Remembering Leo Breiman

Article *in* The Annals of Applied Statistics · January 2011

DOI: 10.1214/10-AOAS427 · Source: arXiv

1 author:

Adele Cutler
Utah State University
45 PUBLICATIONS **7,790** CITATIONS

The Annals of Applied Statistics
2010, Vol. 4, No. 4, 1621–1633
DOI: 10.1214/10-AOAS427

REMEMBERING LEO BREIMAN[1]

By Adele Cutler

Utah State University

Leo Breiman was a highly creative, influential researcher with a down-to-earth personal style and an insistence on working on important real world problems and producing useful solutions. This paper is a short review of Breiman's extensive contributions to the field of applied statistics.

1. Introduction. How many theoretical probabilists walk away from a tenured faculty position at a top university and set out to make their living as consultants? How many applied consultants get hired into senior faculty positions in first-rate research universities? How many professors with a fine reputation in their field, establish an equally fine reputation in a *different* field, *after* retirement? Leo Breiman did all of these things and more. He was an inspiring speaker and a convincing writer, doing both with seemingly boundless enthusiasm, in an unpretentious, forthright manner that he called his "casual, homespun way." He was intelligent and thought deeply about research. But there are a number of bright, talented statisticians. What made Breiman different? For one thing, he was willing to take risks. By and large, statisticians are not great risk-takers. We tend not to stray too far from what we know, tend not to tackle problems for which we have no tools, tend to adopt or adapt existing ideas instead of coming up with completely new ones. Linked to this willingness to take risks was Breiman's unusual creativity. It was not a wild, off-the-wall creativity—it was grounded in a sound knowledge of theoretical principles and directed by an intuition gained by working intensively with data, along with a generous dose of common sense. Breiman was driven by challenging and important real-data problems that people cared about. He didn't spend time publishing things just because he could, filling the gaps just because they were there. Lastly, he was tenacious. He would not give up on a problem until he, or someone else, got to the bottom of what was going on.

Received October 2010.
[1]For Jessica, Rebecca and Mary Lou Breiman.
Key words and phrases. Arcing, bagging, boosting, random forests, trees.

Some of Breiman's ideas have advanced the field in and of themselves (e.g., bagging, random forests) while others have contributed more indirectly (e.g., Breiman's nonnegative garrote [Breiman (1995a)] inspired the lasso [Tibshirani (1996)]). Although his joint work tree-based methods [Breiman et al. (1984)] was arguably his most important contribution to science, he viewed random forests as the culmination of his work. I consider myself privileged to have been able to work with Leo Breiman for almost 20 years, as his student, collaborator and friend, and I'm honored to have been asked to write this review of his contributions to applied statistics. I have divided the paper into roughly chronological sections, but these have considerable overlap and are intended to be organizational rather than definitive. I kept biographical details to a minimum; those interested in a biography are referred to Olshen (2001). I do not feel qualified to discuss Breiman's work on the 1991 Census adjustment [Breiman (1994)] and have omitted a few other isolated pieces of work such as Breiman, Tsur and Zemel (1993); Breiman, Meisel and Purcell (1977); Breiman (1999b); Breiman and Cutler (1993).

2. Early work. Breiman was born in New York City in 1928 and educated in California, receiving his Ph.D. in mathematics from UC Berkeley in 1954. He earned tenure as a probabilist in the UCLA Mathematics Department but soon after, he "got tired of doing theory and wanted something that would be more exciting" (personal communication) so he resigned. At this time, Breiman was already interested in classification, co-authoring a paper on the convergence properties of a "Learning Algorithm" [Breiman and Wurtele (1964)]. Curiously, the paper had only two references, one of which was to some early work by Seymour Papert, who was later to become one of the pioneers of artificial intelligence and co-author of an influential (and controversial) book on perceptrons [Minsky and Papert (1969)].

After resigning, the first thing Breiman did was to write his probability book [Breiman (1968)] and then, with no formal statistical training, he proceeded to spend the next 13 years as a consultant. As well as some work in transportation, he worked for William Meisel's division of Technology Services Corporation, doing environmental studies and unclassified defense work. It's difficult to imagine making such a transition today, but one can speculate that it was in part, *because* he did not have a background in applied statistics that Breiman was so successful at consulting. Certainly the prediction problems on which he worked, some of which are mentioned in Breiman (1984) and Section 3 of Breiman (2001c), would have been a challenge for the tools and computers of the time. In Breiman (2001c), he acknowledges Meisel for helping him "make the transition from probability theory to algorithms."

