Bridging Breiman's Brook: From Algorithmic Modeling to Statistical Learning

Giles Hooker, Lucas Mentch
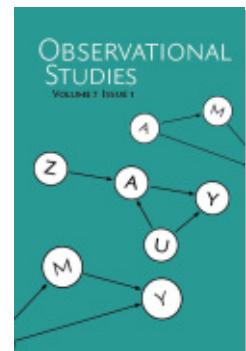
Bridging Breiman's Brook:
From Algorithmic Modeling to Statistical Learning

Giles Hooker   gjh27@cornell.edu

Department of Statistics and Data Science
Cornell University
Ithaca, NY 14850 USA

Lucas Mentch   lkm31@pitt.edu

Department of Statistics
University of Pittsburgh
Pittsburgh, PA 15260 USA

Abstract

In 2001, Leo Breiman wrote of a divide between "data modeling" and "algorithmic modeling" cultures. Twenty years later this division feels far more ephemeral, both in terms of assigning individuals to camps, and in terms of intellectual boundaries. We argue that this is largely due to the "data modelers" incorporating algorithmic methods into their toolbox, particularly driven by recent developments in the statistical understanding of Breiman's own Random Forest methods. While this can be simplistically described as "Breiman won", these same developments also expose the limitations of the prediction-first philosophy that he espoused, making careful statistical analysis all the more important. This paper outlines these exciting recent developments in the random forest literature which, in our view, occurred as a result of a necessary blending of the two ways of thinking Breiman originally described. We also ask what areas statistics and statisticians might currently overlook.

Keywords: Random Forests, Data Science, Machine Learning

1. Introduction

Twenty years after its initial publication, Breiman's "Two Cultures" essay (Breiman et al., 2001) remains an entertaining and enlightening read, though we are struck by how uncontroversial new readers in the modern era may likely find many of its prescriptions. We begin our commentary by comparing Breiman's views at the time with current perspectives on data science and argue that while the field of statistics may have fallen short on some fronts, Breiman would no doubt be pleased with a number of strides made in recent decades. In our view, these developments have been made not because of a switch in statistical thinking from data to algorithmic modeling, but because rigorous statistical analysis has allowed these methods to be understood within more traditional statistical frameworks. We detail how the willingness to expand the statistical toolbox without sacrificing fundamental principles has led to a rapid development of important recent studies on random forests and urge the statistical community to continue these efforts to better understand cutting-edge methodologies in deep learning. Throughout the remainder of this paper, quotes and refer-

ences including only a page number are in reference to Breiman's original essay (Breiman et al., 2001).

2. Data Science: Breiman's Final Prediction

"If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools." – Page 199

The above quote that appeared in the abstract of Breiman's essay provides a reasonably accurate summary of thoughts that he expounds upon in the pages that follow. To most modern data analysts, however, the aim of "using data to solve problems" almost certainly feels like the credo of what we would now refer to as 'data science', of which the field of statistics constitutes only a proper subset. In the decades following the publication of Breiman's essay, the interest in solving problems via data analyses has continued to explode along with the size and scale of the data available. Predictably, this has led to the current climate in which researchers from across at least dozens of academic fields might rightfully claim to be working on projects intersecting with data science.

In our view, however, the field of statistics has seen substantial growth in its own right. Machine learning, a field described by Breiman in 2001 as having "occurred largely outside statistics ... mostly [by] young computer scientists" (p. 200) is now a standard research interest listed on the majority of CVs by young statisticians on the faculty job market. We note that while this interest sometimes includes topics quite far from what would have been described as Machine Learning in 2001, the eagerness on the part of statisticians to plant their flag in the field is a promising development in the eyes of those sympathetic to Breiman's arguments for a more intense focus on algorithmic modeling . Leading statistics journals like The Annals of Statistics, Biometrika, The Journal of the American Statistical Association (JASA), and The Journal of the Royal Statistical Society (JRSS), have shown a willingness, if not an enthusiasm, to publish advances in machine learning research so long as the core contributions are fundamentally grounded in statistics. At the same time, the Journal of Machine Learning Research (JMLR) has been established that regularly contains individuals on the editorial staff who are unmistakable statisticians, providing researchers a another top-tier journal venue. Measures of model accuracy as measured via test or cross-validation error – an aspect Breiman felt was distinctly lacking (p. 204) – are now commonplace in modern statistical publications.

Individuals and departments have also been outspoken in embracing the data science revolution. Bin Yu, in her 2014 IMS Presidential address, urged statisticians to proudly proclaim themselves data scientists, arguing that while statistics may not *be* data science, no one does more of the job (Yu, 2014). Top statistics departments in the United States including, for example, Carnegie Mellon, Yale, and Cornell University, now include 'Data Science' in the official department name. Thus, while the field of statistics has perhaps not expanded quickly enough in the minds of some, we think that Breiman would be very pleased with the willingness to embrace new directions exemplified in recent decades.

3. Merging Data and Algorithmic Modeling

"There is an old saying "If all a man has is a hammer, then every problem looks like a nail." The trouble for statisticians is that recently some of the problems have stopped looking like nails." – Page 204

The foundation of Breiman's arguments lie in the distinction between what he refers to as "data modeling" as opposed to "algorithmic modeling". Data modeling, according to Breiman, is the traditional statistical modeling context wherein a basic framework like $Y = f(X) + \epsilon$ is assumed and simple parametric models like linear regression, logistic regression, or the Cox model are utilized to estimate f. The algorithmic modeling culture, by contrast, utilizes tools like decision trees and neural networks and *"considers the inside of the box complex and unknown. Their approach is to find a function $f(x)$ – an algorithm that operates on x to predict the responses y"* (p. 199). *"The commitment to data modeling,"* Breiman argues, *"has prevented statisticians from entering new scientific and commercial fields where the data being gathered is not suitable for analysis by data models"* (p. 200). *"The best available solution to a data problem might be a data model; then again it might be an algorithmic model"* (p. 204).

