INFO 575
Data Scaling, Applications and Ethics

Spring 2016
Information School, University of Washington
Lectures: Monday (C), Tuesday (A) 5:30 - 8:20, BLD 070

Course Website (A): https://canvas.uw.edu/courses/1041433
Course Website (C): https://canvas.uw.edu/courses/1056835
Instructor: Jevin West
Office Hours: Wednesdays, 1:30 - 2:30 pm, MGH 310E

TA: Lavi Aulck
Office Hours: TBA, MGH 310

Course Description

This is the third course in the Data Science series. The first two (INFX 573 and INFX 574) are required to take this class[1]. INFX 573 lays the foundation in statistics and probability. INFX 574 concentrates on machine learning. This course (INFX 575) focuses on data scaling, applications in data science and ethical considerations. This will include exercises and lectures on cloud computing (AWS, Azure), web scrap-

[1]There may be some exceptions made, but this will require the instructors permission

ing, SQL and NoSQL databases, and information visualization. Students will apply methods from the fields of network science, natural language processing and information theory to data sets provided in class and data sets selected by the students. The goal of this class is to gain experience within the full pipeline of data science – from data gathering and storage to analysis and presentation.

The primary output for the class will be a group project. There will also be three quizzes and three assignments. The programming languages will primarily be R and Python, but there will be some programming in javascript and unix shell scripting.

Each year, there is a different theme for the data exercises. This year the theme is **Data Science for Science**. We will use data and exercises from this field. However, students will be allowed to pick any data set for their group project.

Course Outline

The course outline includes the following topics. Each topic is a class in and of itself. The class will move quickly through each topic so that students are at least exposed to these important topics in data science.

- Cloud Computing (AWS, Azure)
- Web Scraping
- Big Data Management (SQL and NoSQL)
- Data Processing (MapReduce, Spark)
- Natural Language Processing
- Applications in Network Science & Information Theory
- Information 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. 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 (3)	120 pts
Quizzes (3)	90 pts
Participation	90 pts
Total	**500 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

March 28,29: Syllabus, Logistics, Cloud Computing

Week 2

April 4,5: Web Scraping, MySQL

DUE: Quiz #1, Team Assignments

Week 3

April 11,12: NoSQL

DUE: Assignment #1 (web scraping)

Week 4

April 18,19: MapReduce, Hadoop

DUE: Quiz #2 (NoSQL, MapReduce), Proposal

Week 5

April 25,26: Data Ethics, D3

Guest Speaker: TBA (Monday 7:30)

DUE: Assignment #2 (Ethics Essay)

Week 6

May 2,3: Applied NLP (topic modeling, LDA)

Optional: Mid-quarter evaluations

DUE: Preliminary Data Analyis

Week 7

May 9,10: Applications in Information Theory

DUE: Quiz #3 (jargon paper)

Week 8

May 16,17: Spark

DUE: Assignment #3 (jargon distance)

Week 9

May 23,24: Applications in Network Theory

Guest Speaker: TBA (Tuesday, 7:30)

DUE: Final Paper

Week 10

Tuesday, May 31: Presentations

DUE: Presentations, Website, Team Evaluations

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
Team Assembly	5	Week 2
Proposal	20	Week 4
Preliminary Data Analysis	25	Week 6
Final Paper	100	Week 9
Presentation	30	Week 10
Website	15	Week 10
Team Evaluation	5	Week 10
Total	**200**	

*Updated rubrics for each aspect of the group project can be found on canvas. Below are some general guidelines.

Proposal

- Timely submission (5 pts)
- Motivation: Why are you investigating this question? (3 pts)
- Data: Where will you get your data? How will you acquire your data (e.g., API, scraping)? Provide links to your data. (3 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? (3 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. (3 pts)
- Plan: What is your plan for finishing this project by the end of the quarter? Who will be responsible for the different tasks? (3 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. This should be 3 - 4 pages. 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 and data description (10 pts)
- Next steps of the analysis (5 pts)
- Length: approximately 2-3 pages

Paper

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

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)

Website

- Timely submission (5 pts)
- Team and project description (1 - 2 paragraphs) (5 pts)
- Project figure (5 pts)

Team Evaluation

You will be asked to evaluate each member's effort and contribution to the project. You can break this up in percentages or by area of project (e.g., writing, analysis, data management). You also provide explanations for each. The evaluation can be as short as one sentence or as long as 1 page. We will not share these evaluations with other team members.

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