3. Classification and regression trees. One of the early problems Breiman worked on as a consultant was to classify ship types from the peaks of radar profiles. The observations had different numbers of peaks and the number of peaks and their locations depended on the angle the ship made with the radar. After "a lot of head-scratching and a lot of time just thinking" the idea of a classification tree came to him "out of the blue." After this, Meisel's research team began using trees regularly. Charles Stone was brought on board, became interested in trees, and worked with Breiman to improve accuracy. In the early to mid-1970s, Breiman and Stone came up with the breakthrough idea of using cross validation to prune large trees.

It's difficult to obtain published work from Breiman's consulting years, but by 1976, Breiman and Meisel published an early version of regression trees [Breiman and Meisel (1976)] which broke down the data space into regions and fitted a linear regression in each region. Regions were split using a randomly oriented plane and an F-ratio was used to determine if the split had significantly reduced the residual sum of squares; if not, another random split was tried. In retrospect, the idea of using randomly chosen splits seems a good 20 years ahead of its time. The statement "many typical data analytic problems are characterized by their high dimensionality... and the lack of any a priori identification of a natural and appropriate family of regression functions" [Breiman and Meisel (1976)] was a clear indicator of Breiman's future research directions.

In 1976, Breiman met Jerome Friedman, a high-energy particle physicist, and soon Friedman was also working as a consultant for TSI. Both Friedman and Stone had connections to Richard Olshen, and the four started to collaborate. Apparently, they decided to publish their research as a book because they believed the work was unlikely to be published in the standard statistical journals.

In 1980, Stone and Breiman joined the UC Berkeley Statistics Department, and the group experimented with different splitting criteria, refined the cross-validation approach, and came up with the idea of surrogate splits. Several things set this work apart from other early work on trees. First, they did painstaking experiments. As they report in Breiman and Friedman (1988), "In the course of the research that led to CART, almost two years were spent experimenting with different stopping rules. Each stopping rule was tested on hundreds of simulated data sets with different structures." Second, they kept applications in the foreground of their work, due in part to Breiman's years as a consultant. Third, they had what Breiman referred to as "some beautiful and complex theory." The book, priced low to make it accessible, was published in 1984 [Breiman et al. (1984)].

4. ACE and additive models. I once heard Charles Stone express regret that the CART group had not written a follow up book of "things we tried

that didn't work." I expect such a book could have prevented a number of researchers from reinventing the wheel, but few would want to read such a book, much less write it. In fact, after completing Breiman et al. (1984), Breiman admitted to being "completely fed up with thinking about trees." Breiman and Friedman continued to talk, because both were interested in high-dimensional data analysis, and soon they came up with the Alternating Conditional Expectations (ACE) algorithm [Breiman and Friedman (1985)]. For predictor variables $X_1, \ldots, X_p$ and response Y, ACE defines $\phi_1^\star, \ldots, \phi_p^\star$ and $\theta^\star$ to minimize

$$\mathrm{E}\left[\theta(Y) - \sum_{j=1}^{p} \phi_j(X_j)\right]^2$$

under the constraint $\mathrm{Var}(\theta) = 1$. Estimates $\phi_1^\star, \ldots, \phi_p^\star$ and $\theta^\star$ were obtained using an iterative optimization procedure involving (nonlinear) smoothing to estimate each of the transformations while holding the others fixed. This was an application of the Gauss–Seidel algorithm of numerical linear algebra. A simpler version, taking θ as the identity, is the familiar "backfitting" algorithm [Hastie and Tibshirani (1986), Buja, Hastie and Tibshirani (1989)].

ACE was the first in a series of papers Breiman wrote on smoothing and additive models. Breiman and Peters (1992) compared four scatterplot smoothers using an extensive simulation. Building on the spline models used in Breiman and Peters (1992), Breiman's Π method [Breiman (1991)], with the colorful acronym "PIMPLE," fit additive models of products of (univariate) cubic splines. Hinging hyperplanes [Breiman (1993b)] fit an additive function of hyperplanes, continuously joined along a line called a "hinge." According to Breiman (1993a), while ACE provided the "first available method for fitting additive models to data," it had some difficulties. For small sample sizes, the results were "noisy and erratic." The nonlinearity of the smoother combined with the iterative algorithm led to results that were "difficult to analyze and sometimes mildly unstable." So Breiman went back to the drawing board, adapting a spline-based method using stepwise deletion of knots [Smith (1982)], resulting in Breiman (1993a). This paper contains early thoughts on using cross-validation to measure instability: "If transformations change drastically when one or a few cases are removed, then they do not reflect an overall pattern in the data." These early ideas of instability ultimately led to some of Breiman's most influential work.

