Course Recap

Everything Data
CompSci 216 Spring 2019
Today’s class

• Course Recap

• Course Feedback

• Project Work
Announcements (Thu. Apr 18)

• Grading done
  – HW 1-8
    • Graded
    • Rubrics with solutions in \textit{READONLY} repo
  – Labs 1-11
• Regrades processed*
• Project Critiques
  – Due Tuesday 4/23 at 5:00am

• Please do course evals
  – Complete course evaluations on \url{http://assessment.trinity.duke.edu/students-course-evaluations}

• Info on final project critiques, posters, report, and grading will be posted on Piazza
Data and ...

Health

Sports

Culture

Social Good

Science
Goal of this course

Learn concepts and tools needed for data driven inquiry and decision making
Data Cleaning

Data can have inconsistent formatting, errors and missing values.

- Regular Expressions (lec 1, lab 1)
- Google OpenRefine (lab 1)
- Record linkage (lec 3, lab 3)
Modeling/Representing Data

Data does not always fit in a single table or a spreadsheet

• Relational Databases (lec 2, lab 2)
• Graph Data (lec 8, lab 9)
Querying Data

High level languages for querying relational databases

- SQL (lec 2, lab 2)
- Postgres
Quantifying Uncertainty

Probability and Statistics help quantify whether data provides evidence for a fact.

- Probability distributions
- Hypothesis Testing
- Bayes Theorem
Learning (lec 5, lab 4)

Predict a label (Y) for unseen data points based on labeled training data

- Naïve Bayes Classifiers
- K-NN Classification
- Regression
- Decision Trees, SVMs, etc.
Clustering (lec 7)

When data are unlabeled, clustering algorithms can help reveal high level patterns in data

• K-means
• Hierarchical Clustering
Analyzing Text Data (lec 6, lab 6)

Counting words and identifying salient tokens can help reveal interesting patterns

- Bag of words model
- Language models
- Tokenization
- Stemming
- TFIDF
- Document similarity
Analyzing Graph Data (lec 8, lab 8)

Descriptive statistics about graphs reveal interesting properties of social interactions

- Degree Distribution
- Clustering coefficient
- Graph models and Power law
Analyzing Massive Data (lec 12)

Parallel processing and approximation help analyze massive amounts of data quickly

- The Map-reduce model
- Spark
- Sampling
Visualizing Data (lec 9)

The right visualization helps tell a story with the data

- Python visualization
- Viz Challenge
The dark side of the force...
Data can mislead …

Quantify the support and uncertainty before drawing conclusions from data

• Fact Checking (lab 2)
• Train-test Split / Overfitting (lab 4)
• P-Hacking (lab 5)
• Hypothesis tests (hw 4)
Easy to abuse ...

Data analytics can ...

... be unfair

... be unethical

... violate privacy

• Privacy Violations (lec 3)
• Differential Privacy (lec 10)
• Fairness (lec 11)
Data Analysis in the real world

• Choose the languages/tools/libraries
• Stay in sync with other members of your team
• Identify and fix the bugs/bottlenecks in your code and sometimes other people’s code
• Cope with the systems issues that arise during a project to work at scale (lec 12)
• Problems are open-ended - often up to you to redefine problems in a better way
Data Analysis in the real world

- Git
- Jupyter
- RTM (there’s a letter missing in there)
- Google/GitHub/Stack Overflow et al.
- Open source Python toolkits like scipy and numpy
Hope you learned *something* about everything data

What’s next?
Take more specialized CS courses in …
  … databases (316)
  … algorithms (330)
  … machine learning (371, 671)
  … AI (570)
  … algorithm in the real world (290.02)
  …