INFX 598 C/D Course Syllabus:

Data Science "in the Wild"

Winter 2013
University of Washington School of Information
Lectures: Tuesday and Thursday 1:30-3:20, MGH 420

Course Website: http://canvas.uw.edu
Instructor: Prof. Joshua Blumenstock
Office Hours: Tuesdays, 3:30-5:00, 370E Mary Gates Hall
Contact: (206) 685-8746, joshblum@uw.edu

Course Description

This course offers students an introduction to the growing field of "Data Science" as practiced by leading data scientists in industry and research. As "big data" become the norm in modern business and research environments, there is a growing demand for individuals who are able to derive meaningful insight from large, unruly datasets. This requires a diverse mix of skills, from data munging, wrangling, and storage; to machine learning and econometrics; to effective visualization and communication.

Through a combination of hands-on exercises and guest lectures by experts in the field, this course provides an overview of several key concepts, skills, and technologies used by practicing data scientists. While the lectures will be intelligible to a general audience, successful completion of assignments requires college-level exposure to statistics and programming.

Prerequisites

A data scientist is often referred to as someone who knows more statistics than a computer scientist and more computer science than a statistician. Students enrolled in the course must have college-level exposure to both statistics and programming. Students who do not meet the following requirements, and in particular the programming requirement, will find it difficult to satisfactorily complete problem sets, and should consider taking the course at a later date.

Programming: Students should be able to comfortably program in a high level programming language like Java, python, php, or C/C#/C++. Note that html, javascript, and VBA are not sufficient in this context. "Comfortably" implies that students should be able to write simple programs from scratch, like a web scraper, or a text parser, or a simple game of scrabble or tic-tac-toe. Real-world programming experience (e.g. a summer internship that involved writing code) will be more useful than formal CS coursework.
Statistics: Students should have had prior exposure to regression methods and should understand concepts of hypothesis testing and statistical significance. Experience with R or MatLab will be especially valuable in completing the problem sets.

Course Outline

- Introduction to data science
- Identifying questions and developing an empirical framework
- Data capture, munging, ETL, storage, and organization
- Basic analytics: Distributions, t-tests, and the importance of basic statistics
- Advanced analytics: Data mining, machine learning, network analysis
- Visualizing and Communicating Data
- Scaling to terabytes and petabytes
- Perspectives from industry and academia

Assignments and Grading

Students are **required** to read the material assigned each week before section, and will be expected to actively engage in discussion of these materials.

Course grades will based on a final group project, two problem sets, discussion leadership, and overall classroom participation. The two problem sets will be programming assignments, using R and Python. The final project will be done in groups of 3-4 students; further details will be provided at the start of the quarter.

- Group final project: 40%
- Technical problem set 1: 20%
- Technical problem set 2: 20%
- Class Discussion Leadership: 10%
- Class Participation: 10%
- Extra Credit: 4%

Grading Policy

- All assignments are to be submitted on Canvas by 12pm on the due date.
- Assignments turned in *up to* 24 hours after the due date will be penalized 20%.
- Any assignments turned in more than 24 hours late will receive no credit.

A Note on Programming Languages

Most in-class code demonstration will involve either R or Python. Guest lecturers may give examples in other languages but will explain what the code means. Homework assignments will generally require R or Python. If you would prefer to use a different language successfully, you may do so, but you will be "on your own" and should not expect technical support.

Academic Integrity Policy

Discussion with instructors and classmates is encouraged, but each student must turn in individual, original work and cite appropriate sources where appropriate.

Readings and Class Discussion

Required and optional readings will be announced in class and posted on the course website. Students will be assigned by the instructor into groups of 2-4, and each group will be responsible for guiding classroom discussion for a given week (two classes). Unless otherwise noted by the instructor, a discussion will last roughly 20-30 minutes, and will consist of (i) a 10-15 minute summary of all required **and optional** readings for that week, and (ii) 10-15 minutes of open discussion, moderated by the discussion leaders.

Over the course of the quarter, we will host several prominent visiting speakers. On days when one or more speakers are present, there will be no discussion of the readings (in such instances, all readings for a week will be discussed on the day without a speaker). Instead, the discussion leaders will be responsible for researching the visiting speaker, meeting the speaker and walking him/her to the classroom, and moderating the classroom discussion and Q&A following the speaker's presentation. Moderators should prepare several intelligent, challenging questions to ask the speaker immediately after the talk finishes, and if ever there is a lull in the discussion.