5. Multivariate techniques. While all Breiman's work was multivariate, some was more clearly affiliated with traditional multivariate techniques. In 1984, Breiman and Ihaka released a technical report [Breiman and Ihaka (1984)] describing a nonlinear, smoothing-based version of discriminant analysis. The work was never published but it motivated the work on "Flexible Discriminant Analysis" by Hastie, Tibshirani and Buja (1994).

In his consulting days, one of the problems Breiman studied was next-day ozone prediction. One of his ideas was to represent each day as a mixture of "extreme" or "archetypal" days. For example, an archetypal sunny day would be as sunny as possible, an archetypal rainy day would have as much rain as possible, an archetypal foggy day would have fog for as long as possible, and so on. Most days would not be archetypal—they would fall in between the archetypes, resembling each to a greater or lesser extent. For data $\{x_i, i = 1, \ldots, N\}$, the problem was to find archetypal points $\{z_k, k = 1, \ldots, K\}$ to minimize

$$\sum_i \left\| x_i - \sum_k \alpha_{ik} z_k \right\|^2$$

subject to the constraints $\alpha_{ik} \geq 0, \sum_k \alpha_{ik} = 1$, while also constraining the z_k's to fall on or inside the convex hull of the data. The problem can be solved using an alternating least squares algorithm [Cutler and Breiman (1994)]. Archetypes have been used as an alternative to cluster analysis or principal components in numerous disciplines.

The final method in this section is a paper on multivariate regression, whimsically called "curds and whey" [Breiman and Friedman (1997)]. To predict correlated responses, Breiman and Friedman considered predicting each response by a linear combination of the ordinary least squares (OLS) predictors rather than the OLS predictors themselves. The method worked by transforming into canonical coordinates, shrinking, then transforming back. Cross-validation was used to choose the amount of shrinkage.

6. Subset selection in linear regression. Breiman had a longstanding interest in submodel selection in linear regression, revealing itself in Breiman and Meisel (1976), which used an early version of a regression tree to estimate the "intrinsic variability" of the data, with the goal of effectively ranking the predictive capabilities of subsets of independent variables. Breiman and Freedman (1983) looked at determining the optimal number of regressors to minimize mean squared prediction error. Again, using prediction error as the gold standard, Breiman (1992) and Breiman and Spector (1992) contained careful and thorough simulation studies for the X-fixed and X-random situations.

As Efron (2001) mentioned, Leo's "openness to new ideas whatever their source" was an attractive feature of his work. One example of this openness was that in the early 1990s, Leo got interested in neural nets and started participating in the Neural Information Processing Systems (NIPS) conference and workshops. Neural nets were not really a new idea, but they were enjoying new popularity among computer scientists, physicists and engineers, who in Leo's view were turning out "thousands of interesting research papers

related to applications and methodology" [Breiman (2001c)]. To this active community, Leo brought his considerable statistical background, experience with trees and subset selection, and perspective from years of dealing with real data and thinking about how to do it better. This led to Leo's most productive years, in part facilitated by his retirement from the UC Berkeley Statistics Department in early 1993, about which he said, "So far retirement has meant that I've got more time to spend on research" (personal communication).

The first work to appear from this period, stacking [Breiman (1996a)], was stimulated by Wolpert (1992) and first appeared as a technical report in 1992. In Breiman (1996a), he said, "In past statistical work, all the focus has been on selecting the "best" single model from a class of models. We may need to shift our thinking to the possibility of forming combinations of models." In the case of stacking, this was a linear combination of predictors. Each predictor was based on what Wolpert called the "level 1 data" [Wolpert (1992)]. Breiman (1996a) considered a family of models indexed by $k = 1, \ldots, K$. For example, k might be the number of variables in a subset selection method or k might index a collection of shrinkage parameters $\{\lambda_k, k = 1, \ldots, K\}$ for ridge regression. For data $\{x_{1n}, \ldots, x_{pn}, y_n, n = 1, \ldots, N\}$, each of the K predictors were fit to the data with observation n omitted (leave-one-out cross validation) to give k predictions of y_n, namely $z_{kn}, k = 1, \ldots, K$, which were the "level 1 data." The "stacked" predictor was $\sum_k \alpha_k z_{kn}$ where $\alpha_k \geq 0, k = 1, \ldots, K$, were chosen to minimize

$$\sum_n \left(y_n - \sum_k \alpha_k z_{kn} \right)^2.$$

Breiman considered stacked subsets and stacked ridge regressions and concluded that both were better than the existing method (choosing a single model by cross-validation). However, stacking improved subsets more than it improved ridge, which Breiman suggested was due to the greater instability of subset selection.

