**Introduction to Data Science**

Instructor: Bill Howe, Phd

Email: billhowe@cs.washington.edu

Office: Allen 450

Period: Oct 16 - Dec 18, 2012

Class Meeting Time: Tuesdays, 6:00-9:00 p.m.

Class Meeting Place: TBD

Web: http://homes.cs.washington.edu/~billhowe/datascience

Office Hours: by appointment

**Course Overview:** Commerce and research is being transformed by data-driven discovery and prediction. Skills required for data analytics at massive levels – scalable data management on and off the cloud, parallel algorithms, statistical modeling, and proficiency with a complex ecosystem of tools and platforms – span a variety of disciplines and are not easy to obtain through conventional curricula. Tour the basic techniques of data science, including both SQL and NoSQL solutions for massive data management (e.g., MapReduce and contemporaries), algorithms for data mining (e.g., clustering and association rule mining), and basic statistical modeling (e.g., logistic and non-linear regression).

**Learning Objectives:** By the end of this course, students will be able to:

* Explain the major trends in technology, business, and science behind the terms data science and big data.
* Explain the relative strengths and weaknesses of the scalable data analytics platforms in use today, including databases, MapReduce-based platforms, NoSQL solutions, cloud services.
* Given a specification of a large-scale analytics problem, design an application to solve it.
* Apply a selection of machine learning algorithms and suggest scalable implementations of them.
* Identify and articulate the underlying concepts independently of specific technologies.

**Course Structure:** Each class will consist of a 1-hour lecture, a 1-hour case study and demonstration of a specific system, and 1-hour of discussion and hands-on work. Each week, we will consider a category of scalable data platform through a lecture and consider a representative example from this category in detail through a demonstration. Students will be asked to complete a hands-on homework assignment based on the material presented in class and, in some cases, come prepared to discuss assigned reading. The reading assignments will generally be research papers from relevant computer science conferences.

**Student Assessment:** Assignments: 80%, Participation: 20%. Participation will be a combination of attendance and discussion involvement; in class and online involvement will both contribute. Assignments will typically not be graded purely in terms of correct/incorrect answers, but students will be expected to demonstrate effort and critical thinking. In this course, discussion between students about assignments is not discouraged: the goal is to learn as much as possible in a short time, and discussion is a very efficient way to do this. Some assignments may be completed in groups, depending on students’ experience level. In these cases, a portion of the grade will be based on peer review by one’s group members.

**Programming Assignments:** The assignments will involve programming, typically in Java and Python. You will not be expected to be experts in these particular languages; programming experience in any language should be sufficient. We will provide a command-line linux-based environment accessible via ssh with which to complete the assignments. This environment will be hosted on Amazon Web Services; see next section. You are free to use a different platform or programming environment, but you will be responsible for managing the environment yourself.

**Cloud Computing:** You will receive a $100 credit for Amazon Web Services with which to complete the assignments. We will review how to use the features of AWS you will need.

**Textbook:** None. All materials will be on the website. We will use a combination of relevant research papers, documentation from relevant systems, and readings from the web.

**Prerequisites:** You are expected to have intermediate proficiency inprogramming, databases, and basic statistics. The quiz you completed as part of the application is an appropriate measure of the background you are expected to have.

**Specific Course Topics:**

*Week 1:* Introduction: big data definitions, data science definitions, history and context, technology landscape

Reading: None

3 different problems

[http://www-conf.slac.stanford.edu/xldb2012/talks/xldb2012\_wed\_1635\_MichaelStonebraker.pdf](http://www-conf.slac.stanford.edu/xldb2012/talks/xldb2012_wed_1635_MichaelStonebraker.pdf" \t "_blank)

*Week 2*: Relational data model, relational algebra, algebraic optimization, query processing, challenges

Reading: TBD

Data models

Operators: Select Project Join GroupBy

Join Algorithms

NoSQL = NoJoin

Matrix multiplication in LINQ

http://research.microsoft.com/apps/pubs/default.aspx?id=171321

Matrix multiplication in SQL

MadSkills paper

Tool: LINQ, SQLShare

*Week 3:* MapReduce introduction, Hadoop, MR vs. RDBMS, Elastic MapReduce

Reading: Dean et al, MapReduce: Simplified Data Processing on Large-Scale Clusters

http://usenix.org/events/osdi04/tech/full\_papers/dean/dean.pdf

Reading: Stonebraker et al. "MapReduce and Parallel RDBMS: Friend or Foe?" CACM 2011 - <http://cacm.acm.org/magazines/2010/1/55743-mapreduce-and-parallel-dbmss-friends-or-foes/fulltext>

Reading: Performance analysis of parallel DBs and MapReduce, Pavlo et al. SIGMOD 09 - <http://database.cs.brown.edu/sigmod09/benchmarks-sigmod09.pdf>

Tool: Elastic MapReduce

Assignment: 1000 genomes

*Week 4:* Advanced MapReduce algorithms: Join, matrix multiply; communication costs; algorithm design

Reading: Join Algorithms in MapReduce

*http://horicky.blogspot.com/2010/08/designing-algorithmis-for-map-reduce.html*

[*http://research.microsoft.com/apps/pubs/default.aspx?id=171321*](http://research.microsoft.com/apps/pubs/default.aspx?id=171321)

[*http://horicky.blogspot.com/2010/08/designing-algorithmis-for-map-reduce.html*](http://horicky.blogspot.com/2010/08/designing-algorithmis-for-map-reduce.html)

Matrix multiplication in Hadoop

*Week 5*: Scalable text processing, scalable search

Reading: Jimmy Lin

Assignment:

*Week 6*: Graph analytics: recursive queries, graph mining, social networks

Reading: TBD

*Week 7:* Visualization, data presentation, storytelling

Assignment: Tableau

*Week 7*: Key-value stores and NoSQL; tradeoffs of SQL, NoSQL, and NewSQL

Reading: http://[www.cidrdb.org/cidr2011/Papers/CIDR11\_Paper32.pdf](http://www.cidrdb.org/cidr2011/Papers/CIDR11_Paper32.pdf)

Reading: <http://db.csail.mit.edu/projects/cstore/abadi-sigmod08.pdf>

Reading: Dremel (Google) <http://vldb.org/pvldb/vldb2010/pvldb_vol3/R29.pdf>

*Week 8:* Basic statistical analysis at scale: experiment design, sampling, regression

Descriptive statistics

Experiment design / hypothesis testing

Sampling

Regression

Introduction to R

Introduction to Octave

Comparison to Excel

Scaling up to SQL

Scaling up to Hadoop

Reading:

Assignment: Introduction to R and Octave

*Week 9*: Machine learning toolbox at scale: k-means clustering, multi-dimensional scaling

Reading: TBD

Assignment: Image analysis

Linear regression

Least-angle regression

Week 10: Mini-Project Presentations

**The instructor reserves the right to alter the syllabus if circumstances dictate.**