But while the distinction between data and algorithmic models may have seemed natural at the time, 20 years later, it feels arbitrary and unnecessary: $Y = f(X) + \epsilon$ *is* a data model! Throughout the essay, Breiman seems to repeatedly conflate "data models" and "parametric models". It's not clear why Breiman seemed to feel that tools like decision trees cannot or should not be implemented in conjunction with at least some minor model assumptions, and his reference to Grace Wahba's foundational, and very statistical, work on nonparametric smoothing as outside of mainstream statistics, at least in 2001, seems strange.

Breiman himself goes so far as to admit that one the primary goals in statistics is to obtain information about the underlying data generating mechanism (p. 203, 214). And insofar as this goal is concerned, it's important to stress that neither the data or algorithmic modeling approach is of any value when carried out at the extreme. A data model, though perhaps interpretable, holds no value if the model itself is not reflective of the process that generated the original data. On the other hand, an algorithmic model that amounts to an impenetrable black-box, while it may produce accurate predictions, is merely a potentially more efficient version of nature and therefore of limited scientific value – we haven't learned anything about the mechanisms and variables driving those predictions.

In our view, the enormous progress that has been made by statisticians in recent decades came not from abandoning the traditional principals of data modeling, but by merging the cultures Breiman describes. The willingness on the part of statisticians to expand the traditional modeling toolbox without sacrificing core data analytic and uncertainty quantification principles has led to tremendous gains in our knowledge and understanding of modern supervised learning techniques. Random forests – the apple of Breiman's eye – is certainly one of the methodologies that has benefitted most from this merged mindset.

4. Advancements with Random Forests

'*So forests are A+ predictors. But their mechanism for producing a prediction is difficult to understand. … Using complex predictors may be unpleasant, but the soundest path is to go for predictive accuracy first, then try to understand why.*" – Page 208

The second sentence in the above quote is perhaps the best summary of the later half of Breiman's essay and arguably the fundamental principle that has stimulated the recent progress in understanding the success and inner workings of black-box learning algorithms like random forests. At the time, much of the traditional statistical modeling culture revolved around formulating parametric models and making continual tweaks until the model seemed to fit best. The process begins with an explicit and interpretable model and modifications are made to try and improve accuracy. Breiman – stemming from his belief that "*…the models that best emulate nature in terms of predictive accuracy are also the most complex and inscrutable.*" (p. 208) – urged statisticians to consider the process in reverse: begin with a an accurate but potentially obscure model and perform follow-up investigation to uncover the impact of individual predictors on the response.

Breiman kickstarts these efforts in his work with random forests and out-of-bag (oob) variable importance, demonstrating that "*a model does not have to be simple to provide reliable information about the relation between predictor and response variables*" (pp. 209-211). Interestingly, in doing so, he himself implicitly begins to utilize traditional data modeling thinking in order to extract insight from complex models. In the following sections, we detail how similar thinking has recently led to a rapid advancement in our understanding of random forests as well as the ways in which they can be reliably used to further our understanding of the underlying processes.

4.1 Statistical Understanding of Random Forests

"*My kingdom for some good theory.*" – Page 205

Breiman's preference for algorithmic models did not lead him to abandon statistical theory; indeed his essay explicitly recognizes the theoretical developments underpinning both smoothing splines and support vector machines (although it unaccountably ignores a large statistical literature on other nonparametric smoothing methods) and he sketched such an analysis of random forests. The two decades since then have seen statistics slowly fill in theoretical details particularly about the random forest method. It's unclear whether Breiman's essay, or their reliable "off-the-shelf" performance (or a combination of the two) that motivated a statistical focus on random forests in particular, but our own interest in these methods was also influenced by recognizing an appealing macroscopic structure that could be used to develop mathematical theory.

The pace of these developments at times felt glacial, due somewhat to the difficulty of developing an analysis that accounts for the greedy splitting strategies involved in building trees. Theoretical headway has come from making a number of modifications, some arguably beneficial, others purely focussed on mathematical tractability:

1. Methods that provide guarantees on the (geometic) size of the leaves of individual trees, usually obtained by introducing randomly chosen splits. Originally, this in-

volved generating tree structures largely at random (Biau et al., 2008; Biau and Devroye, 2010; Biau, 2012). More recently, the *regularity* conditions of Wager and Athey (2018) can be enforced by choosing a split at random with some probability while otherwise allowing data-informed splitting methods to be used. Controlling leaf dimensions provides a similar control over within-leaf bias, where the tree represents the underlying relationship as constant and hence forms an important component of consistency results.

2. Methods that enforce independence between tree structures and leaf values. We think of this as avoiding the problem of trees "chasing order statistics" or finding structures that isolate exceptionally (and randomly) large values of the response. Choosing tree structures completely at random and letting the data dictate leaf values naturally provides this and facilitates a representation of random forests as kernel methods (Scornet, 2016). Wager and Athey (2018) propose splitting the data between samples used to determine tree structure and those assigning values to leaves. The theoretical advantage of this separation is that it allows the values given to leaves to be unbiased for the average response in the leaf. This completes a control of bias when combined with controlling the size of the leaves. The authors also argue for performance improvements in addition to theoretical guarantees based on reducing edge bias. However, this improvement is counteracted by reducing the size of the sample available for building trees (and hence their depth) and we thus regard that trade-off as unclear. The standard CART split criterion (Breiman et al., 1984) seems to remain the overwhelmingly most popular choice in practice.