Building on stacking [Breiman (1996a)] and using some of his experiences from Breiman (1992) and Breiman and Spector (1992), Breiman introduced the nonnegative garrote [Breiman (1995a)]. For data as before and original OLS coefficients $\{\hat{\beta}_k\}$, the nonnegative garrote chose $\{c_k\}$ to minimize

$$\sum_n \left(y_n - \sum_k c_k \hat{\beta}_k x_{kn} \right)^2$$

subject to the constraints $c_k \geq 0$ and $\sum_k c_k \leq s$. This was a much simpler idea than stacking because it did not use Wolpert's "level 1 data" [Wolpert (1992)] and k ranged over the predictor variables instead of denoting the size

of a subset or the value of a shrinkage parameter. Breiman found [Breiman (1995a)] that the garrote had consistently lower prediction error than subset selection, and sometimes better than ridge regression. Breiman's ideas about instability, first mentioned in Breiman (1993a), led him to characterize of ridge regression as stable, subset selection unstable, and the garrote intermediate. Breiman remarked that "the more unstable a procedure is, the more difficult it is to accurately estimate PE (prediction error)" and speculated about finding a "numerical measure of stability." Bühlmann and Yu (2006) showed some interesting results for the garrote in a boosting context. However, the largest impact of the garrote was that it inspired the lasso [Tibshirani (1996)], which is currently the method of choice, in part because of garrote's dependence on $\{\hat{\beta}_k\}$.

Breiman's notions of stability were further explored in Breiman (1996b). He compared ridge regression, subset selection and two versions of garrote and stated, "Unstable procedures can be stabilized by perturbing the data, getting a new predictor sequence... and then averaging over many such predictor sequences." The types of perturbation he considerd are leave-one-out cross-validation, leave-ten-out cross-validation and adding random noise to the response variable. He stated [Breiman (1996b)] "we do not know yet what the best stabilization method is."

7. Bagging. Breiman released an early version of Breiman (1996b) in June 1994, but by September of the same year he released yet another technical report in which he had already resolved some of the questions raised in Breiman (1996b). He called the report "Bagging Predictors" and it was to be published as Breiman (1996c). The name comes from "bootstrap aggregating" because in bagging, the data were perturbed by taking bootstrap samples and the resulting predictors were averaged (aggregated) to give the "bagged estimate." The classification version aggregates by voting the predictors. In November 1994, Breiman presented bagging as part of a Tutorial at the NIPS conference, where it was immediately embraced by the neural net community. According to Google Scholar, citations of Breiman (1996c) already exceed 6300, slightly higher than Efron's 1979 bootstrap paper [Efron (1979)]. The simplicity and elegance of bagging made it appealing in a community where new ideas tended to be technically complex.

In bagging, each predictor was fit to a bootstrap sample, so roughly 37% of the observations were not included in the fit ("out-of-bag"). In an unpublished technical report Breiman (1997b) described how to use these for estimating node probabilities and generalization error.

Although bagging trees improved the accuracy of trees, Breiman liked the simple, understandable structure of individual trees and was not ready to give up on them. Noting that trees have "the disadvantage that the splits get

noisier as you go down" (personal communication), he worked with Nong Shang [Shang and Breiman (1996)] to try to improve the stability of trees by estimating the joint density of the data and basing the splits on this density estimate instead of directly on the data. One of the problems of this method was that density estimates depended on numerous parameters and Breiman referred to it later [Breiman (1998b)] as a "complex and unwieldy procedure." Another attempt, described in Breiman (1998b), was to generate new "pseudo-data" by randomly choosing an existing data point and moving its predictor variables a small step towards a second randomly-chosen data point. The new predictor values, together with the response for the original data point, gave the pseudo-data. The step size was chosen to be uniform on the interval $(0, d)$ where d was a parameter of the method. Although the results appeared promising, the method did not give improvements on large datasets and the paper was never published.

Breiman tried to improve upon bagging in a number of other ways. His "iterated" or "adaptive" bagging [Breiman (2001b)] was designed to reduce the bias of bagged regressions by successively altering the output values using the out-of-bag data. Naturally, this biases the out-of-bag generalization error estimates, but Breiman found that for the purpose of bias reduction it worked well [Breiman (2001b)]. In a similar vein, Breiman (2000a) provided an alternative to bagging by combining predictors fit to data for which only the output variables have been perturbed. It's not clear whether these ideas would have endured because Breiman did not release code and they were discarded once he discovered random forests [Breiman (2001a)].

