1st Reading for COMP6321 Machine Learning

Prior to the Midterm 1 you must the read and understand the main arguments of the following classic paper:

- Leo Breiman (2001). **Statistical modeling: The two cultures.** *Statistical science*, 16(3):199–231. (**PDF**)

The paper was written by Leo Breiman (1928–2005), inventor of bootstrap aggregation ('bagging'), random forests, and many other practical techniques.

The Breiman paper is conceptual. The target audience is statisticians. It is a critique of an analytical approach that is common in statistics. Breiman refers to technical and non-technical concepts, but there is no real math. Several of the concepts referred to will only be **covered in Lectures 4 and 5**, but many of the arguments and examples should be understandable beforehand.

Requirements

You must read the sub-sections indicated below.

Required?	Section
Yes	ABSTRACT
Yes	1. INTRODUCTION
Yes	2. ROAD MAP
Yes	3. PROJECTS IN CONSULTING
Yes	3.1 The Ozone Project
Yes	3.2 The Chlorine Project
Yes	3.3 Perceptions on Statistical Analysis
Yes	4. RETURN TO THE UNIVERSITY
Yes	4.1 Statistical Research
Yes	5.0 THE USE OF DATA MODELS
Yes	5.1 An Example
Yes	5.2 Problems in Current Data Modeling
Yes	5.3 The Multiplicity of Data Models
Yes	5.4 Predictive Accuracy

Required?	Section
Yes	6.0 THE LIMITATIONS OF DATA MODELS
Yes	7.0 ALGORITHMIC MODELING
Yes	7.1 A New Research Community
Yes	7.2 Theory in Algorithmic Modeling
Yes	7.3 Recent Lessons
Yes	8.0 RASHOMON AND THE MULTIPLICITY OF GOOD MODELS
Yes	9.0 OCCAM AND SIMPLICITY VS. ACCURACY
Yes	9.1 Growing Forests for Prediction
Yes	9.2 Forests Compared to Trees
Yes	9.3 Random Forests are A+ Predictors
Yes	9.4 The Occam Dilemma
Yes	10.0 BELLMAN AND THE CURSE OF DIMENSIONALITY
Yes	10.1 Digging It Out in Small Pieces
No	~~10.2 The Shape Recognition Forest~~
Yes	10.3 Support Vector Machines
Yes	11.0 INFORMATION FROM A BLACK BOX
No	~~11.1 Example I~~
No	~~11.2 Example II~~
No	~~11.3 Example III~~
Yes	11.4 Remarkes about the Examples
Yes	12.0 FINAL REMARKS
Yes	GLOSSARY
No	~~ACKNOWLEDGEMENTS~~
No	~~Comment by D. R. Cox~~
No	~~Comment by Brad Efron~~
No	~~Comment by Bruce Hoadley~~

Required?	Section
No	~~Comment by Emanuel Parzen~~
No	~~Rejoinder by Leo Breiman~~

Reading the "Comments" and the "Rejoinder" is optional, but very informative. They present a rare opportunity to see top minds in the field debate the merits of statistical and algorithmic approaches.

It might be interesting to know that the "Dempster" that Brieman ridicules in the final remarks is Arthur P. Dempster, co-inventor of the EM algorithm that we studied in class.

Expectations

You will see many unfamiliar terms in the manuscript. That is OK, you do not need to chase down technical definitions for every concept mentioned. Any important terminology will be mentioned in class. Terms like "cross-validation" you will understand by the time of Midterm 1.

However, you should strive to understand the *arguments* that Breiman is making. For example, if you notice the term "residual analysis" you may think "Oh no! Are we expected to know how to do 'residual analysis'?" The answer is *no*, you will not be expected to know what residual analysis is just from reading this paper, but you should take it upon yourself to learn the *purpose* of residual analysis, which is enough to understand Breiman's argument.

Example midterm questions related to this reading

- How are "data models" typically validated? How are "algorithmic models" typically validated?
- Give three examples of what Breiman calls "data models" and three examples of what he calls "algorithmic models"
- After having worked as a consultant, what were Breiman's "perceptions" about how to work with data?
- How was predictive accuracy measured in Breiman's "Ozone project"?
- Describe the modeling approach that Breiman's team used in the "Chlorine project"
- Describe an example where theory in algorithmic modeling led to an important advance.
- What learning algorithm does Breiman rate as "A+ for prediction" but "F for interpretability", and what are his reasons?
- State 'Occam's dilemma' as Breiman describes it.
- Describe the main symptom of 'model instability'.
- What is the "straight jacket" that Breiman claims statisticians are imposing on themselves? Why does it matter?
- What is Breiman's argument against doing dimensionality reduction?