By 12pm on the Monday following second class, discussion leaders are also responsible for posting a 400-600 word entry on the class blog/wiki that summarizes the required and optional readings for the week. Discussion leaders must also post a separate 400-600 word summary of the most salient points from the class/guest lectures and in-class discussion.

Additional Resources

While several textbooks on Data Science are currently being written (see first few links below), to date there is no great textbook that is suitable for this course. Data Science is an emerging field with an amorphous identify, so the readings for this course will be assembled from the best of what is out there. For those interested in digging deeper, I recommend the following:

- Rajaraman, Leskovec, & Ullman (unpublished): Mining of Massive Datasets. Online at http://infolab.stanford.edu/~ullman/mmds/book.pdf
- Provost & Fawcett (unpublished): Data Science for Business: Fundamental principles of data mining and data-analytic thinking
- Friedman, Hastie, Tibshirani (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition. Online at: http://www-stat.stanford.edu/~tibs/ElemStatLearn/
- Whitten, Frank, Hall (2011). Data Mining: Practical Machine Learning Tools and Techniques (3rd Edition). Morgan Kauffman.
- Torgo, Luis (). Data Mining with R: Learning with Case Studies
- Adler, Joseph (). R in a Nutshell: A Desktop Quick Reference
- Lutz and Ascher (): Learning Python (O'Reilly)

Assignments

Group Project (40%): Students will form groups of 2-4, and complete **one** of the following assignments.

A. Analysis of a dataset "in the wild"
B. Sectoral Analysis

Each group must submit a 200-300 word description of the proposed project by 12pm on January 24.

Problem Set 1 (20%): Programming Assignment 1: Data processing and basic analytics

The goal of this problem set is to familiarize yourself with implementing basic statistical analysis on a small dataset using R. It is important that you read Chapters 1-3 carefully, and work to understand exactly how the authors are using R to manipulate the underlying data. If you merely hack your way through the exercises without working through the material in the chapters, you will come away with a very incomplete understanding of how to perform such analysis.

Assignment: After reading Chapters 1-3 (pp 1-64) of Everitt & Hothorn, complete the following exercises (2 points each, 18 points total) : 1.1, 1.2, 1.3, 1.4; 2.1, 2.3, 2.4; 3.1, 3.2. Extra credit (2 points): 3.3. Submit two files:

1) A polished .pdf with your solutions. 2 points will be given for overall presentation, so take care to appropriately label your graphs and tables. At the top of your submission, write your name and the names of any students with whom you worked, if applicable. Also write an estimate of how long, in hours, it took you to complete the exercise -- this will not affect your grade, it is for calibration.
2) Your commented R code as a plain text file

Problem Set 2 (20%): Programming Assignment 2: Advanced analytics and visualization

- Details TBD

Classroom Discussion Leadership (20%):

- See description in above section, "Readings and Class Discussion"
- All students are expected to attend each class and contribute intellectually to the class environment, irrespective of who is leading classroom discussion.

Extra Credit (4%): Several popular books have been written on topics thematically relevant to this course. Excerpts from some of these books are required reading. For extra credit, write a short (400-800 words) review of one of the books listed below. The review should summarize the core points of the book and provide a critical analysis that draws attention to strengths and shortcomings of the book. Please do not review a book that you have previously read.

- Davenport & Harris. Competing on Analytics
- Davenport, Harris, & Morison. Analytics at Work
- Ian Ayres. Super Crunchers: Why Thinking-by-Numbers Is the New Way to Be Smart
- Steven Baker. The Numerati
- Thomas Redman. Data Driven: Profiting from Your Most Important Business Asset
- Sasha Issenberg: The Victory Lab
- Nate Silver: The Signal and the Noise

Course Structure, Readings, and Assignments*

*** SUBJECT TO REVISION – CHECK CANVAS FOR UP-TO-DATE VERSION!**

Section 1: Introduction to Data Science

January 8: What is Data Science?