3. Changing the sampling structure. Breiman's original random forest algorithm obtained trees from bootstrap samples of the data. The last five years has seen a switch to employing subsamples instead. This is motivated by the connection between random forests based on subsamples and classical U-statistic constructions and allowed both Mentch and Hooker (2016) and Wager and Athey (2018) to develop central limit theorems for the predictions that random forests give. These have been refined by analyses in Peng et al. (2019) who also provided Berry-Esseen results and Zhou et al. (2019) who extended the analysis to subsampling *with* replacement and thereby obtained improved variance estimates. Scornet et al. (2015) also utilize subsampling to obtain what is arguably the "best" consistency result to date, in that it covers estimators very closely resembling the form of random forests originally proposed by Breiman (2001).

The switch to subsampling can be regarded as assisting Breiman's objective of reducing correlation between trees – we'd like subsamples large enough that trees can be built to reasonable depths (to improve tree-level accuracy) but small enough that between-tree correlation is mitigated. While this creates another tuning parameter, Zhou et al. (2019) found that using subsamples with replacement also reduced the sensitivity of results to this fraction, and that the optimal subsample size is not always that of the training data.

We note that, similar to more classical non-parametric smoothing results, these central limit theorems are not necessarily centered at a target $f(X) = E(Y|X)$. The U-

statistic argument applies quite generally to averages of any sort of learner, but specific analysis is required to make $E\hat{f}(X) - f(X)$ small enough to be ignorable for inference (e.g. Wager and Athey, 2018). In nonparametric smoothing, removing the bias from a central limit theorem generally involves under-smoothing relative to the optimal predictive model (Eubank, 1999). While we readily acknowledge the value in being able to perform strictly formal inference (see below for examples) we also believe that an analysis of stability has value in itself.

These developments allow us to provide results that state, for a given point of interest x,

$$\zeta_n(x)^{-1}(\hat{f}_n(x) - \eta_n(x)) \sim N(0,1) \tag{1}$$

where particular expressions for the mean and standard deviation functions $\eta_n(x)$ and $\zeta_n(x)$ depend on particulars of the modification of the random forest algorithm. However, the general framework creates the basis for the formalized inference we describe below.

None of the above modifications are strictly necessary in the sense that (1) is provably false without them. As noted above, Scornet et al. (2015) have produced a consistency result for estimators nearly identical to the original random forest algorithm. Such minor modifications have largely been motivated as a means to mathematical tractability, although in some cases they may also provide genuine improvements in predictive or inferential performance. These ideas have also been accompanied with a wide range of modifications to random forests that repurpose them for estimating individual treatment effects (Wager and Athey, 2018), survival analysis (Gordon and Olshen, 1985; Segal, 1988; Davis and Anderson, 1989; Hothorn et al., 2004; Zhu and Kosorok, 2012; Steingrimsson et al., 2016), quantile regression (Meinshausen, 2006) and multivariate outcomes (Segal and Xiao, 2011), with similar analyses to those we outline above. Breiman's ideas about using co-in-leaf status as a measurement of proximity between points have also been taken up to provide local models (Athey et al., 2019), prototypes (Tan et al., 2020) and prediction intervals (Lu and Hardin, 2019); see also Zhang et al. (2019) for prediction intervals based on conformal methods.

These results have all been focussed around random forest-like methods, but they are hardly the only machine learning methods available. Though not shy about promoting his own work, Breiman's essay specifically discusses boosting (Freund et al., 1996; Freund and Schapire, 1997; Friedman, 2001) as a rival methodology with real potential. These methods received early attention from analogies to the LASSO in linear models (Friedman, 2001) and initial analysis in Bühlmann (2002). Boosting is able to accommodate a wide array of model forms and restrictions (Breiman's comments suggest he might not approve of the latter), some of which are explored in Hothorn et al. (2010); Lou et al. (2013); Zhou and Hooker (2018). Nonetheless, theoretical analysis remains rudimentary and a long way from (1). Early analyses in this direction in Zhou and Hooker (2018) try to make boosting look more like random forests and may provide a starting point; Bayesian methods (Chipman et al., 2010; Ročková and Saha, 2019; Hill, 2011) have also seen significant development.

4.2 Real-world Understanding from Random Forests

"Framing the question as the choice between accuracy and interpretability is an incorrect interpretation of what the goal of a statistical analysis is. The point of a model is to get useful information about the relation between the response and predictor variables. Interpretability

is a way of getting information. But a model does not have to be simple to provide reliable information about the relation between predictor and response variables; neither does it have to be a data model." – page 209

While statistical understanding of the properties of random forests has focused on analyzing models that are increasingly similar to Breiman's original ideas, this has not been the case for the tools that he advocates for deriving knowledge from them. He acknowledges that the tools he advocates result in algebraically-complex models that do not allow human inspection, but proposes that we should first find a model that predicts well and then reverse-engineer what it does. This proposal has been directly attacked by some (e.g. Rudin (2018)), but the specifics of Breiman's approaches have also been shown to be potentially misleading. In fact, identifying, or testing, the important covariates in a random forest is a challenging procedure with many popular methods exhibiting distinct flaws. Here we review some of these and point to ways in which variable importance can be validly assessed, and point to how improved statistical understandings of random forests have contributed to these recommendations.

Breiman's original proposals for eliciting information about the relation between predictor and response variables was via his out-of-bag variable importance measures. These proceed on a tree-by-tree basis:

1. Choose those data points that were not included in the sample that the tree constructed. Call this data set Z.

2. Form a new data set Z^π by permuting the values of the variable of interest (X_1 throughout the discussion below, with $X = (X_1, X_{-1})$).