8. Boosting and arcing. While Breiman developed bagging, Freund and Schapire worked on AdaBoost [Schapire (1990), Freund (1995), Freund and Schapire (1996)]. Breiman referred to the AdaBoost algorithm as "the most accurate general purpose classification algorithm available" [Breiman (2004b)]. Like bagging, AdaBoost combined a sequence of predictors. Unlike bagging, each predictor was fit to a sample from the training data, with larger sampling weights given to observations that had been misclassified by earlier predictors in the sequence. The predictions were combined using performance weights. In a personal communication, Breiman wrote, "Some of my latest efforts are to understand Adaboost better. Its really a strange algorithm with unexpected behavior. Its become like searching for the Holy Grail!!" In his quest, Breiman produced a series of papers [Breiman (1997a, 1998a, 1998c, 1999a, 2000b, 2004b)]. He noted in Breiman (1998a) that if AdaBoost "is run far past the point at which the training set error is zero, it gives better performance than bagging on a number of real data sets." This was a great mystery and Breiman was determined to get to the bottom of it. In Breiman (1998a), Breiman constructed a more general class of algorithms "arcing," of which AdaBoost, ("arc-fs") was a special case. One contribution

of Breiman (1998a) was that Breiman removed the randomness of boosting by using a weighted version of the classifier instead of sampling weights. Focusing on bias and variance, he concluded that "Arcing does better than bagging because it does better at variance reduction" [Breiman (1998a)], but Schapire et al. (1998) gave examples in which the main effect of AdaBoost was to reduce bias and proposed their own reasons for why boosting worked so well. Breiman thought the explanation was incomplete [Breiman (1999a)].

Breiman's work on half and half bagging [Breiman (1998c)] was stimulated by one of the referees of Breiman (1998a), who commented that the probability weight at a given step was equally divided between the points misclassified, and those correctly classified, at the previous step. In Breiman (1998c) Breiman divided the data into two parts, one containing "easy" points, the other "hard" points, based on previous classifiers in the sequence. He randomly sampled an equal number of cases from both groups and fitted a classification tree. For the first time, the tree was grown deep (one example per terminal node), which he later carried over to random forests [Breiman (2001a)].

In Breiman (1997a), he showed that AdaBoost is a "down-the-gradient" method for minimizing an exponential function of the error. Independently, Hastie, Tibshirani and Friedman (2000) presented "The Statistical View of Boosting."

About his "Infinity Theory" paper [Breiman (2000b)], Breiman stated in August 2000: "I've been compulsively working on a theory paper about tree ensembles which I got sick and tired of but knew that if I didn't keep going... it would never get finished." The paper was released as a technical report, cited by Zhang (2004) and Bühlmann and Yu (2003), among others. A later version was published as Breiman (2004b) and in this paper Breiman showed that the population version of AdaBoost was Bayes-consistent. In the meantime, several publications, including Hastie, Tibshirani and Friedman (2000), suggested that AdaBoost could overfit in the limit and Jiang (2004) showed that in the finite sample case, AdaBoost was only Bayes-consistent if it was regularized.

9. Random forests. In the light of boosting, Breiman spent a lot of time trying to improve individual trees [Shang and Breiman (1996), Breiman (1998b)] and bagged trees [Breiman (2000a, 2001b)]. He also worked very hard to understand what was going on with boosting [Breiman (1997a, 1998a, 1998c, 1999a, 2000b, 2004b)]. However, he never seriously produced a boosting algorithm for practical use, and I believe the reason was that he wanted a method that could give meaningful results for data analysis, not just prediction, and he didn't think he could get this by combining dependent predictors. The culmination of his work on bagging and how to improve

