Review of *Statistical Modeling: The Two Cultures*

Grace Yi Chen

In this paper, the author introduced two cultures: the data modeling culture that assumes the data are generated by a given model, and the algorithmic modeling culture which treats the data mechanism as unknown and try to use a black box to model it. The goals of doing statistics are to use data to predict and get information about the underlying data mechanism. Depending on different problems and data, he believes that statisticians should be more open to different methods that would be the most appropriate for each problem. He argued that data modeling culture adopted by most of the statisticians may not be practical and suitable for current real-world problems, especially with the increase of dimensionality and complexity of the data generating mechanisms. So instead of focusing on the model interpretability, he advocated that we should focus on the model predictive accuracy. In addition, 4 professors gave commentaries to this paper and Breiman also gave response. Manny Parzen and Bruce Hoadley are more in agreement with Brad Efron and D. R. Cox having reservations.

I think this paper is well-written and Breiman clearly makes his argument about the two cultures in statistical modeling. I agree with him that new analysis models should be developed to solve the increasing complex problems. When the problem is about prediction, prediction accuracy might be a better metric to evaluate models instead of evaluations based on goodness of fit or unbiasedness of the model coefficients. We should not constrain ourselves with familiar tools and problems when facing these new problems. However, I also agree with Efron and Cox that one of our ultimate goals is to understand the nature mechanism in the black boxes and we should aim to understand more and more about the natural black boxes. In addition, I agree with Cox that statistical theoretical journals like JASA and Biometrics should focus more on the technique rather than the explanation of the applied analysis process. Moreover, Breiman argued that data modeling culture suffer from Rashomon as models with similar RSS may tell different stories about the inherited nature mechanism. This is also true for algorithmic modeling methods because these methods are hard to honestly compare based on simulations with many parameters to tune.

Question:
When both data modeling and algorithmic modeling methods work, how should we compare them, and which one should we prefer?