3. Compare the predictive accuracy of using the current tree to predict the response in Z to that when using Z^π.

The idea of permuting features can be applied generally, but specifically permuting out-of-bag data does, by definition, require an ensemble structure constructed via bootstrapping or at least some form of resampling. It also means we measure accuracy of individual members of the ensemble, not the ensemble as a whole. This general approach was critiqued by numerous studies including both Strobl et al. (2008) and Hooker (2007) for also breaking any relationships between X_1 and X_{-1}. In particular it may lead to values in Z^π that are far from any observed data points, thereby evaluating the ways in which a random forest extrapolates at least as much as the signal that it picks up. Despite these early warnings, the method proved (and unfortunately remains) popular as a variable screening tool; a stereotypical example can be found in Díaz-Uriarte and De Andres (2006) where features measured as least important are successively removed. This prompted us to repeat and formalize the critique with further studies in Hooker and Mentch (2019). In those studies, as in many of the earlier works, we found that pairs of correlated features were likely to have their importance estimates inflated, despite having an impact that would conventionally be regarded as smaller. We stand by our recommendation not to use these methods.

The other common measure of variable importance among random forest implementations is specific to tree structures and measures the improvement in accuracy that is obtained each time the tree is spit (e.g. Friedman, 2001); a variable accumulates the improvements

113

associated with those splits that use it. This has been critiqued for favoring variables with more potential split points in Strobl et al. (2007)[1] and we admit to having been surprised by the size of the effect found in simulation in Zhou and Hooker (2019). This, along with other papers (Li et al., 2019; Loecher, 2020) have attempted to correct for this bias by using out-of-sample data, but the precise interpretation of this measure is also not clear outside of the very specific cases in Scornet (2020).

Some of the flaws in the above methods are surprising while others, in retrospect, appear foreseeable. So are there ways to more appropriately and rigorously assess variable importance? Yes, though none are as computationally or algorithmically simple as those above.

The primary problem with the out-of-bag approaches outlined above is that models are constructed only on the original data – inspections are made by altering (permuting) test data, but never by altering the training data and reconstructing the model. We refer to this general idea as "permute-and-repredict" in Hooker and Mentch (2019). In most practical applications, however, when one asks whether a particular feature – say X_1 – is "important", they are really asking whether the model estimate (learning procedure, algorithm) could be just as accurate without it. Investigating this question directly thus necessarily involves the more computationally intensive idea of rebuilding the model under alterations to the training data.

In our earlier work on random forests (Mentch and Hooker, 2016), we proposed a hypothesis test for variable importance based on exactly this idea: we construct one random forest on the original data, then another where X_1 is removed from the dataset and compare the difference in predictions between the two models. Noticing that predictions are sometimes affected by the dimension of the feature space, we advocated that the test be carried by permuting X_1 in the data used by the second random forest rather than removing it from the model entirely. In more recent work, Coleman et al. (2019) proposed a more computationally efficient version of this test that measures the difference in accuracy between the two forests rather than the difference in raw predictions. This kind of idea as also appeared in the conformal inference literature where Lei et al. (2018) proposed $LOCO$ (Leave-One-Covariate-Out) tests that follow exactly the same intuition and in a model-agnostic test recently proposed in Williamson et al. (2020).

While the "drop-and-rebuild" and "permute-and-rebuild" approaches both improve upon the "permute-and-repredict" framework, neither is perfect in practical settings because in addition to breaking the relationship between the feature and response, we also break the relationship between the feature and all remaining features. A natural further fix is thus to replace X_1 with a copy generated from its distribution *conditional* on X_{-1}. This was, to our knowledge, first advocated in Strobl et al. (2008) and holds some resemblance to the new literature on knockoffs (Barber et al., 2015), conditional randomization tests (Candès et al., 2018; Liu and Janson, 2020), and holdout randomization tests (Tansey et al., 2018). However, while such an approach is certainly theoretically appealing, the task of generating conditional simulations of $X_1|X_{-1}$ is non-trivial; a number of potential methods exist to do this, but more work needs to be done here and we anticipate this research area will remain very important in the coming years.

---

1. This critique also motivated the GUIDE procedure (Loh, 2002).

Finally, we note that the definition of variable importance itself remains somewhat murky and, much to our frustration, commonly undefined in practical applications. Conditional generation implicitly suggests that we are primarily interested in the predictive ability of X_1 *beyond that provided by the other covariates*. We could, instead, simulate $X_{-1}|X_1$, or some subset of the covariates. To this end, Shapley values (Lundberg and Lee, 2017) have been suggested as representing the marginal improvement in predictive accuracy given by X_1, averaged over the order of inclusion. Alternative functional ANOVA methods (Hooker, 2007), relate variable importances to statistically-familiar types of sums of squares, allow the examination of interactions and can account for covariate distributions, albeit at a potentially prohibitive computational cost.

4.3 Formalized Inference from Random Forests

"an algorithmic model can produce more and more reliable information about the structure of the relationship between inputs and outputs than data models." – page 200

While statistical theory and insight has helped to understand and improve our methods of extracting information from a random forest, we can go further and start to use random forests explicitly for statistical inference. The CLT in (1) can be used to provide confidence intervals about individual predictions (Mentch and Hooker, 2016) or other quantities averaged over some covariate distribution (Wager and Athey, 2018). However, ideas such as variable importance suggest hypotheses about structure along the lines of

$$H_0: \ \eta(X_1, X_{-1}) = \eta^*(X_{-1}) \tag{2}$$

for some function η^*. Here we can use the extension of (1) to multivariate or process distributions over the input space. In particular, Mentch and Hooker (2016) provide formal tests of this form by evaluating the differences

$$D(X) = \frac{1}{B} \sum (T_b^\omega(X) - T_b^\pi(X)) \tag{3}$$

between trees T^ω build with the original data, and T^π that used a permuted version of X_1. Since the $D(X)$ have a joint multivariate central limit theorem, we can form a χ-square test of their values at a set of equation points.

