Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

The Data Science Process: One Culture

Bin Yu & Rebecca Barter

To cite this article: Bin Yu & Rebecca Barter (2020) The Data Science Process: One Culture, Journal of the American Statistical Association, 115:530, 672-674, DOI: 10.1080/01621459.2020.1762615

To link to this article: https://doi.org/10.1080/01621459.2020.1762615

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 115, NO. 530, 672–674: Comment
https://doi.org/10.1080/01621459.2020.1762615

The Data Science Process: One Culture

Bin Yu[a,b,c] and Rebecca Barter[a]

[a]Statistics Department, University of California Berkeley, Berkeley, CA; [b]EECS Department, University of California Berkeley, Berkeley, CA; [c]Chan Zuckerberg Biohub, San Francisco, CA

We would like to congratulate Professor Brad Efron for the well-deserved honor of winning the 2019 International Statistics Prize, and for his thought-provoking paper discussing the diverging paths of so-called traditional statistical regression and pure prediction approaches. Professor Efron has made vast contributions to the field of statistics spanning several decades, and we enjoyed reading his take on the current state of the fields of statistics and machine learning.

In this discussion, we provide a portrayal of our own perspectives on the intersection of statistics and machine learning methods, and compare them with Professor Efron's. While we came from very similar beginnings as Professor Efron, rather than arriving at his view of a "discrepancy" or a "radical difference" between "traditional statistical regression methods" and "pure prediction" methods, we arrived at a surprisingly contrasting view that unifies these two approaches. At the very end of his paper, Professor Efron seems "hopeful" about the "reunification" of these two paradigms. However, in his view, efforts toward this reunification are "just underway," with the primary impediments being a lack of theory. In our contrasting view, we are much further along the path of reunification, with the theoretical underpinnings being less critical than—but still important, and following on from—empirical evidence in today's reality-rooted era.

While our views are quite different, we certainly agree with many of the points that Professor Efron makes in his paper. For instance, Professor Efron appears weary of relying solely on *random* splits when implementing the train/test paradigm that is pervasive in the world of pure prediction algorithms. Specifically, he points out that there are instances where instead of providing an honest estimation of the error, a random split will be far too optimistic, as in the case of time-dependent data. He argues that a more realistic split in such scenarios is not random, but rather uses an early/late split where the early data are used to train and the later data are used to test. Indeed, a random split in many cases ensures that the training and test sets are identically distributed, but fails to ensure that they are independent (i.e., a random split does not resemble two independent samples taken from the same population). Such a random split certainly does not resemble the relationship between the data used to train an algorithm and the future data that it will be used on (future data are, at best, an independent sample from the same population, but are often collected from a different source or under different circumstances). While Professor Efron uses this example to focus on how prediction is easier than extrapolation or interpolation (based on the observation that a random split yields better test-set performance than the early/late split in time-dependent data), we view this as an example of the importance of critical thinking. In our view, the train/test paradigm should not focus on randomness, it should instead focus on representing the way the predictive algorithm will be used. The training set should be representative of the current data being used to train the algorithm, and the test set should be representative of the future data to which the algorithm will be applied. So while we agree with Professor Efron's point that a random split is not always appropriate, we argue that it is not the *randomness* of the split that is critical to the pure prediction paradigm, but the *representativeness* of the split (relative to the prediction problem).

We were particularly delighted to read Professor Efron's reflections on Leo Breiman's 2001 two cultures paper, including his own discussion and perspectives at the time, along with how they have been updated in the two decades since. Interestingly, Leo's paper, and author Yu's own interactions with him, were the starting point for her own foray into machine learning 20 years ago. The lessons learned and perspectives gained from this journey have become a focal part of Yu's own research and philosophy that she has both passed on to—and developed together with—her students (including coauthor Barter). Today looking back on Brieman's work, the main difference between our view and Professor Efron's is that his view focuses on the differences between machine learning (or "pure prediction") methods and "traditional regression" methods, whereas our own view focuses on the connections that bridge these two cultures, and the expansions building on both.

Professor Efron highlights his perspective on the differences between "pure prediction" and "traditional regression methods" via six criteria presented in Table 5. For instance, he views traditional regression models as focusing on scientific truth with pure prediction methods focusing instead on prediction accuracy; he views traditional regression models as focusing on low dimension problems while pure prediction methods focus on high dimensional problems; and he views traditional regression problems as utilizing optimal inference theory such as maximum likelihood and Neyman–Pearson while pure prediction methods focus instead on training/test paradigms.

CONTACT Bin Yu *binyu@berkeley.edu* Statistics Department, University of California Berkeley, CA 94720.

