Larry Wasserman organizing an almost urgent seminar meeting to discuss the paper. As a statistician with a background in industrial statistics and very little experience in prediction, I was not sure I understood what was going on during the seminar, but I clearly felt the excitement in the room. Little did I know how critical the "two cultures" would be to my own research.

From Description to Prediction

In the early 2000s, I was involved in a large biosurveillance collaboration between CMU and University of Pittsburgh, focused on early disease detection — my first predictive endeavor. We were studying the possibilities of using wavelets for predicting non-well-behaved time series (see Goldenberg et al., 2002), and an interesting challenge arose: while wavelets were terrific for capturing the patterns in a time series (a descriptive task), their ill fit at the edges of a time series was problematic when the goal was extrapolating into the future. Prediction was appearing to be a different creature. As I moved to University of Maryland's business school in 2002, and started teaching and writing a textbook on data mining for MBA-level courses, I encountered strange discrepancies between how linear regression was presented in statistics textbooks and taught in statistics courses, versus how it is used for prediction. In working out to reconcile the differences between the data mining approach (prediction) and

the statistics approach (description, estimation), I discovered that the distinction Breiman made was not only profound, but also an entry to an even deeper tunnel: the difference between statistical modeling for prediction and causal explanation.

From Estimation to Causal Explanation

Collaborating with researchers in social science, behavioral sciences and econometrics – in my case, information systems researchers – exposes a statistician to a different world in which statistical modeling is used first and foremost for testing causal hypotheses. The mantra “correlation is not causation” chanted in statistics circles is replaced with theoretical argumentation that relies on domain knowledge and causal arguments. Such collaborations have led me to discover startling differences between the process of statistical modeling as performed by social scientists asking causal questions, statisticians asking descriptive questions, and machine learning researchers focusing on generating predictions. I also discovered that econometricians and social scientists have their own language and processes for “data analysis”, which rely on different ways of modeling the world. In place of data originating from “generating processes” they have constructs, causal structures, and underlying theory that are tied to data through careful operationalization. Papers that rely on regression models often intentionally retain statistically insignificant covariates. Interaction terms are called moderators, and covariates have different types and roles, such as “mediators” and “control variables”.

Breiman’s paper brilliantly looked at two different fields that claim to be experts of “data modeling”, and discovered a large gap in how they go about this task. He explains how he obtained this birds-eye view by working in consulting prior to becoming an academic. Rather than condemn one community for “not properly using statistics” or defying a mantra, I believe he came with an open mind, asking “why are they doing things differently?”. That was my own approach when trying to understand my social-science-trained or econometrics-trained colleagues, whose work was thrilling and interesting, yet statistically perplexing.

To Explain or To Predict? Two Cultures and Two Languages

After several years of investigating these statistical modeling differences by engaging with scholars in various disciplines, and giving presentations that were received with a mix of nods of approval and furious disagreement, I finally wrote “To Explain or To Predict?” (Shmueli et al., 2010). The paper expands Breiman’s point about “algorithmic modeling” vs. “data modeling” into a comprehensive discussion of the entire data analysis process, from study design to deployment, and going back to the philosophy of science debates on prediction vs. understanding. The paper identifies the sources of differences between statistical modeling for causal explanation vs. prediction, how this translates into different tradeoffs on several dimensions (causation-association, theory-data, retrospective-prospective, bias-variance, and most recently I’ve added average-individual observation), and how these differences translate into different considerations, choices, and actions in each step of the data analysis process.

A key complaint from a reviewer on my initial submission was — shocking at first, but then discovered as enlightening — the lack of notation. Breiman used a few diagrams in

his paper, trying to explain visually what is difficult to write in equations and statistical terminology. I resorted to using two sets of fonts, one to denote theoretical constructs and theoretical causal relationships, and the other denoting (the standard) random variables, their realizations, and associations. The lack of vocabulary for describing our different representations of data, operations, models, processes, and theories, is, in my view, a serious blind spot of our field.

Another example of lack of sufficient vocabulary is in causal analysis. The recent introduction of Judea Pearl’s *do(.)* operator and grammar (e.g., in Pearl, 2009) highlights the lack of notation for representing causal relationships. Similarly, Rubin’s potential outcomes framework and counterfactual causal analysis (e.g. Rubin, 2005) uses unusual notation, with either superscripts or parenthetical arguments to convey quantities that do not exist in other statistical domains.

From Two Cultures to Multicultural

Once the differences between prediction and causal explanation are clarified, a thrilling opportunity arises: might we be able to advance each of these different “cultures” by borrowing and adapting from each other?

For example, how can prediction be used for theory building? Parzen (2001) commented on Breiman’s paper:

The two goals in analyzing data which Leo calls prediction and information I prefer to describe as “management” and “science”. Management seeks profit... Science seeks truth.

But is this distinction useful or true? Can theory help seeking profit and might prediction help seeking truth?

Yu and Kumbier (2020) recently developed a workflow for “veridical data science” that “uses predictability as a reality check”. In a recent comment on Efron’s paper “Prediction, Estimation, Attribution” (Efron, 2020a,b), Friedman et al. (2020) write: “we do not see any fundamental tension between prediction and estimation/attribution... We agree that interpretable prediction models may sometimes be helpful in getting closer to the ground truth”. They further write,

modern prediction algorithms go hand-in-glove with the more traditional estimation models... If you fit a boosting model, and it explains 70% of the (test-set) variance compared to the 30% explained by your parametric model, that suggests you may have left out some variables, some interaction, and possibly need some nonlinearities.