Alternatively, the functional ANOVA methods in Hooker (2007) were turned into formal tests in Mentch and Hooker (2017). Specifically, we interpret the hypothesis in (2) as

$$H_0: \ \eta(X_{i,1}, X_{k,-1}) = \eta(X_{j,1}, X_{k,-1})$$

over a set of covariate values $(X_{i,1}, X_{i,-1})$. This creates a grid of function evaluations for which H_0 is a linear contrast that can be again tested via a multivariate χ-square statistic. The same techniques can be extended to examine interactions by testing hypotheses of the form

$$H_0': \ \eta(X_{i,1}, X_{j,2}, X_{k,-(1,2)}) = \eta_1(X_{i,1}, X_{k,-(1,2)}) + \eta_2(X_{j,2}, X_{k,-(1,2)}).$$

Any of these testing frameworks require evaluation at a large number of covariate values, making variance estimates unstable, and Mentch and Hooker (2017) resorted to random

projection methods (Srivastava et al., 2016) to stabilize them. They also rely heavily on asymptotic normality results and on a fixed set of evaluation points. Coleman et al. (2019) considered using the two forests that contribute to (3), but developed a permutation test based on randomly swapping trees between forests and measuring the change in differences in error, meaning that far fewer trees need to be constructed, thus allowing for very large test sets.

We hope to show by these methodological developments that Breiman's notions of "algorithmic modeling" can be accommodated and used readily within classical statistical paradigms, and refer to Chipman et al. (2010) for parallel Bayesian developments. Some of the methods have described lean heavily on at least the ensemble structure of random forests, but others are more general provided a result of the form of (1) is available. When "algorithmic models" are understood as non-parametric regression methods, the chasm that Breiman perceived within statistics can be crossed with only a few well-placed stepping stones.

4.4 Why do Random Forests Work?

"But when a model is fit to data to draw quantitative conclusions... the conclusions are about the model's mechanism, and not about nature's mechanism." – Page 202.

Despite the enormous progress made on the statistical front in recent years, the question that has perhaps proved the most elusive is also the simplest: Why do random forests seem to work so well in practice? In their recent review paper, Biau and Scornet (2016) note that "present results are insufficient to explain in full generality the remarkable behavior of random forests." And indeed, studies have long demonstrated the surprisingly robust accuracy of the random forest method. The most recent and impressive large-scale empirical comparison of methods was provided in Fernández-Delgado et al. (2014) where the authors compare a total of 179 classifiers across 121 datasets – the entirety of the UCI database (Dua and Graff, 2017) at the time. They found that not only did random forests perform the best overall, but 3 of the top 5 classifiers were some variant of random forests, leading them to declare that "the random forest is clearly the best family of classifiers."

Until very recently however, few if any explanations for random forest success had been offered that pushed beyond Breiman's original intuition of reducing the between-tree correlation. Wyner et al. (2017) focused on classification settings and hypothesized that their success was due to interpolation – a kind of purposeful overfitting that, in their view, was beneficial because perhaps real-world data are not as noisy as statisticians generally believe.

Mentch and Zhou (2020a) were critical of this explanation, noting in particular that in regression settings, the opposite appears to be true: random forests appear to be *most* useful in settings where the signal-to-noise ratio (SNR) is low. The authors proposed an alternative explanation based on degrees of freedom wherein the additional node-level randomness in random forests implicitly regularizes the procedure, making it particularly attractive in low SNR settings. Based on this idea, they demonstrate empirically that the advantage of random forests over bagging is most pronounced at low SNRs and disappears at high SNRs; when the trees in random forests are replaced with linear models, these kind of implicit

regularization claims and can be proved formally. This finding deals a serious blow to the widely held belief shared by both Breiman (2001) and Wyner et al. (2017) that random forests simply "are better" than bagging as a general rule. Perhaps most surprisingly, the authors also demonstrate that when a linear model selection procedure like forward selection incorporates a similar kind of random feature availability at each step, the resulting models can be substantially more accurate and even routinely outperform classical regularization methods like the lasso.

In follow-up work, Mentch and Zhou (2020b) build on this implicit regularization to introduce another surprising result: the predictive accuracy of both bagging and random forests can sometimes be dramatically and systematically improved (particularly at low SNRs) by including additional noise features in the model which, by construction, hold no additional information about the response beyond that contained in the original features. The authors refer to these kinds of procedures as *augmented bagging* and show that even the formal hypothesis testing procedures above will, correctly but counterintuitively, recognize large groups of noise features as making a significant contribution to the predictive accuracy of the model.

This last point is quite surprising and easy to misinterpret; readers seeing this for the first time might reasonably ask whether these kinds of hypothesis tests should really be trusted if they routinely identify noise features as predictively significant. We stress, however, that these tests are merely evaluating whether there is a significant improvement in accuracy when additional noise features are added to the model. This does indeed appear to be the case in some settings – models with additional noise features really *can* make the models more accurate. Thus, in those settings, by identifying such noise features as predictively helpful, it's important to realize that these tests are not making an error. Rather, as discussed at length in Mentch and Zhou (2020b), it is not the tests themselves that are wrong but the common (mis)interpretations of those tests.

