Big Data

http://acberg.com/bigdata

CS 790-134 (Fall 2013)

M/W 3:30-4:45 in Frederick P. Brooks Jr Building, Room 009

Professor: Alex Berg
http://acberg.com
aberg@cs.unc.edu

office hours: Monday after class and by appointment.

(description) Applications of computational science are becoming ubiquitous across industry and academia (from Google to gene sequencing to the NSA). In part this is due to the availability of very large-scale data that requires processing to extract useful information. The course will cover algorithms for learning from "big data" where dataset size, or the complexity of the features extracted from the data, require special care. (objectives) The course objectives are to familiarize students with algorithms for learning from big data, as well as some practical systems issues that arise in running such algorithms on real-world clusters. A major aspect of the course will be a project centered around a research topic of interest to the student that requires large-scale learning or data-mining.

(text) There is no required text. Instead there will be a series of research papers to read and discuss in class. These will be announced in class and on the webpage.

(target audience) The course is targeted toward graduate students in computer science or students doing computational work on big data in other fields. (prerequisites) There are no prerequisites, but basic algorithms, programming, linear algebra, and calculus will be assumed. I will try to increase the enrollment cap to accommodate as many students as are interested in taking the course, but this effort may be hampered by the size of available classrooms.

(course requirements) Students will read and summarize (in a few sentences) research papers that will be discussed in class. Over the semester, there will be 2-3 small computational exercises to make some of the algorithms more concrete, as well as a programming assignment using hadoop

(map-reduce) on a cluster (probably Amazon's EC2). A major component of the course is a research project related to big data and a topic of interest to the student. This will involve a project proposal, updates, final presentation, and write-up. Students will work as individuals on the exercises, and in small groups on the assignment and projects. (grading criteria) There will be no midterms or final. Grading will be 50% final project, 30% participation in class discussion + summaries + exercises, and 20% on the assignment. (honor code) Students are always expected to cite collaborators and sources.

A preliminary list of course topics follows. It may be adjusted to fit interests of the class:

- Streaming and online approximation of statistics over one dimension, multiple dimensions – *Exercise*
- Overview of classification
- Streaming and online learning of classifiers (Newton, SGD, L-BFGS) – *Exercise*
- Decision stumps, trees, and boosting
- Dimensionality reduction, Hashing, Feature Learning
- Coarse parallelism on clusters (map-reduce,hadoop,EC2) – **Assignment**
- parallel algorithms for statistics and classifiers
- parallel/streaming algorithms over large-graphs, statistics and subgraph search
- special case: sequence alignment and search (words/genes)
- special case: convolutional neural networks
- **Project**

First reading assignment:
For the second class, please read the papers listed on the course page.