Readings:

- Executive summary of: McKinsey Global Institute. Big data: The next frontier for innovation, competition, and productivity. Online at http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation

January 10: Industry Overview

Guest Speaker: Bob Davis (General Manager, Office 365, Microsoft)

DUE: Background and interests (online quiz through canvas)

Readings:

- Thomas Davenport (2006). "Competing on Analytics", Harvard Business Review, Jan. 2006, Vol. 84 Issue 1, pp. 99-107. http://offcampus.lib.washington.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=19117901&site=ehost-live
- Chapter 1 of Provost & Fawcett: Data Science for Business

Section 2: Identifying questions and developing an empirical framework

January 15: Business experiments

Readings:

- Andrew Gelman: "There are four ways to get fired from Ceasars" http://andrewgelman.com/2012/12/there-are-four-ways-to-get-fired-from-caesars-1-theft-2-sexual-harassment-3-running-an-experiment-without-a-control-group-and-4-keeping-a-gambling-addict-away-from-the-casino/
- Anderson & Simester (2011). "A Step-By-Step Guide to Smart Business Experiments", *Harvard Business Review*, pp. 99-105
- Davenport (2009). "How to Design Smart Business Experiments", *Harvard Business Review* pp. 69-76.
- Ariely (2004). "Why Businesses Don't Experiment", *Harvard Business Review*, p. 34
- Bertrand et al. (2009). "Does Ad Content Affect Consumer Demand?" *Alliance*, 14:3, p.18
- INTRODUCTION TO: Bertrand M.; Karlan D.; Mullainathan S.; Shafir E.; Zinman J. (2012) "What's advertising content worth? Evidence from a consumer credit marketing field experiment" *Quarterly Journal of Economics*, 125(11) pp. 263-

- [optional] Bertrand M.; Karlan D.; Mullainathan S.; Shafir E.; Zinman J. (2012) "What's advertising content worth? Evidence from a consumer credit marketing field experiment" *Quarterly Journal of Economics*, 125(11) pp. 263-306
- [Optional] Selections from Gerber & Green: Field Experiments

January 17: Developing an Empirical Framework

Guest Speaker: Scott Golder (Staff Sociologist, Context Relevant)

Readings:
- Chapter 2 of Provost & Fawcett: Data Science for Business
- [Optional] Chapter 4 of Bernard: Social Research Methods
- [Optional] Alamar and Mehrotra, "Beyond 'Moneyball': The rapidly evolving world of sports analytics" Online at http://www.analytics-magazine.org/special-articles/391-beyond-moneyball-the-rapidly-evolving-world-of-sports-analytics-part-i

Section 3: Data capture, munging, storage, and organization

January 22: Data Capture and ETL

Guest Speaker: Andrew Borthwick (Principal Scientist and Director of Data Research, Intelius)

Readings:
- "Duplicate Record Detection: A Survey", by Elmagarmid, et. al.
- "Record Linkage: Similarity Measures and Algorithms" by Koudas, et. al.
- [optional] Andrew's work at Intelius: here for blocking and here for pairwise decision making.
- Bibliographic database assignment? De-dup the authors

January 24: Storage and Organization: Databases, Scalable SQL and NoSQL

DUE: 1-paragraph group project description

Guest Speaker: Bill Howe (Director of Research, Scalable Data Analytics, eScience Institute and Affiliate Assistant Professor, Department of Computer Science and Engineering, UW)

Readings:
- Stonebraker et al (2010). "MapReduce and Parallel DBMS's: Friends or Foes?" Online at http://database.cs.brown.edu/papers/stonebraker-cacm2010.pdf
- Rick Cattell, "Scalable SQL and NoSQL Data Stores", *SIGMOD Record*, December 2010 (39:4)
- [Optional] Cohen et al. (2009) "MAD Skills: New Analysis Practices for Big Data" Online at: http://db.cs.berkeley.edu/papers/vldb09-madskills.pdf

Section 4: Basic analytics

January 29: Distributions, t-tests, and the importance of basic statistics

Readings

- Chapters 1-3 of: A Handbook of Statistical Analyses Using R
- Chapters 1 & 2 of Freedman, Pisani, and Purvis
- [Optional] Chapters 6 & 19 of Freedman, Pisani, and Purvis
- [Optional] H. Stern: "Statistics and the College Football Championship," *The American Statistician*, 2004.

January 31: How far can basic statistics get you?

