The Two Cultures:
Statistics and Machine Learning in Science

R. Kass, January 2021

In the 1990s, Leo Breiman noticed an identifiable distinction between the way statisticians and computer scientists analyze data, and it bothered him. It bothered him a lot. There were two cultures: the computer scientists focused on algorithms and prediction, largely ignored traditional statistics, and plunged ahead with inventive methods, often succeeding. Breiman recognized flaws in the more traditionally-oriented statistical culture and, in his 2001 *Statistical Science* paper, advocated for greater emphasis on the predictive approach. This is perhaps not surprising, given his description of his personal trajectory, where an early, formative project consulting for the EPA (after giving up his successful career as a probabilist) had him predicting next-day ozone concentrations based on hundreds of variables. In any case, Breiman suggested that results based on models were frequently over-interpreted. He also warned that, in comparison to their predictively-oriented counterparts, traditional statisticians were often too timid.

My reaction to the paper now is almost the same as it was 20 years ago. Breiman's observations, which I just summarized, were important, and timely, but I did not buy his analysis of them: by making prediction primary and explanation secondary, he was using one component of science to mischaracterize the whole; and his fundamental criticism of statistical models, which was based on the unrealizable standard that they should represent well-established mechanisms, willfully ignored best statistical practices. His article thus struck me as an odd combination of wise and narrow-minded, so I was reassured to read the comments of two of my heroes, David Cox and Brad Efron, who seem to have reacted similarly. Efron began, "At first glance, Leo Breiman's stimulating paper looks like an argument against parsimony and scientific insight, and in favor of black boxes with lots of knobs to twiddle. At second glance, it still looks that way."

Because it did say important things, and possibly also because of its iconoclastic outlook and tone, Breiman's paper has achieved landmark status, but the disagreeable parts still seem, to me, to be sending bad messages.

Machine Learning and Artificial Intelligence

As Breiman recognized, and I repeated above, there are identifiable differences in the approaches to data analysis among those trained in statistics and those trained in computer science. Still, I like to say that machine learning is the intersection of statistics and computer science, emphasizing it is entirely part of statistics and entirely part of computer science. An important way the world has changed is that each of the disciplines of statistics and computer science has incorporated big parts of what the other field has to offer. As a result, machine learning is thriving in many good ways, and the penetration of statistics into computer science should be considered a victory.

I have a favorite anecdote from about 15 years ago, around the time that the Machine Learning Department was established here at Carnegie Mellon. We held a retreat to find out what we all were working on, and where we could find common ground. At some point during our discussions, my computer scientist colleague Roni Rosenfeld announced, and I'm paraphrasing, "I've figured out the difference between statisticians and computer scientists: statisticians attack problems with 10 parameters and try to get it right; computer scientists attack problems with 10 million parameters and try to get an answer." I thought it was an insightful comment, very much in the spirit of Breiman's article, but I like to add that now we are all attacking large problems *and* trying to get it right.

In 2017, the dean of our School of Computer Science decided to create a website devoted to artificial intelligence at Carnegie Mellon, and he emailed many faculty to ask if they would be willing to be listed there. I replied in the affirmative but added, to his assistant I assumed, that I was about as far from artificial intelligence as someone in machine learning could be. The dean then wrote me back with a good-humored explicative version of "nonsense," saying, "We all know that $\mathrm{AI} = \text{ Statistics } + \text{ Optimization}$." His comment occurred just as deep neural networks were becoming dominant, and that abbreviated description of artificial intelligence may now seem limited. On the other hand, statisticians and computer scientists are keenly aware of the strengths and shortcomings of deep neural networks, and many are actively working to understand them better.

Outside the community of experts, however, there is an all-too-familiar lack of appreciation for what machine learning in general, and deep neural networks in particular, can and can not do, manifested often in an apparent feeling that sprinkling some machine learning dust on a problem in data analysis will magically produce great results. The abuse of machine learning approaches is every bit as bad as the abuse of classical statistical methods such as linear regression, and sometimes in much the same manner. Contrary to what Breiman intimated, predictively-oriented approaches have not cured the wishful analysis disease.

