Introduction To Big Data Analysis
STAT 29000, Fall 2014

Class times: MWF, 8:30 AM -- 9:20 AM, in SC 183
(STAT 29000-002; Banner CRN 68731)

Semester: Fall 2014

Prerequisite: none

Credits: 3.0

Primary Audience: sophomores in the Purdue Statistics Living-Learning Community

Description: An introduction to statistical data analysis. Computational tools for representing, extracting, manipulating, interpreting, transforming, and visualizing data, especially big data sets. Practical experience in effectively communicating insights about data. Topics include: the R environment, visualizing data, UNIX, bash, regular expressions, SQL, XML and scraping data from the internet, as well as selected advanced topics, as time permits, e.g., linear models, time series, parallel computation with distributed data, etc. No prior computational or statistical experience is necessary.

Week 1: introduction to the R platform; data types; parameters; missing values; indexes; recycling; help and documentation systems; CRAN

Week 2: factors, tapply; further discussion of arrays, data.frames, lists; importing and exporting data from csv files; strings and dates; functions

Week 3: best practices for data visualization; graphics; cartograms

Week 4: subsetting data; the family of apply functions; data transformations; verifying and cleaning data

Week 5: generating random numbers; connections with probability; simulations

Week 6: inference; regression; linear models; ANOVA; brief overview of: multiple linear regression, generalized linear models, and logistic regression

Week 7: brief overview of: clusters; resampling; factor analysis

Week 8: introduction to time series, autocorrelation, ARIMA models

Week 9: Grep, pattern matching, and regular expressions; Gawk

Week 10: Introduction and immersion into the UNIX Operating System; shells and shell scripting

Week 11: XML and extracting/scraping data from the web; parsing data; XPath and XPointer

Week 12: Databases, SQL/MySQL, and interacting with databases from R

Week 13: overview of Hadoop, parallelism, distributed data, MapReduce

Week 14: Thanksgiving vacation

Week 15: Final project preparation time

Week 16: Final project presentations