Statistics with Big Data: Beyond the Hype

Joseph Rickert[1,*]

1. Revolution Analytics
*Contact author: joseph.rickert@revolutionanalytics.com

Keywords: R, statistics, big data, Hadoop

This past year the hype surrounding "Big Data" has dominated the "Data Science" world and made quite a stir in the main-stream media. Entire industries have grown up around technologies such as Hadoop and MapReduce promising astonishing insights from "crunching" large amounts of data. Putting hyperbole aside, what are the theoretical and practical challenges involved in working with very large data sets and what tools exist in R to help meet these challenges? In this talk, I will offer some ideas about how to think of big data from a statistical point of view, make some suggestions on computer architectures for facilitating working with *R* and large data sets, and show some examples of R code used to analyze large data sets including **biglm**, **Rhadoop** and **RevoScaleR code**. I will also illustrate how very large data sets are forcing developers to rethink basic algorithms by describing the rxDTrees algorithm in the **RevoScaleR** package that builds classification and regression trees on histogram summaries of the data.