Excessive Cautiousness in Statistics

Breiman felt compelled to follow a grand tradition in statistics: hand-wringing. Hand-wringing in statistics has two kinds of concerns: abuse of methods, and detachment from real-world problems. The former is sometimes accompanied by suggested remedies within introductory courses; the latter is often coupled with advice about advanced training. I myself have indulged in this pastime (Brown and Kass, 2009; Kass, 2011).

Breiman was especially worried about the timidity of statisticians. Emery Brown and I (Brown and Kass, 2009) adopted the cultural characterization in discussing the problem from the perspective of graduate training:

> Somehow, in emphasizing the logic of data manipulation, teachers of statistics are instilling excessive cautiousness. Students seem to develop extreme risk aversion, apparently fearing that the inevitable flaws in their analysis will be discovered and pounced upon by statistically trained colleagues. Along with communicating great ideas and fostering valuable introspective care, our discipline has managed to create a culture that often is detrimental to the very efforts it aims to advance.

Like Breiman and others, we advocated greater involvement with real problems, though we stressed collaboration as opposed to consulting. While I still agree with Breiman on this point, I would not want revolutionary change: we continue to need a discipline devoted to the details.

The Dangers of Statistical Modeling

In a 1983 paper, DuMouchel and Harris suggested a Bayesian approach to combining information from diverse studies of cancer risk (DuMouchel and Harris, 1983). The paper was both interesting and bold: they applied a hierarchical two-way ANOVA model, which seems very simple by today's standards but was cutting-edge back then, and they dared to put on the same footing 5 disease assessments, from human lung cancer, to skin cancer in mice, to mutagenesis in mouse cell lines, and also 9 carcinogens ranging from roofing tar, to diesel emissions, to cigarette smoke. Their 5×9 array of data, consisting of individual-study risk assessments and standard errors, had empty cells for the risks of human lung cancer due to 6 of the carcinogens. Their fundamental goal was to provide an improved method of extrapolating to humans (for those 6 empty cells) the information provided by non-human carcinogen risk assessments.

Along with two others, I was invited to comment on the article (Kass, 1983). Their study remains interesting, to me, because it highlights the strengths and limitations of modeling efforts, the Bayesian aspect being less important than the modeling (though, again, back then the use of “Bayes" was, inevitably, emphasized). I wrote:

> The great benefit of [their] approach to this problem is that it makes precise the assessments of relevance and uncertainty in related results. ... Yet, [it] has its difficulties. ... Lurking beside each analysis are the interrelated dangers of oversimplification, overstated precision, and neglect of beliefs other than the analyst’s.

That criticism of models is consistent with Breiman’s, but I did not, and do not, see the lurking presence of these dangers as a reason to abandon such modeling efforts altogether. After expanding on my concerns—some of which had to do with formalization of a risk-management policy—I added that “I would hesitate to apply the model without additional theoretical or empirical knowledge." Nonetheless, my judgment about the shortcomings of their effort did not diminish my appreciation for the work of DuMouchel and Harris, and I ended my comment with a congratulatory sentiment because, under the right circumstances, their approach could be illuminating.

What Can Statistical Models Accomplish?

The pragmatic defense against the “lurking dangers" in statistical models is to work at identifying them and figuring out the likely magnitude of their effects. An excellent illustration is the great case study by Mosteller and Wallace of the Federalist papers (Mosteller and Wallace, 1964).

The Federalist was a collection of 85 political papers written collectively by Alexander Hamilton, John Jay, and James Madison, under the pseudonym Publius, arguing for ratification of the U.S. Constitution. It remains an important work on the political philosophy of the Constitution’s framers, but, because the papers were not individually signed, it has also remained somewhat uncertain who wrote each one. Prior to the work of Mosteller and Wallace, historians generally agreed on the authorship of 70 papers, and there were 3 that were apparently co-authored in some way, but the authorship of the other 12 was disputed: for each, it was either Hamilton or Madison. Mosteller and Wallace set out to resolve the dispute based on usage of selected words.