Readings

- Chapter 3 of: A Handbook of Statistical Analyses Using R
- [Optional] Excerpts from Huff: How to Lie With Statistics
- [Optional] Chapters TBD of Torgo: Data Mining with R: Learning with Case Studies

February 5: Regression

Readings

- Chapters 3 & 4 of Provost & Fawcett: Data Science for Business
- Chapter 6 of: A Handbook of Statistical Analyses Using R
- [Optional] Excerpts from Kennedy: A Guide to Econometrics

Section 5: Advanced analytics

February 7: Machine Learning I: Introduction

DUE: Problem Set 1

Readings:

- Chapter 1 of: Friedman, Hastie, Tibshirani (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition. Available online at: http://www-stat.stanford.edu/~tibs/ElemStatLearn/
- Chapter 1 of: Mining of Massive Datasets. Online at http://infolab.stanford.edu/~ullman/mmds/book.pdf
- [Optional] Haydn Shaughnessy, "How Semantic Clustering Helps Analyze Consumer Attitudes": http://blogs.hbr.org/research/2010/07/every-day-in-the-english.html

February 12: Machine Learning 2: Supervised Learning

Readings:

- Chapter 8 of The Signal and the Noise: "Less and Less and Less Wrong" (Bayes' Theorem)
- [Optional] C. Haruechaiyasak: "A Tutorial on Naive Bayes Classification", 2008.
- [Optional] "Polonium: Tera-Scale Graph Mining and Inference for Malware Detection." Duen Horng (Polo) Chau et al. *Proccedings of SIAM International Conference on Data Mining (SDM) 2011*. April 28-30, 2011. Mesa, ArizonaSK-Learn User Guide, Sections

February 14: Machine Learning 3: Unsupervised Learning

Readings:

- Chapter 7 [Read 7.1, 7.2, 7.3] of: Mining of Massive Datasets. Online at http://infolab.stanford.edu/~ullman/mmds/book.pdf
- [Optional] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein (1998). "Cluster analysis and display of genome-wide expression patterns" *Proceedings of the National Academy of Sciences*. Vol. 95 pp. 14863-14868
- [Optional] Rajkumar Venkatesan (2007). "Cluster Analysis for Segmentation", *Darden Business Publishing*

February 19: Social Network Analysis

Readings:

- [Optional] N. Godbole et. al: "Large-scale sentiment analysis for news and blogs", *International Conference on Weblogs and Social Media*, 2007.
- [Optional] L. Page, et. al: "The PageRank citation ranking: Bringing order to the web", Stanford, 1999.
- [Optional] Albert R, Jeong H, Barabasi AL. (1999): "Diameter of the World Wide Web". *Nature*, 401:130-131

Visualizing and Communicating Data

February 21: Visualizing Quantitative Information

Guest Speaker: Jock Mackinlay (Senior Director, Visual Analysis, Tableau Software)

Readings:

- Excerpts from Edward Tufte, "Visual Display of Quantitative Information"
- WSJ Guide to Information Visualization
- [Optional] Excerpts from Beautiful Data

February 26: Communicating Results Effectively

Readings

- "How to Lie with Charts" and "How to Lie with Maps"
- TBD

February 28: Text Mining

Readings:

- TBD

Scaling to terabytes and petabytes

March 5: Scaling: What works and what doesn't (and what might in the future)

Readings:

- Readings TBD

- Cloudera overview

March 7: Common applications and tools: MapReduce, Hadoop, Hive, and Alternatives

Readings:

- Chapter 2 ("Large-Scale File Systems and Map-Reduce") of: Mining of Massive Datasets. Online at http://infolab.stanford.edu/~ullman/mmds/book.pdf
- [Optional] Heimstra and Hauff, "MapReduce for Information Retrieval: Let's Quickly Test This on 12 TB of Data", In: M. Agosti et al. (Eds.): *CLEF2010*, pp. 64-69, 2010
- [Optional] Dean and Ghemawat, "MapReduce: A Flexible Data Processing Tool", *Communications of the ACM*. January 2010.

Perspectives from industry and academia

March 12: Group Project Presentations

Readings:

- Review Chapters of: McKinsey Global Institute. Big data: The next frontier for innovation, competition, and productivity.
- Group Project Reports

March 14: The Future and Ethics of Data Science and Big Data

Readings:

- Chris Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", Wired magazine. Online at http://www.wired.com/science/discoveries/magazine/16-07/pb_theory
- Counterpoint to Anderson TBD