INFO 370
Introduction to Data Science

Fall 2015

Information School, University of Washington

Lectures: Tuesday and Thursday 1:30 - 3:20, BLD 070

Lab: Wednesday 1:30 - 2:20, MGH 430

Course Website: https://canvas.uw.edu/courses/1012299

Instructor: Jevin West

Office Hours: Wednesdays, 11:00 - 12:00 pm, MGH 310E

TA: Lavi Aulck

Office Hours: Monday, 3:00 - 4:00 pm, MGH 310

Course Description

This course offers students a practical, hands-on introduction to the growing field of "Data Science. 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 data. This requires a heterogeneous mix of skills, from data wrangling and ETL (extracting, transforming, and loading data); to statistical modeling; to machine learning and econometrics; to effective visualization and communication; and awareness of the ethical and privacy implications of the

data revolution. Through a combination of data-intensive exercises and projects, this course provides an overview of key concepts, skills, and technologies used by data scientists. Interested students must have college-level exposure to statistics and programming.

Prerequisites

This course will involve programming and statistical analysis in R and python. Students are highly encouraged to familiarize themselves with R and python prior to the first day of class. CSE 142 and STAT 221 (or equivalent course) are required. If you have not had these courses, please talk to the instructor. Please note, students who do not meet these requirements will find it difficult to complete required assignments and excercises.

- **Programming**: Students should be able to comfortably program in a high level programming language like Java, python, javascript, php, R, or `C/C#/C++`. Note that html 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.
- **Statistics**: Students should have had introductory coursework in both probability and statistics prior to enrolling in this course. At a minimum, students should have an operational understanding of hypothesis testing, statistical significance, and regression analysis.

Course Outline

- Concepts and opportunities in data science
- Working with data (EDA)
- Storage and scaling (cloud computing)

- Empirical frameworks and experimental design
- Basic analytics and statistical inference
- Basics in machine learning and pattern recognition
- Interactive data visualization
- Data ethics and privacy

Assignments and Grading

Course grades will consist of a group project, assignments, quizzes, labs, and participation. Details can be found on the canvas course website. Assignments and the group project are due at the beginning of class. Late assignments will receive a significant grade reduction. Rubrics will be provided on the course website. Each assignment type will be weighted as follows:

Group Project	200 pts
Assignments	100 pts
Quizzes	100 pts
Participation	100 pts
Lab	50 pts
Total	**550 pts**

Academic Integrity

Discussion with instructors and classmates is encouraged, but each student must turn in individual, original work and cite appropriate sources where appropriate. See UW policy for guidelines on academic conduct.

Readings and Participation

Required and optional readings will be announced in class and posted on the course website one week in advance of lecture. The goal of these readings is to deepen your knowledge of Data Science, as the topic is so broad that not everything can be covered in class. Students who come to class unprepared detract from everyones ability to learn in an active and engaged environment. To help foster this environment, I may call on random students to solicit opinions of the readings or to summarize a core concept. Students who are clearly unprepared (or who are absent) will miss this opportunity to earn full credit participation. If you are going to miss class, you need to let me know ahead of time. If there are extenuating circumstances and you have to miss class at the last minute, let me know as soon as possible. If you do miss class, you may be asked to write an essay about the class topic for that day. There is a university-wide Data Science Seminar. I encourage you to this and similar events across campus.

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 current literature and book chapters. Readings will be posted on the class website. Make sure to check the readings before class because they will be updated throughout the quarter. Final readings for each class will be posted 1 week before class. For those interested in digging deeper, I recommend the following:

- Jure Leskovec, Anand Rajaraman, and Jeff Ullman. *Mining of Massive Datasets*, chapter 1. Cambridge University Press, 2012
- Foster Provost and Tom Fawcett. Data science and its relationship to big data and data-driven decision making. *Big Data*, 1(1):51–59, 2013

- Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2005

- Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005

- Luis Torgo. *Data Mining with R: Learning with Case Studies*, pages 1–38. Chapman and Hall/CRC, 2010-11-09

- Joseph Adler. *R in a nutshell: A desktop quick reference*. " O'Reilly Media, Inc.", 2010

- Mark Lutz. *Learning python*. " O'Reilly Media, Inc.", 2013

Schedule

**** Subject to revision – check canvas for up-to-date information***

Week 1

October 1: Syllabus, Logistics, What is Data Science?