Their method was to analyze word counts within blocks of text for 165 selected words. They began with a Poisson assumption, but settled on negative binomial for the bulk of their main study. Thus, for each paper, and each of the two possible authors, each word count distribution had 2 free parameters. After considerable exploratory work, and a helpful reparameterization, they found tractable priors, based on 5 hyperparameters, which assumed usage rates were independent across words but allowed for correlation in usage rates by Hamilton and Madison. They used the 70 known-authorship papers to estimate the parameters (165×4 parameters) and developed methods to compute, for each of the 12 disputed-authorship papers, the Bayes factor in favor of Madison versus Hamilton (Kass and Raftery, 1996). Their results settled the matter, in favor of Madison, decisively for 10 of the 12 papers, and strongly for the other 2.

The care that went into Mosteller and Wallace's development of their main model was already exemplary. They then considered whether particular words might have had excessive effects, whether problems with the data were likely, what the effects of correlation across words would be, and whether their results were consistent with those from simpler methods, which did not involve their distributional assumptions. They also used other data, from subsequent writing, for many of their supplementary analyses. Thus, the totality of their work required a book-length treatment. Though they even devoted two pages to the possibility of fraud, or gross errors, in the end, they felt their very large Bayes factors were "believable."

Although we can't require every model-based statistical study to be as thorough and careful as that of Mosteller and Wallace, we should expect them all to follow roughly the same kinds of steps. As my colleague Valérie Ventura says, we tell our PhD students to find out what it takes to "break" the model. We then evaluate the extent to which an analysis survives, and reconsider our conclusions.

Notice, though, that in summarizing the results of the Mosteller and Wallace study I did not report the values of their Bayes factors. Instead, I simply used the qualitative descriptors "decisive" and "strong." While this follows advice in Kass and Raftery (1996) that is specific to Bayes factors, in practice, the move from quantitative to qualitative is far more general, as I said in commenting on a paper by Brad Efron (Kass, 2010):

> In the scientific applications I am familiar with, statistical inferences are important, even crucial, but they constitute intermediate steps in a chain of inferences, and they are relatively crude. ... Furthermore, statistical uncertainty is typically smaller than the unquantified aggregate of the many other uncertainties in a scientific investigation. ... To be convincing, the science needs solid statistical results, but in the end only a qualitative summary is likely to survive.

> For instance, in Olson et al. (2000), my first publication involving analysis of neural data, more than a dozen different statistical analyses—some of them pretty meticulous, involving both bootstrap and MCMC—were reduced to the main message that among 84 neurons recorded from the supplementary eye field, “Activity reflecting the direction of the [eye movement] developed more rapidly following spatial than following pattern cues.” The statistical details reported in the paper were important to the process, but not for the formulation of the basic finding.

I added, with reference to Jeffreys (see also Kass, 2009):

> If science is such a loose and messy process, and inferences so rough and approximate, where does all the statistical effort go? In my view, Jeffreys got it right. State-of-the-art analyses may take months, but they usually come down to estimates and standard errors.

In other words, we must judge models by their rough and approximate inferences in the context of scientific problems that are richer, or messier, than our idealized settings.

We can, however, gain something by continuing to think the way we do, statistically, about these situations. For instance, commenting on Breiman, Efron said, “Following Fisher’s lead, most of our current statistical theory and practice revolves around unbiased or nearly unbiased estimates (particularly MLEs)," and he went on to say that good experimental design helps produce favorable conditions for this statistical presumption. Thus, when I said, as quoted above, “statistical uncertainty is typically smaller than the unquantified aggregate of the many other uncertainties in a scientific investigation," I should have acknowledged, explicitly, that “the many other uncertainties" must include adjustment for bias: in assessing results, I mostly think about potential sources of bias, and I mentally adjust statistical inferences by accounting for them, in the spirit of Mosteller and Wallace, and “breaking the model."