it, and his work trying to understand boosting, was a method Breiman called "random forests" (RF) [Breiman (2001a)]. Random forests fit trees to independent bootstrap samples from the data. The trees were grown large (for classification) and at each node independently, m predictors were chosen out of the p available, and the best possible split on these m predictors was used. As a default for classification, Breiman settled on choosing $m = \sqrt{p}$. In RF we see a synthesis of the bagging ideas (bootstrapping), along with ideas that came from boosting (growing large trees), and Breiman's understanding of how to increase instability (randomly choosing predictors at each node) to get more accurate aggregate predictions. Once he came up with RF, Breiman stopped working on new algorithms and started work on how to get the most out of the RF results. He developed measures of variable importance and proximities between observations. Together, we developed a program for visualizing and interpreting RF results (available from http://www.math.usu.edu/~adele/forests/cc_graphics.htm). Chao Chen and Andy Liaw worked with Breiman on ways to adjust RF for unbalanced classes [Chao, Liaw and Breiman (2004)]. Vivian Ng worked with him on detecting interactions [Ng and Breiman (2005)]. In his last technical report, Breiman showed consistency for a simple version of RF [Breiman (2004a)]. But the work on RF did not stop when Breiman died. Several extensions have been published; for example, Diaz-Uriarte and Alvarez de Andres (2006) developed a variable selection procedure, Meinshausen (2006) introduced quantile regression forests, and Hothorn et al. (2006), Ishwaran et al. (2008) considered forests for survival analysis. Although theory is still thin on the ground, Lin and Jeon (2006) showed that RF behaves like a nearest neighbor classifier with an adaptive metric and Biau, Devroye and Lugosi made some progress on consistency in a paper dedicated to Breiman's memory [Biau, Devroye and Lugosi (2008)]. Numerous applied articles have appeared and even a number of YouTube videos. I believe Breiman would be truly delighted at the popularity of the method.

10. Software. Leo developed his own code, invariably in fortran. I collaborated with him on the random forests fortran code and documentation http://www.math.usu.edu/~adele/forests/cc_home.htm. Andy Liaw and Matt Wiener developed an interface to R [Liaw and Wiener (2002)]. Although Leo supported the R release and admired the free-software philosophy of R, he regarded R as a tool for "Ph.D. statisticians" and he wanted his code to also be available with an easy to use graphical user interface (GUI). GUI-driven software for classification and regression trees and random forests is available from Salford Systems. Versions of trees, random forests and archetypes are available in R (packages rpart, randomForests [Liaw and Wiener (2002)], and archetypes [Eugster and Leisch (2009)]).

11. Textbooks. In addition to his papers, Breiman wrote three textbooks [Breiman (1968, 1969, 1973)], the first of which is in SIAM's "Classics of Mathematics" series. Perhaps even more impressive is the fact that other scholars are now writing texts that refer extensively to Breiman's work, including trees, bagging and random forests [see Berk (2008), Hastie, Tibshirani and Friedman (2009) and Izenman (2008)].

12. Philosophy. Breiman passionately believed that statistics should be motivated by problems in data analysis. Comments such as

> If statistics is an applied field and not a minor branch of mathematics, then more than 99% of the published papers are useless exercises. [Breiman (1995b)]

show how deeply he believed that statistics needed a change of direction. When he heard that Breiman (1998a) was to be published with discussion in *The Annals of Statistics*, he commented that "it would sure liven things up... maybe get some blood moving in the statistical main stream of asymptopia" (personal communication).

Although it is not widely cited, I believe Breiman's "Two Cultures" paper [Breiman (2001c)] is one of his most widely read, at least among statisticians. The paper contained Breiman's views about where the field was going and what needed to be done. To conclude, he said:

> 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. [Breiman (2001c)]

REFERENCES

Berk, R. (2008). *Statistical Learning from a Regression Perspective.* Springer, New York.

Biau, G., Devroye, L. and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.* **9** 2039–2057. MR2447310

Breiman, L. (1968). *Probability Theory.* Addison-Wesley, Reading, MA. [Republished (1991) in *Classics of Mathematics.* SIAM, Philadelphia, PA.] MR0229267

Breiman, L. (1969). *Probability and Stochastic Processes with a View Toward Applications.* Houghton Mifflin, Boston, MA. MR0254942

Breiman, L. (1973). *Statistics: With a View Toward Applications.* Houghton Mifflin Harcourt, Boston, MA. MR0359089

Breiman, L. (1984). Nail finders, edifices, and Oz. Technical Report 32. Dept. Statistics, Univ. California, Berkeley, CA. Neyman–Kiefer Memorial Volume.

Breiman, L. (1991). The Π method for estimating multivariate functions from noisy data. *Technometrics* **33** 125–143. MR1110355

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *J. Amer. Statist. Assoc.* **87** 738–754. MR1185196

Breiman, L. (1993a). Fitting additive models to regression data: Diagnostics and alternative views. *J. Comput. Statist. Data Anal.* **15** 13–46. MR1202297

Breiman, L. (1993b). Hinging hyperplanes for regression, classification, and function approximation. *IEEE Trans. Inform. Theory* **39** 999–1013. MR1237723

Breiman, L. (1994). The 1991 census adjustment: Undercount or bad data? *Statist. Sci.* **9** 458–537.