Week 2

October 6: Introduction to R and Python

DUE: Data Science Example #1

Required Readings:

- Luis Torgo. *Data Mining with R: Learning with Case Studies*, chapter 1, pages 1–38. Chapman and Hall/CRC, 2010-11-09
- McKinsey Company Report. Big data: The next frontier for innovation, competition, and productivity. pages 1–15, 2011

October 7: LAB: Introduction to Python

October 8: Web scraping, data ingestion, regular expressions

DUE: Quiz #1

Required Readings:

- Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014

Week 3

October 13: Data Science for Business

DUE: Group Assembly

Required Readings:

- Thomas H Davenport. Competing on analytics. *Harvard Business Review*, 84(1):98, 2006

Optional Readings:

- Foster Provost and Tom Fawcett. Data science and its relationship to big data and data-driven decision making. *Big Data*, 1(1):51–59, 2013

October 14: LAB: Web Scraping, APIs

October 15: Empirical Frameworks and Experimental Design

DUE: Data Science Example #2

Required Readings:

- Jim Gray. A transformed scientific method. *The fourth paradigm: Data-intensive scientific discovery*, pages xvii–xxxi, 2009

Week 4

October 20: Random Variables, Probability Distributions

DUE: Assignment #1

Required Readings:

- Morris H DeGroot and Mark J Schervish. *Probability and statistics*, chapter 3, pages 93–186. Addison-Wesley, 3 edition, 2002

Required Readings:

- Jim Pitman. *Probability*, chapter 1-3. Springer-Verlag, 1993

October 21: LAB: Sampling Distributions

October 22: Central Limit Theorem, Bayes' Theorem

Required Readings:

- Nate Silver. *The signal and the noise: Why so many predictions fail-but some don't*, chapter 8. Penguin, 2012

Optional Readings:

- OpenIntro statistics. *Diez, David M and Barr, Christopher D and Cetinkaya-Rundel, Mine*, chapter 4. CreateSpace independent publishing platform, 2012
- Watch Kahn Academy lesson on Central Limit Theorem and Sampling Distributions

Week 5

October 27: Hypothesis Testing, Statistical Tests

DUE: Group Proposal

Required Readings:

- Foster Provost and Tom Fawcett. *Data Science for Business: What you need to know about data mining and data-analytic thinking*, chapter 3-4. " O'Reilly Media, Inc.", 2013
- OpenIntro statistics. *Diez, David M and Barr, Christopher D and Cetinkaya-Rundel, Mine*, chapter 5. CreateSpace independent publishing platform, 2012
- James D. Leeper. Choosing the correct statistical test

Optional Readings:

- Hal S Stern. Statistics and the college football championship. *The American Statistician*, 58(3), 2004

October 28: LAB: Statistical Tests

October 29: Introduction to Regression

DUE: Quiz #2

Required Readings:

- OpenIntro statistics. *Diez, David M and Barr, Christopher D and Cetinkaya-Rundel, Mine*, chapter 7-8. CreateSpace independent publishing platform, 2012

Optional Readings:

- Gareth James, Daniela Witten, and Trevor Hastie. *An Introduction to Statistical Learning: With Applications in R.*, chapter 3. Taylor & Francis, 2014

Week 6

November 3: Introduction to Machine Learning

Guest Speaker: Prof. Joshua Blumenstock

DUE: Group Assembly

Required Readings:

- Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. *The elements of statistical learning: data mining, inference and prediction*, chapter 1. Springer, 2005
- Jure Leskovec, Anand Rajaraman, and Jeff Ullman. *Mining of Massive Datasets*, chapter 1. Cambridge University Press, 2012

Optional Readings:

- H Shaughnessy. How semantic clustering helps analyze consumer attitudes. *Retrieved, October*, 4:2010, 2010

November 4: LAB: Machine Learning in Azure

November 5: K-nearest neighbors and dimensionality

DUE: Data Science Example #2

Required Readings:

- TBA

Week 7

November 10: Databases in the age of Big Data

Optional: Mid-quarter evaluations

Required Readings:

- NoSQL Databases

November 11: LAB: UW Holiday (No Lab)

November 12: Distributed Computing, MapReduce, Hadoop

Guest Speaker: Ian Wesley-Smith

Required Readings:

- TBA

Week 8

November 17: Network Analytics (Ranking, Community Detection)

DUE: Preliminary Data Analysis

Required Readings:

- Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: bringing order to the web. 1999

Optional Readings:

- Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010
- Duncan J Watts. A twenty-first century science. *Nature*, 445(7127):489–489, 2007

November 18: LAB: AWS (EC2, S3, IAM)

Required Readings:

- AWS EC2
- AWS S3

November 19: Cloud Computing (Overview, AWS, Azure, Hadoop)

Required Readings:

- What is cloud computing?

Week 9

November 24: Information Visualization, D3

DUE: Quiz #3

Required Readings:

- Perception in Visualization by C. Healey

Optional Readings:

- Mark Monmonier. *How to lie with maps*. University of Chicago Press, 2014

- Gerald Everett Jones. *How to lie with charts*. LaPuerta Books and Media, 2011
- Toby Segaran and Jeff Hammerbacher. *Beautiful data: the stories behind elegant data solutions*. " O'Reilly Media, Inc.", 2009
- Edward R Tufte and PR Graves-Morris. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT, 1983
- Dona M Wong. *The Wall Street Journal Guide to Information Graphics: The Dos and Don'ts of Presenting Data, Facts, and Figures*. WW Norton & Company, 2013

November 25: LAB: No Lab (Thanksgiving)