With these limitations of statistical inference in mind, we can adopt a neo-Jeffreys attitude: statistical models should provide trustworthy results in the sense that they get us into the right ballpark with their estimates and standard errors (and significance tests or Bayes factors).

The Statistical Paradigm

In my view, Breiman was right to put statistical models at the center of modern statistics. However, a pair of subtleties in Breiman's framework could easily be missed. First, there are two distinct aspects of what he calls algorithmic modeling: an algorithmic approach to analysis, and assessment using prediction. It is worth considering them separately. Second, it is not clear when statistical models can be used within the algorithmic culture; apparently, sometimes.

Some years ago, a computer scientist and his PhD student came to me for advice about analyzing some neural data. As they were describing what they had already done I had to stop them, to inquire how they had fit the data. They said they had used EM. I couldn't help suggesting alternative verbiage: they had applied EM to implement maximum likelihood for fitting their model to the data. This is a nice example of the contrast between algorithmic and statistical thinking. The contrast is very important because, in my observation, and in the observation of my book co-authors Emery Brown and Uri Eden (Kass et al., 2014, page 2), exclusive reliance on algorithmic thinking often deprives quantitative analysts of the most fundamental ways that statistics could enrich their approach:

> Many researchers have excellent quantitative skills and intuitions, and in most published work statistical procedures appear to be used correctly. Yet, in examining these papers we have been struck repeatedly by the absence of what we might call statistical thinking, or application of the statistical paradigm, and a resulting loss of opportunity to make full and effective use of the data. These cases typically do not involve an incorrect application of a statistical method (though that sometimes does happen). Rather, the lost opportunity is a failure to follow the general approach to the analysis of the data, which is what we mean by the label "the statistical paradigm."

So, what is this statistical paradigm? In our book we devoted 14 pages to a concise summary of it, including standard components (statistical modeling is an iterative process that incorporates assessment of fit and is preceded by exploratory analysis) and some teachings that are often not explicitly emphasized (important data analytic ideas are sometimes implemented in different ways; measuring devices often pre-process the data; data analytic techniques are rarely able to compensate for deficiencies in data collection; simple methods are essential). Following Brown and Kass (2009) and Kass (2011), we began by saying,

> After numerous conversations with colleagues, we have arrived at the conclusion that among many components of the statistical paradigm, summarized

below, two are the most fundamental:

1. Statistical models are used to express knowledge and uncertainty about a signal in the presence of noise, via inductive reasoning.

2. Statistical methods can be analyzed to determine how well they are likely to perform.

We quickly pointed out that the statistical models we were talking about could be either parametric or nonparametric. In Brown and Kass (2009), similarly to Breiman, we focused on nonparametric regression as an archetype, supplying the usual model

$$Y_i = f(x_i) + \varepsilon_i,$$

and said,

> One can dream up a smoothing method, and apply it, without ever referencing a model—indeed, this is the sort of thing that we witness and complain about in neuroscience. Meanwhile, among statisticians there is no end of disagreement about the details of a model and the choice among methods (What space of functions should be considered? Should the ε_i random variables enter additively? Independently? What class of probability distributions should be used? Should decision theoretic criteria be introduced, or prior probabilities?). The essential component that characterizes the discipline is the introduction of probability to describe variation in order to provide a good solution to a problem involving the reduction of data for a specified purpose. This is not the only thing that statisticians do or teach, but it is the part that identifies the way they think.

Again, Breiman was right in his characterization, except that he seemed to be limiting his criticism to parametric modeling which, from a modern perspective, distorts what we should consider "statistical thinking" and, thus, the statistical modeling culture.

