Course Proposal

STA141C Big Data & High Performance Statistical Computing

Units: 4

Format: 3 lectures per week

Discussion: 1 hour per week

Prerequisites: 141A or ECS40

Catalog Description: High-performance computing in high-level data analysis languages; different computational approaches and paradigms for efficient analysis of big data; interfaces to compiled languages; R and Python programming languages; high-level parallel computing; MapReduce; parallel algorithms and reasoning.

Course Topics & Content:

- Understanding efficiency - concepts, analysis, idioms, profiling (3 lectures)
- UNIX Shell computational model (3 lectures)
    - shell language & commands
    - pipes & redirection
    - pre-processing data
    - R connections/streams.
- Using compiled code to improve speed (C, C++ and Rcpp?). (4 lectures)
- Alternative Language for Data Analysis: Python (12 lectures)
    - Syntax, Computational model
    - Processing records sequentially (not vectorized)
    - Modules for Data Analysis (Pandas, NumPy, SciPy)
    - Visualization (Seaborn)
    - Machine learning modules (scikit-learn).
- Parallel computing strategies & technologies, (6 lectures)
    - Data locality and distribution
    - Cluster
    - Multicore
    - Cloud (AWS)
    - MapReduce, Hadoop, Spark, distributed data models.
    - Hive (database on Hadoop) & Pig

Algorithms & Statistical Methods may include

- Alternating Direction Method of Multipliers (ADMM) for statistical methods
- Bag of Little Bootstraps

- Streaming algorithms.
- Statistical/Machine Learning methods.
- Graph/network methods and algorithms
- MCMC
- Approximate results.

Possible Optional Topics:

- Using massively parallel Graphical Processing Units (GPUs) for statistical algorithms.
- Java and MapReduce, Mahout,
- The Julia language.

Grading: A series of 4 - 6 computer assignments. The instructor may choose to use participation (in class, discussion section and online forum). Similarly, the instructor may include a mid-term and/or final exam. Given the practical computer-based nature of this course, an instructor may deem written exams inappropriate.

Learning Outcomes: Students learn to reason about computational efficiency in high-level languages. They will be able to use different approaches, technologies and languages to deal with large volumes of data and computationally intensive methods.

Illustrative Reading:

- Advanced R, Wickham.
- Parallel R, McCallum & Weston.
- Python for Data Analysis, Weston.
- Hadoop: The Definitive Guide, White.

Course Overlaps: ECS158 covers parallel computing, but uses different technologies and has a more technical, machine-level focus. ECS145 covers Python, but from a more