Breiman, L. (1995a). Better subset regression using the nonnegative garrote. *Technometrics* **37** 373–384. MR1365720

Breiman, L. (1995b). Reflections after refereeing papers for NIPS. In *The Mathematics of Generalization: Proceedings of the SFI/CNLS Workshop on Formal Approaches to Supervised Learning, Volume 1992* (D. H. Wolpert, ed.). Westview Press, Boulder, CO. MR1353248

Breiman, L. (1996a). Stacked regressions. *Mach. Learn.* **24** 49–64.

Breiman, L. (1996b). Heuristics of instability and stabilization in model selection. *Ann. Statist.* **24** 2350–2383. MR1425957

Breiman, L. (1996c). Bagging predictors. *Mach. Learn.* **24** 123–140.

Breiman, L. (1997a). Arcing the edge. Technical Report 486. Dept. Statistics, Univ. California, Berkeley, CA.

Breiman, L. (1997b). Out-of-bag estimation. Technical report. Dept. Statistics, Univ. California, Berkeley, CA.

Breiman, L. (1998a). Arcing classifiers. *Ann. Statist.* **26** 801–849. MR1635406

Breiman, L. (1998b). Using convex pseudo-data to increase prediction accuracy. Technical Report 513. Dept. Statistics, Univ. California, Berkeley, CA.

Breiman, L. (1998c). Half & half bagging and hard boundary points. Technical Report 534. Dept. Statistics, Univ. California, Berkeley, CA.

Breiman, L. (1999a). Prediction games and arcing algorithms. *Neural Comput.* **11** 1493–1517.

Breiman, L. (1999b). Pasting small votes for classification in large databases and on-line. *Mach. Learn.* **36** 85–103.

Breiman, L. (2000a). Randomizing outputs to increase prediction accuracy. *Mach. Learning* **40** 229–242.

Breiman, L. (2000b). Some infinity theory for predictor ensembles. Technical Report 577. Dept. Statistics, Univ. California, Berkeley, CA.

Breiman, L. (2001a). Random forests. *Mach. Learn.* **45** 5–32.

Breiman, L. (2001b). Using iterated bagging to debias regressions. *Mach. Learn.* **45** 261–277.

Breiman, L. (2001c). Statistical modeling: The two cultures. *Statist. Sci.* **16** 199–231. MR1874152

Breiman, L. (2004a). Consistency for a simple model of random forests. Technical Report 670. Dept. Statistics, Univ. California, Berkeley, CA.

Breiman, L. (2004b). Population theory for boosting ensembles. *Ann. Statist.* **32** 1–11. MR2050998

Breiman, L. and Cutler, A. (1993). A deterministic algorithm for global optimization. *Math. Programming* **58** 179–199. MR1216490

Breiman, L. and Freedman, D. (1983). How many variables should be entered in a regression equation? *J. Amer. Statist. Assoc.* **78** 131–136. MR0696857

Breiman, L. and Friedman, J. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–598. MR0803258

Breiman, L. and Friedman, J. (1988). Comment on "Tree-structured classification via generalized discriminant analysis" by W. Y. Loh and N. Vanichsetakul. *J. Amer. Statist. Assoc.* **83** 725–727. MR0963799

Breiman, L. and Friedman, J. (1997). Predicting multivariate responses in multiple linear regression. *J. Roy. Statist. Soc. Ser. B* **59** 3–54. MR1436554

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees.* Wadsworth, New York. MR0726392

Breiman, L. and Ihaka, R. (1984). Nonlinear discriminant analysis via scaling and ACE. Technical Report 40. Dept. Statistics, Univ. California, Berkeley, CA.

Breiman, L. and Meisel, W. S. (1976). General estimates of the intrinsic variability of data in nonlinear regression models. *J. Amer. Statist. Assoc.* **71** 301–307.

Breiman, L., Meisel, W. S. and Purcell, E. (1977). Variable kernel estimates of multivariate densities. *Technometrics* **19** 135–144.

Breiman, L. and Peters, S. (1992). Comparing automatic smoothers (A public service enterprise). *Int. Statist. Rev.* **60** 271–290.

Breiman, L. and Spector, P. (1992). Submodel selection and evaluation in regression. The X-random case. *Int. Statist. Rev.* **60** 291–319.