November 26: No Class (Thanksgiving)

Week 10

December 1: Data Ethics

DUE: Ethics Essay

Required Readings:

- Danah Boyd and Kate Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679, 2012
- Neil M Richards and Jonathan H King. Big data ethics. *Wake Forest L. Rev.*, 49:393, 2014

December 2: LAB: Presentation Prep

December 3: Project Presentations

DUE: Group Presentations

Week 11

December 8: Applications in Data Science

DUE: Final Paper

Required Readings:

- D. Vilhena, J. Foster, M. Rosvall, **J.D. West**, J. Evans, and C. Bergstrom. Finding cultural holes: How structure and culture diverge in networks of scholarly communication. *Sociological Science*, 1(June):221–238, 2014

December 9: LAB: Project Website

December 10: Data Science Opportunities, Job Interviews, Next Steps

DUE: Team Evaluation

Group Project

Students will form groups of 3-4 and jointly find, extract and analyze a '' dataset in the wild" Students may either form groups of their own choosing, or may ask the instructor to be randomly paired with other group members.

The goal of the project is to find an original dataset and question, analyze the data and communicate the findings in a written report and presentation in front of the class. Students should seek to perform a novel analysis on a dataset of their choosing, to reveal previously undiscovered patterns and answer previously unknown questions. Exceptional examples of such analysis include many of the links posted on canvas, recent Kaggle competitions, and other bite-sized projects such as Pulse of the Nation, Soda vs. Pop, some of the more thoughtful OK Cupid blog posts, etc. The points and due dates for the group project are as follows:

Assignment	Points	Due Date
Team Assembly	10	Oct. 13
Proposal	30	Oct. 27
Preliminary Data Analysis	20	Nov. 5
Presentation	30	Dec. 3
Final Paper	100	Dec. 8
Website	15	Dec. 9
Team Evaluation	5	Dec. 10
Total	**200**	

Updated rubrics for each part of the group project can be found on canvas. Here are general guidelines for each part.

Paper

- Timely submission (10 pts)
- Overall ambition, creativity, novelty, and difficulty of project (20 pts)
- Clarity of empirical questions (10 pts)

- Analysis and choose of suitable analytic framework (20 pts)
- Use of appropriate and compelling visualizations (20 pts)
- Clarity and organization of written report (20 pts)
- Length: approximately 2000 - 3000 words
- Note: Points for exceptional effort: These will be awarded if the project drastically exceeds expectations in any of the above categories. This could occur if, for instance, students choose to work with an unusually messy or large dataset, create unusually compelling animations or visualizations, implement unusually difficult algorithms, or present their work in an unusually compelling report or presentation.

Proposal

- Timely submission (5 pts)
- Motivation: Why are you investigating this question? (5 pts)
- Data: Where will you get your data? How will you acquire your data (e.g., API, scraping)? Provide links to your data. (5 pts)
- Question: What question are you asking? Can you ask your question given your data and time constraints? Has your question been asked before? What is the novelty in your question? (5 pts)
- Analysis: What kind of analysis could you perform on the data? This doesnt have to be fully worked out. You will do this in the preliminary analysis report. (0 pts)
- Previous work: What other work has been done around your project? Provide at least 4 citations. (5 pts)
- Plan: What is your plan for finishing this project by the end of the quarter? Who will be responsible for the different tasks? (5 pts)

- Length: approximately 2- 3 pages

Preliminary Data Analysis

For the preliminary data analysis, you will conduct a first-pass descriptive analysis of your dataset as a way of exploring your data. You will produce a set of tables and graphs that explore your data and that contain summary statistics about the data. Questions to answer for each data source include:

1. What information/features/characteristics do you have for each observation?
2. What are the min/max/mean/median/sd values for each of these features? What is the distribution of the core features (show a histogram)?
3. Are there obvious trends in the data (over time, across subgroups, etc.), and are the differences statistically significant?
4. What are the other salient aspects of the data (e.g. geospatial factors, text content, etc.)
5. Provide a bullet-list of the next 5-10 tasks you will perform in analyzing your dataset.

Grading criteria consists of the following:

- Timely submission (5 pts)
- Clarity of figures and/or tables (5 pts)
- Clarity of methods description (5 pts)
- Next steps of the analysis (5 pts)
- Length: approximately 2-3 pages

Presentation

- Motivation (5 pts)
- Research Question (5 pts)
- Data Details (5 pts)
- Methods of Analysis (5 pts)
- Results and interpretation (5 pts)
- Conclusion and Future Directions (5 pts)

Project Website

- Timely submission (5 pts)
- Team and project description (5 pts)
- Project figure (5 pts)

Data

There are unlimited data sets on the web to use for this project. Here are some examples to help generate some ideas:

- Seattle Government Open Data
- Stanford Large Network dataset
- US Government Open Data
- Twitter data)
- Wikipedia
- Movie database
- Government data
- Yahoo data
- Mobile phone data
- Best Buy data
- Medical data
- Flickr data