However, this observation does confound two separate questions: whether the feature improved the predictive accuracy of a model, and whether it is associated with the response by real-world processes. Alternatively: is the feature *uniquely* helpful: do we need to include the same measurements, or would the current value of the microwave background radiation be just as useful? "What is your model for the data?" – the question Breiman objected to – in fact becomes essential: are we measuring a property of the world that generated the data, or the algorithm used to analyze it? Both are reasonable to measure, but they are not the same, and can't be assessed by the same methods. Crucially, Breiman's standard of predictive accuracy, while applying to properties of the algorithm, is not necessarily indicative of real-world processes. We expect that deeper studies in the general area of variable importance will continue to remain an important topic of study in future years as data becomes larger and more complex and black-box models are increasingly relied upon as a result.

5. Illustration

"It maybe revealing to understand how I became a member of the small second culture... My experiences as a consultant formed my views about algorithmic modeling." – page 200

Brieman illustrates the development of his views on how to approach data through a series of examples and David Cox, his strongest critic, also notes the paucity of data that actually appears in the highest profile statistics journals. The conversation between statistical theory and applied practice is of vital importance, particularly in discussing broader questions about modeling philosophy, and we here provide a précis of a project to which we contributed and which we believe demonstrates well the fusion of both machine learning methods and statistical thinking. The interested reader is referred to Coleman et al. (2020) for full details.

The ebird project at the Cornell Laboratory of Ornithology (www.ebird.org) collects data from amateur birdwatchers across the United States. This data repository has been used to estimate over-all changes in bird abundance, provide maps of bird migration, and explore the dependence of bird distributions on ecological features. It has been a significant motivation for our own research. The project we describe used these data to assess the impact of local weather conditions on the timing tree swallows' (*Tachycineta bicolor*) fall migration. Specifically, daily maximum temperature was expected to drive tree swallow migration through its (unobserved) effect on the abundance of flying insects that make up the greatest part of their food source.

Modeling migration effects is made challenging by complex spatio-temporal effects along with high-order interactions among other landscape and land-use features (Zuckerberg et al., 2016). This makes machine learning models particularly useful, and these have be extensive explored, for example in Fink et al. (2020). In this project we had a data set of 25727 observations collected after day 200 on each year from 2008 through 2013 in six wildlife refuges along the north eastern United States coast. These observations recorded the presence or absence of tree swallow sightings, the time and day of the recording, and the time and distance effort put into data collection. We also collected local landcover and elevation summaries from remote sensing systems as well as the daily maximum temperature, providing a total of 30 variables for each observation.

Within this data set, daily maximum temperature is strongly associated with day of year to the extent that there is little evidence of its predictive utility. This relationship also makes permuting the maximum temperature values problematic in creating unrealistically warm December days and very chilly Augusts. Instead, we replace maximum temperature with its difference from the average for that day and location over the six year period. This orthogonalization isolates the effect of weather from other temporal drivers, both allowing a machine learning algorithm to make use of the signal and reducing the effect of permuting the difference.

We assessed whether maximum temperature had a discernible affect on tree swallow abundance by examining the differences in predictions between tree ensembles obtained from the original data and those obtained using a permuted version of maximum temperature difference. This was done at 25 time-locations pairs in each of the wildlife refuges in which we used a multivariate version of (1) to develop a χ^2 test for each 25-vector of prediction differences. This revealed strong evidence of an effect at 4 of the regions with much attenuated evidence in the most northerly and southerly regions where we expect migration timing and climate to already reduce the role of temperature.

This study illustrates the combining of the modeling cultures that Breiman describes: statistical considerations play a key role in designing the covariates we used, the development

of our tests, and our assessment of evidence, but we avoided the need for extensive and complex models by using algorithmic tools: treating these as statistical models is not a large leap. Debates around causality and fairness in machine learning have equally demonstrated the need to combine schools in this way and we expect the gap that Breiman perceived to become increasingly invisible.

6. What is the New Algorithmic Modeling?

"This is a fascinating enterprise, and I doubt if data models are applicable. Yet I would enter this in my ledger as a statistical problem." – page 214

Breiman's essay clearly displays his frustration with the conservativism of a statistics community unwilling to engage with unfamiliar ideas and methods. This same sentiment is not unfamiliar to us but it is also hardly unique to statistics, as many statisticians who have received reviews from the computer science conferences promoted by Breiman can attest. We do not subscribe to the sentiment that "Statistics lost Machine Learning", but the statement does lead us to ask what other areas we might currently be going overlooked.

An obvious answer is "deep learning" (viewed at the time of Breiman's essay as being somewhat played out) and indeed we are aware of starkly little statistical involvement in this intensely hot area of machine learning research where the greatest performance gains over the past decade – Breiman's primary metric of validity – have come from. The work that we are aware of (e.g. Barron and Klusowski, 2018; Bai et al., 2020) focusses on shallow networks and largely ignores the algorithms used to learn weights. This seems analogous to the original work on random forests and we hope that it will similarly lead to steadily more practical results, but also expect progress will require new ways of representing the problem.

However, rather than the mathematical properties of neural networks, the aspect we most worry about statisticians ignoring is the type of data where these tools excel: images, text, sound and other natural data types that are both highly structured – e.g. pixels are next to other pixels – and where individual numbers have little to no individual interpretive value. (The large statistical literature on medical imagining requires making images comparable on a value-by-value or at least region-by-region level. Similar comments can be made about inference with many other non-Euclidean objects where defining a distance appears to be a starting point.) We have seen little statistical interest in the way deep learning is applied to these types of complex data objects where questions about manifold structure, the stability of explanation methods, or structuring models that extract information could all benefit from statistical perspectives, even as we freely acknowledge that we, ourselves, have not contributed.

Outside of these specifics, we expect that Breiman would have welcomed the broadening of interests associated with the field of Data Science. Statistics has a long history of worrying about the design or causal issues in data provenance and the biases associated with data collection, and a somewhat more limited engagement with data privacy. Much more of this is sorely needed as the emerging issues in ethical AI are making clear. We think the adjacent areas of data merging and data cleaning are both ripe for more formal study; developing methods and analyzing their consequences. There are surely other tasks of which we haven't

thought, but we really do need to encourage the discipline to expand beyond our collective intellectual comfort zone.

