A Statistical Engineering Approach to Big Data Projects

Roger W. Hoerl
Union College, Schenectady, NY, USA – hoerlr@union.edu

Massive data sets, or Big Data, have become much more common recently, due to improved technology for data acquisition, storage, and processing of data. With the advent of Big Data, several disciplines, not only statistics, have developed new tools to analyze such data, including classification and regression trees (CART), neural nets, methods based on bootstrapping, such as random forests, and various clustering algorithms. These tools make high-powered statistical methods available to not only professional statisticians, but also to casual users. As with any tools, the results to be expected are proportional to the knowledge and skill of the user, as well as the quality of the data. Unfortunately, much of the data mining, machine learning, and Big Data literature may give casual users the impression that if one has a powerful enough algorithm and a lot of data, good models and good results are guaranteed at the push of a button. Conversely, if one applies sound principles of statistical engineering (Anderson-Cook and Lu 2012, Snee et al. 2013) to the Big Data problem, several potential pitfalls become obvious. We consider four important principles of statistical engineering that in our opinion have been either overlooked or underemphasized in the Big Data literature:

• Need for a clear strategy to guide the analysis of Big Data sets and the solution of the associated problems of interest.
• The importance of using sequential approaches to scientific investigation, as opposed to the “one-shot study” so popular in the algorithms literature.
• The need for empirical modeling to be guided by domain knowledge (subject-matter theory), including interpretation of data within the context of the processes and measurement systems that generated it, and
• The inaccuracy of the typical unstated assumption that all data are created equal, and therefore that data quantity is more important than data quality.

Keywords: statistical engineering; big data; complex problems; strategy.