Breiman, L., Tsur, Y. and Zemel, A. (1993). On a simple estimation procedure for censored regression models with known error distributions. *Ann. Statist.* **21** 1711–1720. MR1245765

Breiman, L. and Wurtele, Z. S. (1964). Convergence properties of a learning algorithm. *Ann. Math. Statist.* **35** 1819–1822. MR0178995

Bühlmann, P. and Yu, B. (2003). Boosting with the L2 loss: Regression and classification. *J. Amer. Statist. Assoc.* **98** 324–339. MR1995709

Bühlmann, P. and Yu, B. (2006). Sparse boosting. *J. Mach. Learn. Res.* **7** 1001–1024. MR2274395

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models. *Ann. Statist.* **17** 453–510. MR0994249

Chao, C., Liaw, A. and Breiman, L. (2004). Using random forests to learn imbalanced data. Technical Report 666. Dept. Statistics, Univ. California, Berkeley, CA.

Cutler, A. and Breiman, L. (1994). Archetypal analysis. *Technometrics* **36** 338–347. MR1304898

Diaz-Uriarte, R. and Alvarez de Andres, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7** 3.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR0515681

Efron, B. (2001). Comment on "Statistical modeling: The two cultures" by L. Breiman. *Statist. Sci.* **16** 218–219. MR1861072

Eugster, M. J. A. and Leisch, F. (2009). From Spider-Man to Hero—archetypal analysis in R. *J. Statist. Soft.* **30** 1–23.

Freund, A. (1995). Boosting a weak learning algorithm by majority. *Inform. Comput.* **121** 256–285. MR1348530

Freund, Y. and Schapire, R. (1996). Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference* 148–156. Morgan Kauffman, San Francisco, CA.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statist. Sci.* **1** 297–310. MR0858512

Hastie, T., Tibshirani, R. and Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *J. Amer. Statist. Assoc.* **89** 1255–1270. MR1310220

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2000). Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Statist.* **28** 337–407. MR1790002

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, New York. MR1851606

HOTHORN, T., BÜHLMANN, P., DUDOIT, S., MOLINARO, A. and VAN DER LAAN, M. (2006). Survival ensembles. *Biostatistics* **7** 355–373.

ISHWARAN, H., KOGALUR, U. B., BLACKSTONE, E. H. and LAUER, M. S. (2008). Random survival forests. *Ann. Appl. Statist.* **2** 841–860. MR2516796

IZENMAN, A. (2008). *Modern Multivariate Statistical Techniques.* Springer, New York. MR2445017

JIANG, W. (2004). Process consistency for AdaBoost. *Ann. Statist.* **32** 13–29. MR2050999

LIAW, A. and WIENER, M. (2002). Classification and regression by random forest. *R News* **2** 18–22.

LIN, Y. and JEON, Y. (2006). Random forests and adaptive nearest neighbors. *J. Amer. Statist. Assoc.* **101** 578–590. MR2256176

MEINSHAUSEN, N. (2006). Quantile regression forests. *J. Mach. Learn. Res.* **7** 983–999. MR2274394

MINSKY, M. and PAPERT, P. (1969). *Perceptrons: An Introduction to Computational Geometry.* MIT Press, Cambridge, MA.

NG, V. W. and BREIMAN, L. (2005). Bivariate variable selection for classification problem. Technical Report 692. Dept. Statistics, Univ. California, Berkeley, CA.

OLSHEN, R. (2001). A conversaton with Leo Breiman. *Statist. Sci.* **16** 184–198. MR1861072

SCHAPIRE, R. (1990). The strength of weak learnability. *Mach. Learn.* **5** 197–227.

SCHAPIRE, R., FREUND, Y., BARTLETT, P. and LEE, W. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Ann. Statist.* **26** 1651–1686. MR1673273

SHANG, N. and BREIMAN, L. (1996). Distribution based trees are more accurate. In *Proceedings of the Int. Conf. on Neural Information Processing, Hong Kong* 133–138. Springer, Singapore.

SMITH, P. (1982). Curve fitting and modeling with splines using statistical variable selection techniques. NASA Report 166034. NASA, Langley Research Center, Hampton, VA.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

WOLPERT, D. (1992). Stacked generalization. *Neural Networks* **5** 241–259.

ZHANG, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.* **32** 56–134. MR2051001

DEPARTMENT OF MATHEMATICS
AND STATISTICS
UTAH STATE UNIVERSITY
UMC 3900
LOGAN, UTAH 84322-3900
USA
E-MAIL: adele.cutler@usu.edu