7. Discussion

"*The roots of statistics, as in science, lie in working with data and checking theory against data. I hope in this century our field will return to its roots. There are signs that this hope is not illusory. Over the last ten years, there has been a noticeable move toward statistical work on real world problems and reaching out by statisticians toward collaborative work with other disciplines. I believe this trend will continue and, in fact, **has** to continue if we are to survive as an energetic and creative field.*" – Page 214

Looking around today at the emerging field of data science and the resulting shifts within the statistics community, it's easy to see the impact of Breiman's thinking. As faculty members in statistics departments who are active in data science education at both the undergraduate and graduate levels, we frequently interact with students who are new to the field, many of whom are reading Breiman's seminal essay for the first time. In the vast majority of cases, we find that these young statistics students are puzzled as to why it was ever considered a controversial work. "*Isn't it obvious that this is what we're supposed to be doing?*" Ironically, perhaps the best indication of Breiman's lasting impact through his "Two Cultures" essay is its lessened relevance today.

Just as the field of statistics has begrudgingly opened it's mind to problems outside the classical modeling framework, computer scientists and machine learning researchers deserve credit for expanding their horizons as well. A newfound interest in interpretability (e.g. "explainable AI"), causality, and fairness is commonly on display at each of the most notable machine learning conferences, each of which hearkens back to fundamental ideas and philosophical considerations in traditional statistical modeling.

It's worth noting that while the boundaries of these fields may have softened, it remains the case that their respective directions are not unified. Computer scientists remain largely driven by the "what" (predictive accuracy); statisticians the "how" and "why." We still observe a separation of concerns: global interpretations versus local explanations mirrors scientific knowledge versus actions. In our view, however, so long as we resist dogmatic aversions to alternative ways of thinking, unity shouldn't be seen as a necessity for (or perhaps even the optimal route to) continued progress.

Finally, there are likely some statisticians who, like David Cox's commentary in the original paper, remain skeptical of the recent trends in our field towards data science, fearing that it signifies a surrender of the theoretical school of statistics and traditional modeling. We do not think so; rather it is the realization that both sides might have something to tell the other. Rather than seeing these trends as *shifting away from* classical statistical ideas, we urge skeptical colleagues to consider these recent developments an *expansion of* those ideas. This, we firmly believe, is the central theme of Breiman's essay and is ultimately what allowed for the exciting recent developments detailed in the previous sections.

References

Susan Athey, Julie Tibshirani, Stefan Wager, et al. Generalized random forests. *Annals of Statistics*, 47(2):1148–1178, 2019.

Jincheng Bai, Qifan Song, and Guang Cheng. Efficient variational inference for sparse deep learning with theoretical guarantee. *arXiv preprint arXiv:2011.07439*, 2020.

Rina Foygel Barber, Emmanuel J Candès, et al. Controlling the false discovery rate via knockoffs. *Annals of Statistics*, 43(5):2055–2085, 2015.

Andrew R Barron and Jason M Klusowski. Approximation and estimation for high-dimensional deep learning networks. *arXiv preprint arXiv:1809.03090*, 2018.

Gérard Biau. Analysis of a Random Forests Model. *The Journal of Machine Learning Research*, 98888:1063–1095, 2012.

Gérard Biau and Luc Devroye. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, 101(10):2499–2518, 2010.

Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25(2):197–227, 2016.

Gérard Biau, Luc Devroye, and Gábor Lugosi. Consistency of Random Forests and Other Averaging Classifiers. *The Journal of Machine Learning Research*, 9:2015–2033, 2008.

Leo Breiman. Random Forests. *Machine Learning*, 45:5–32, 2001.

Leo Breiman, Jerome Friedman, Charles J. Stone, and R.A. Olshen. *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1st edition, 1984.

Leo Breiman et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

Peter Lukas Bühlmann. Consistency for l2 boosting and matching pursuit with trees and tree-type basis functions. In *Research report/Seminar für Statistik, Eidgenössische Technische Hochschule (ETH)*, volume 109. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH), 2002.

Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

Hugh A Chipman, Edward I George, Robert E McCulloch, et al. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

Tim Coleman, Wei Peng, and Lucas Mentch. Scalable and efficient hypothesis testing with random forests. *arXiv preprint arXiv:1904.07830*, 2019.

Tim Coleman, Lucas Mentch, Daniel Fink, Frank A La Sorte, David W Winkler, Giles Hooker, and Wesley M Hochachka. Statistical inference on tree swallow migrations with random forests. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69 (4):973–989, 2020.

Roger B Davis and James R Anderson. Exponential survival trees. *Statistics in medicine*, 8(8):947–961, 1989.

Ramón Díaz-Uriarte and Sara Alvarez De Andres. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):1–13, 2006.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Randall L Eubank. *Nonparametric regression and spline smoothing*. CRC press, 1999.

Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1):3133–3181, 2014.

Daniel Fink, Tom Auer, Alison Johnston, Viviana Ruiz-Gutierrez, Wesley M Hochachka, and Steve Kelling. Modeling avian full annual cycle distribution and population trends with citizen science data. *Ecological Applications*, 30(3):e02056, 2020.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer, 1996.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Louis Gordon and Richard A Olshen. Tree-structured survival analysis. *Cancer treatment reports*, 69(10):1065–1069, 1985.

Jennifer Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20:217–240, 03 2011. doi: 10.1198/jcgs.2010.08162.

Giles Hooker. Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732, 2007.

