Rich Gruss
Week 2 Paper

Chapter 1: The Data Science Process

Zumel and Mount describe the organizational structure of a successful data science project, defining roles and stages, while acknowledging that these are fluid and variable. As with software projects, the business sponsor sets the goals and decides whether the project is a success or failure. Some important principles explained in the chapter: 1) Stopping conditions are important. If your goal is too vague, it's difficult to know when your analysis is finished. 2) Typical modeling tasks include classification, scoring, ranking, clustering, finding relations, and characterization. 3) Models are tested by generating a "confusion matrix," which applies the model to new data and compares actual outcomes with predicted outcomes. 4) A "null model" refers to either the existing decision process you're trying to improve upon, or, if there isn't one, the simplest possible process (such as random guessing). Hypothesis testing could be employed to determine whether your model performs significantly better than the null model. 5) The "Bayes Rate" is maximum accuracy you can achieve given your data. If your number of observations is too small or there is too much random variance, your bayes rate may be too high for your model to be useful.