Breiman tied prediction to algorithmic thinking with cross-validation. He was, again, correct in a couple of senses. First, a fundamental feature of the computer science-based culture he was describing is its emphasis on prediction and, second, the way we judge predictive accuracy is to apply cross-validation. This, however, makes a strong assumption, which is appropriate in many technological, business, and policy settings, but is staggeringly strange in most scientific settings, namely, that the test data are statistically equivalent to scientifically relevant new data. As Cox said in his comment, "Often the prediction is under quite different conditions." That is, as we wrote in Kass et al. (2016), "The scientific results that stand the test of time are those that get confirmed across a variety of different, but closely related, situations." We wrote this in describing the value of replication, adding that commonly-occurring discrepancies in the details of data collection and analysis across scientific studies confers a degree of robustness to the findings. In neuroscience, this could involve different kinds of experiments, different recording devices, and even different species of animals. Prediction in the sense of cross-validation remains an important tool, but it is unable to supply the sole support for scientific inference.

What Do Statistical Models Say?

Having recognized the dangers in model-based inference, Breiman ignored the pragmatic response and, instead, simplified his argument for impugning the value of modeling by invoking chance mechanisms; he used the word "mechanism" 14 times. He could have acknowledged that statistical models aim to describe regularity and variability in the data for a problem-specific purpose, but he didn't. Consistently with his Berkeley colleague David Freedman, Breiman took models to be making claims about the way the data were generated, which limits their application and imposes an unrealistically high burden of proof on the modeler. My irritation with this way of thinking about models, and all it implies for teaching as well as research, was a major motivation for Kass (2011). There, I emphasized the conceptual distinction between mathematical models, which live in what I called the "theoretical world," and data, which are in the "real world." I wrote,

> The problem with the chance-mechanism conception is that it applies to a rather small part of the real world, where there is either actual random sampling or situations described by statistical or quantum physics. I believe the chance-mechanism conception errs in declaring that data are assumed to be random variables, rather than allowing the gap [between the theoretical and real worlds]. In saying this I am trying to listen carefully to the voice in my head that comes from the late David Freedman (see Freedman and Zeisel, 1988). I imagine he might call crossing this bridge [from the theoretical to the real world], in the absence of an explicit chance mechanism, a leap of faith. In a strict sense I am inclined to agree. It seems to me, however, that it is precisely this leap of faith that makes statistical reasoning possible in the vast majority of applications.

In that article I claimed to be describing "the dominant contemporary philosophy of statistics." Certainly I was in good company. For example, Mosteller and Wallace, agreeing with some previous authors that models should be considered, at best, approximations, completed the thought by saying (Mosteller and Wallace, 1964, page 49),

> What is good enough for all the rest of applied mathematics is going to have to be good enough for statistical inference. ... [One chooses a model, then] proceeds to use it freely without further question, even though [we know] that it is not, and cannot be, exactly right.

I like to say (Brown and Kass, 2009; Kass et al., 2014) that the quintessential statistical attitude is captured well by the famous dictum of George Box, “All models are wrong, but some are useful."

Methods That Work Well in Practice

Dozens of experiments have found that, when a grating of alternating black and white bars of light moves across the visual field of a primate, many of its neurons in primary visual cortex become active and, furthermore, they do so preferentially for gratings in particular orientations. This is a highly replicated finding.

A multitude of laboratories have reported such orientation-tuned neurons (usually in passing, as part of a study with more elaborate goals) because it is a phenomenon that becomes part of, or can be explained by, theories of early visual processing. In other words, it is used repeatedly to gain insight. In statistics, too, we want repeated insight: although it is very hard to define, insight is at the heart of what we aim for, and methods become accepted as they are used repeatedly to achieve scientific goals.

Statistics, however, is fundamentally different from empirical sciences in two respects. First, we never get evidence to say that a particular analysis of a given set of data was justified; we never find out, in a strong scientific sense, that a method worked well. In statistics, how would we build a situation analogous to replication of orientation tuning? We would want to say that some method produces qualitatively similar results when applied by different statisticians to some class of data sets. But what do we mean by qualitatively similar results? And how do we form the appropriate class of data sets that would be analogous to the primary visual cortex of primates? We must accept this fundamental limitation of statistical methods.