However, in our experience, pure prediction methods are routinely used in practice *together* with classical statistical methods such as PCA and logistic regression. Right or wrong, logistic regression is often even categorized as a machine learning (pure prediction) method, and many modern machine learning books include traditional statistical ideas such as maximum likelihood, linear regression, and logistic regression. We have experienced the use of both traditional regression and pure prediction approaches in low- and high-dimensional problems. The idea that traditional regression methods focus on scientific truth while pure prediction approaches focus on prediction accuracy certainly has some merit, but in our experience the "truth" that traditional regression methods supposedly represent is rarely justified or validated. The mathematical formulae on which traditional regression models are built are rarely based on verifiable domain knowledge. Instead, the primary method for which people find evidence that these mathematical regression formulations represent the truth is often measured by their *predictive accuracy* (which, in Professor Efron's paper, is the primary concern of pure prediction methods, rather than traditional regression models). At the very least, if traditional regression methods do not generate accurate predictions, then it is unlikely that it is adequately capturing any kind of underlying "truth." Moreover, since a reasonable goal is to ensure that conclusions drawn (and predictions generated) are generalizable to new data, the training/test paradigm is just as relevant for traditional regression approaches as it is for the pure prediction approaches.

Professor Efron separates regression approaches into three categories: prediction ("the prediction of new cases"), estimation ("an instrument for peering through the noisy data and discerning a smooth underlying truth"), and attribution ("the assignment of significance to individual predictors").

Professor Efron argues that pure prediction algorithms tend to "focus on prediction, to the neglect of estimation and attribution." However, while traditional regression methods generate attribution via p-values associated with model coefficients, many pure prediction methods, such as random forest, have their own methods of attribution in the form of variable importance (gini or permutation) scores. Professor Efron discusses this idea but dismisses these modern attribution attempts as being suboptimal, dampened by the pure prediction method's focus on prediction using many weak learners that do not prioritize individual strong predictors. Meanwhile, in our own research group we have seen great success using these importance scores to understand attribution. We developed the iterative random forest (iRF) (Basu et al. 2018) approach for identifying higher-order interactions based on decision paths in (pure prediction) random forest models, and have found such approaches to be extremely useful in practical applications including verifying existing and identifying novel gene interactions in drosophila embryonic development. Our experience indicates that prediction and attribution can go hand-in-hand, rather than being at odds with one another. In addition, Professor Efron's focus on estimation implies that there is an underlying "truth" that can be uncovered by a model. As we discussed above, our view has become that no such truth is ever really attainable, whether or not you are working within the realm of traditional regression models or pure prediction models. Our own perspective has transitioned into a framework of accurate *approximation* for a particular domain, and relative to a particular performance metric, rather than *truth-seeking* as was the goal of traditional statistical inference.

Finally, Professor Efron calls for "substantial further development" of machine learning approaches before they are ready for "routine scientific applicability." However, such approaches are already being used successfully in routine scientific applicability. Of course we also agree that further development is always a good thing. In our research group, we have been working toward this goal, building on our own experience gained from many interdisciplinary projects applying pure prediction methods in conjunction with traditional statistical ideas in neuroscience, genomics and precision medicine. As a result of our work, we have developed a framework that we call *veridical* ("truthful") *data science*, which encompasses a unified view of pure prediction and traditional regression approaches under the three principles of predictability, computability, and stability (PCS). The idea of using prediction as a reality check forms the basis of our predictability principle, and ideas from the traditional regression approaches are modernized in the stability principle, which itself bears a strong relationship to and expands on Professor Efron's groundbreaking work on the bootstrap. Stability is a critical component of all stages of a data science project, ranging from the literal words used to formulate the problem (every word in the problem formulation should mean the same thing—have stable meaning—to every individual on a multidisciplinary project team) to data cleaning, and data and model perturbations in the modeling stage. Data-driven approaches are brought into the modern big data era via the computability principle. Our "one culture" veridical PCS data science framework highlights the practice of veridical ("truthful") data science for reliable, reproducible and transparent data-driven decision making and knowledge generation from data for a particular domain problem (Yu and Kumbier 2020). In addition, our paper also introduces and advocates for PCS documentation which records the analytical judgement calls made by humans throughout the data science life cycle, and records relevant domain knowledge and code. Our paper provides a nice summary of our views as this one does for Professor Efron's. The two papers would be an interesting side-by-side read.

Professor Efron's piece gave us much food for thought, allowing us the opportunity to reflect on changing perspectives over time. We appreciate the invitation to provide a discussion for this piece, and we remain ever in debt to the truly innovative contributions that Professor Efron has made to the field of statistics that have now also made their way into the field of machine learning, and in our view, contributed to the unification (rather than the division) of these two fields to arrive at "one culture" in today's field of data science.

Funding

We would like to acknowledge and thank the following funding sources: Office of Naval Research grant N00014-16-1-2664, NSF grants DMS-1613002 and IIS 1741340, and the Center for Science of Information, a US NSF Science and Technology Center, under grant CCF-0939370. B.Y. is a Chan Zuckerberg Biohub investigator.

References

Basu, S., Kumbier, K., Brown, J. B., and Yu, B. (2018), "Iterative Random Forests to Discover Predictive and Stable High-Order Interactions," *Proceedings of the National Academy of Sciences of the United States of America*, 115, 1943–1948. [673]

Yu, B., and Kumbier, K. (2020), "Veridical Data Science," *Proceedings of the National Academy of Sciences of the United States of America*, 117, 3920–3929. [673]