These two points touch on the role of prediction in the scientific endeavor. But there is even more: In Shmueli and Koppius (2011), we asked “how can prediction support theory development?” While it was clear that prediction was useful in practice, social science researchers want to know how prediction might support scientific development. We identified six roles of predictive modeling in scientific scholarship: (1) generating new theory, (2) developing measures, (3) comparing competing theories, (4) improving existing models, (5) assessing relevance, and (6) assessing predictability.

Going Forward by Crossing Borders

Leo Breiman’s “two cultures” paper opened not only an important conversation within the statistics community, but has also set up the basis for an important longer term research project: pointing out the distinction allows us to study the differences and their sources and use them to innovate. New questions arise: can explanatory modeling approaches be useful for advancing prediction? Can predictive algorithms and approaches improve descriptive modeling and causal explanation?

Breiman introduces several examples, showing that tree-based algorithms are sometimes more powerful than statistical models such as regression models. His examples show the ability of trees to extract more information in high-dimensional data (e.g. via random forests), and in some forms provide interpretable models that are user-understandable. In my own work, I’ve found trees to have enormous potential beyond their use for descriptive and predictive modeling: namely, for causal explanation problems using high-dimensional large datasets. For example, in Yahav et al. (2016) we introduce the use of trees as an alternative to propensity score matching (PSM) for addressing self-selection in large-scale impact studies. Trees not only reduce data dredging by avoiding arbitrary user choices of matching parameters, but are also able to point out which variables seem unbalanced, as well as automatically produce local data subsets that are matched on the unbalanced variables. In a different work, we used the sequence of the tree splitting in order to identify potential Simpson’s Paradoxes in large, rich datasets (Shmueli and Yahav, 2018). Several econometricians have also recently started developing specialized trees for causal modeling, such as causal trees for estimating heterogeneity in causal effects (Athey and Imbens, 2016)), or for generating instrumental variables (Yang et al., 2019). While today’s buzzword is Deep Learning, I believe trees still hold further treasures that can be useful in solving new types of problems beyond prediction.

In the opposite direction, there is a host of statistical models used exclusively for estimation and inference, such as path models (also known as “structural equations models” or SEM), which are heavily used in survey studies in marketing, information systems, and other fields. These models tie constructs to observable measurements, and are used for testing causal hypotheses. It is unclear, however, how they might generate predictions. Yet the scientific method requires predictions to test the falsifiability of a model. Might path models benefit from integrating predictive approaches? Towards that end, we first investigated what types of predictions can be generated with path models, and which of those allow evaluating predictive power (Shmueli et al., 2016). There has since been much work exploring and adapting predictive thinking into path models, especially in the marketing and management literature, but there is plenty of room for further development. It would also be useful to identify and study other classes of statistical models used exclusively for non-predictive goals in terms of integrating predictive approaches.

To conclude, I’m grateful for encountering Leo Breiman’s “Two Cultures” paper relatively early in my academic career. The paper has led to many useful advancements and surprising discoveries, and I am certain there are plenty more to come.

References

- Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.
- Bradley Efron. Prediction, estimation, and attribution. *Journal of the American Statistical Association*, 115(530):636–655, 2020a.
- Bradley Efron. Prediction, estimation, and attribution. *International Statistical Review*, 88:S28–S59, 2020b.
- Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Discussion of “prediction, estimation, and attribution” by bradley efron. *Journal of the American Statistical Association*, 115(530):665–666, 2020.
- Anna Goldenberg, Galit Shmueli, Richard A Caruana, and Stephen E Fienberg. Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. *Proceedings of the National Academy of Sciences*, 99(8):5237–5240, 2002.
- Emanuel Parzen. Comment-statistical modeling: The two cultures. *Statistical Science*, 16(3):224–225, 2001.
- Judea Pearl. *Causality*. Cambridge university press, 2009.
- Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.
- Galit Shmueli and Otto R Koppius. Predictive analytics in information systems research. *MIS quarterly*, pages 553–572, 2011.
- Galit Shmueli and Inbal Yahav. The forest or the trees? tackling simpson’s paradox with classification trees. *Production and Operations Management*, 27(4):696–716, 2018.
- Galit Shmueli, Soumya Ray, Juan Manuel Velasquez Estrada, and Suneel Babu Chatla. The elephant in the room: Predictive performance of pls models. *Journal of Business Research*, 69(10):4552–4564, 2016.
- Galit Shmueli et al. To explain or to predict? *Statistical science*, 25(3):289–310, 2010.
- Inbal Yahav, Galit Shmueli, and Deepa Mani. A tree-based approach for addressing self-selection in impact studies with big data. *Management Information Systems Quarterly*, 40(4):819–848, 2016.
- Mochen Yang, Edward McFowland, Gordon Burtch, and Gediminas Adomavicius. Achieving reliable causal inference with data-mined variables: A random forest approach to the measurement error problem. Technical Report 19-20, Kelley School of Business Research Paper, 2019.
- Bin Yu and Karl Kumbier. Veridical data science. *Proceedings of the National Academy of Sciences*, 117(8):3920–3929, 2020.