On the other hand, statistics is also unique because we are able to construct ground truth scenarios and use them to investigate our methods. When, in his famous 1962 paper, “The

Future of Data Analysis," John Tukey urged statisticians to think of their field as experimental, he meant that they should investigate potentially unfavorable situations by using simulation studies to replace intractable mathematical theory (Tukey, 1962). This ability to apply both mathematical theory and simulation in order to understand the behavior of data analytic procedures is a huge strength, allowing us to try to "break the model," thereby giving us confidence in results: it is the second of the two items identified as "most fundamental" in my summary of the statistical paradigm, above. But this activity itself relies on statistical models (for theory and simulations), and it can only have value if we are willing to take the "leap of faith" I claimed was part of the "dominant contemporary philosophy of statistics."

In this context of judging statistical methods, particularly those that are novel and complicated, it is also important to be willing to stand by conclusions drawn from them. This distinguishes what we usually mean by a case study from a typical example found in a methods paper. Mosteller and Wallace summed up (Mosteller and Wallace, 1964, page 2), as follows:

> The main value of our work does not depend much upon one's view of the authorship question used as a vehicle for this case study, although we do add information there. Rather the value resides in the illustrative use of the various techniques and generalizations that emerge from their study. In retrospect, the methodological analysis could have been restricted to sets of papers whose authors are known. Still, *the responsibility of making judgments about authorship in disputed cases adds a little hard realism and promotes care that might otherwise have been omitted.* [Italics added]

Conclusion

Breiman's paper succeeded in calling attention to "cultural" issues that separated two clans of data analysts at the turn of the millennium. Like all good alarms, the paper's most important feature was loudness, and it blasted at about the right time. As for appropriate responses, however, I have sided largely with dissenting comments by Cox and Efron, and I have tried here to clarify many aspects of Breiman's subject that ought to be brought into the conversation. It is a conversation worth continuing. Every serious student, who has even a passing interest in the interface between statistics and computer science, should grapple with these foundational ideas.

References

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

Brown, E. N. and Kass, R. E. (2009). What is statistics? (with discussion). *American Statistician*, 63:105–123.

DuMouchel, W. and Harris, J. (1983). Bayes methods for combining the results of cancer studies in humans and other species (with discussion). *Journal of the American Statistical Association*, 78:293–314.

Freedman, D. and Zeisel, H. (1988). From mouse to man: The quantitative assessment of cancer risks (with discussion). *Statistical Science*, 3:3–56.

Kass, R. E. (1983). Bayes methods for combining the results of cancer studies in humans and other species: Comment. *Journal of the American Statistical Association*, 78:312–313.

Kass, R. E. (2009). The importance of Jeffreys's legacy (comment on Robert, Chopin, and Rousseau). *Statistical Science*, 24(2):179–182.

Kass, R. E. (2010). Comment: How should indirect evidence be used? *Statistical Science*, 25:166–169.

Kass, R. E. (2011). Statistical inference: The big picture (with discussion). *Statistical Science*, 26:1–20.

Kass, R. E., Caffo, B., Davdian, M., Meng, X.-L., Yu, B., and Reid, N. (2016). Ten simple rules for effective statistical practice. *PLoS Computational Biology*, 12:e1004961.

Kass, R. E., Eden, U. T., and Brown, E. N. (2014). *Analysis of Neural Data*. Springer.

Kass, R. E. and Raftery, A. E. (1996). Bayes factors. *Journal of the American Statistical Society*, 90:773–795.

Mosteller, F. and Wallace, D. L. (1964). *Inference & Disputed Authorship: The Federalist*. Republished in 1984 as *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*, Springer.

Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics*, 33:1–67.