Giles Hooker and Lucas Mentch. Please stop permuting features: An explanation and alternatives. *arXiv preprint arXiv:1905.03151*, 2019.

Torsten Hothorn, Berthold Lausen, Axel Benner, and Martin Radespiel-Tröger. Bagging survival trees. *Statistics in medicine*, 23(1):77–91, 2004.

Torsten Hothorn, Peter Bühlmann, Thomas Kneib, Matthias Schmid, and Benjamin Hofner. Model-based boosting 2.0. *Journal of Machine Learning Research*, 11:2109–2113, 2010.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Xiao Li, Yu Wang, Sumanta Basu, Karl Kumbier, and Bin Yu. A debiased mdi feature importance measure for random forests. *Advances in Neural Information Processing Systems*, 32:8049–8059, 2019.

Molei Liu and Lucas Janson. Fast and powerful conditional randomization testing via distillation. *arXiv preprint arXiv:2006.03980*, 2020.

Markus Loecher. Unbiased variable importance for random forests. *Communications in Statistics-Theory and Methods*, pages 1–13, 2020.

Wei-Yin Loh. Regression tress with unbiased variable selection and interaction detection. *Statistica sinica*, pages 361–386, 2002.

Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–631. ACM, 2013.

Benjamin Lu and Johanna Hardin. A unified framework for random forest prediction error estimation. *arXiv preprint arXiv:1912.07435*, 2019.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NIPS*, 2017.

Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999, 2006.

Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research*, 17(1):841–881, 2016.

Lucas Mentch and Giles Hooker. Formal hypothesis tests for additive structure in random forests. *Journal of Computational and Graphical Statistics*, 26(3):589–597, 2017.

Lucas Mentch and Siyu Zhou. Randomization as regularization: A degrees of freedom explanation for random forest success. *Journal of Machine Learning Research*, 21(171): 1–36, 2020a. URL http://jmlr.org/papers/v21/19-905.html.

Lucas Mentch and Siyu Zhou. Getting better from worse: Augmented bagging and a cautionary tale of variable importance. *arXiv preprint arXiv:2003.03629*, 2020b.

Wei Peng, Tim Coleman, and Lucas Mentch. Rates of convergence for random forests via generalized u-statistics. *arXiv preprint arXiv:1905.10651*, 2019.

Veronika Ročková and Enakshi Saha. On theory for bart. In *Proceedings of Machine Learning Research*, volume 89, pages 2839–2848, 2019.

Cynthia Rudin. Please stop explaining black box models for high stakes decisions. *arXiv preprint arXiv:1811.10154*, 1, 2018.

Erwan Scornet. Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62(3):1485–1500, 2016.

Erwan Scornet. Trees, forests, and impurity-based variable importance. *arXiv preprint arXiv:2001.04295*, 2020.

Erwan Scornet, Gérard Biau, Jean-Philippe Vert, et al. Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741, 2015.

Mark Segal and Yuanyuan Xiao. Multivariate random forests. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 1(1):80–87, 2011.

Mark Robert Segal. Regression trees for censored data. *Biometrics*, pages 35–47, 1988.

Radhendushka Srivastava, Ping Li, and David Ruppert. Raptt: An exact two-sample test in high dimensions using random projections. *Journal of Computational and Graphical Statistics*, 25(3):954–970, 2016.

Jon Arni Steingrimsson, Liqun Diao, Annette M Molinaro, and Robert L Strawderman. Doubly robust survival trees. *Statistics in medicine*, 35(20):3595–3612, 2016.

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25, 2007.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC bioinformatics*, 9 (1):307, 2008.

Sarah Tan, Matvey Soloviev, Giles Hooker, and Martin T Wells. Tree space prototypes: Another look at making tree ensembles interpretable. In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, pages 23–34, 2020.

Wesley Tansey, Victor Veitch, Haoran Zhang, Raul Rabadan, and David M Blei. The holdout randomization test: Principled and easy black box feature selection. *arXiv preprint arXiv:1811.00645*, 2018.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Brian D Williamson, Peter B Gilbert, Noah R Simon, and Marco Carone. A unified approach for inference on algorithm-agnostic variable importance. *arXiv preprint arXiv:2004.03683*, 2020.

Abraham J Wyner, Matthew Olson, Justin Bleich, and David Mease. Explaining the success of adaboost and random forests as interpolating classifiers. *The Journal of Machine Learning Research*, 18(1):1558–1590, 2017.

Bin Yu. Ims presidential address: Let us own data science. IMS Annual Meeting, Sydney, Australia, July 9-14, 2014. URL https://imstat.org/2014/10/01/ims-presidential-address-let-us-own-data-science/.

Haozhe Zhang, Joshua Zimmerman, Dan Nettleton, and Daniel J Nordman. Random forest prediction intervals. *The American Statistician*, 2019.

Yichen Zhou and Giles Hooker. Boulevard: Regularized stochastic gradient boosted trees and their limiting distribution. *arXiv preprint arXiv:1806.09762*, 2018.

Zhengze Zhou and Giles Hooker. Unbiased measurement of feature importance in tree-based methods. *arXiv preprint arXiv:1903.05179*, 2019.

Zhengze Zhou, Lucas Mentch, and Giles Hooker. v-statistics and variance estimation. *arXiv preprint arXiv:1912.01089*, 2019.

Ruoqing Zhu and Michael R Kosorok. Recursively imputed survival trees. *Journal of the American Statistical Association*, 107(497):331–340, 2012.

Benjamin Zuckerberg, Daniel Fink, Frank A La Sorte, Wesley M Hochachka, and Steve Kelling. Novel seasonal land cover associations for eastern north american forest birds identified through dynamic species distribution modelling. *Diversity and Distributions*, 22(6):717–730, 2016.