Introduction & Forming Groups

CSE481 Data Science Capstone

Tim Althoff

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Our plan for zoom

- Turning on video is optional but very appreciated ☺
- Let's make this engaging! Ask your questions through zoom chat!
  - If you know the answer, feel free to reply in the chat ☺
  - I will ask you questions, too! Use chat to reply by default, but we will also a variety of activities to report back from via mic/audio.

- Tim's office hours will be right after class on Tuesdays to answer any questions. Additional TA office hours.
- You will need a working camera and microphone to participate in course activities and group project.
- **We expect all students to participate in all lectures. You cannot appropriately participate in this class through asynchronous recordings.**

Data contains value and knowledge

Data Science

- **But to extract the knowledge data needs to be**
  - **Stored (systems)**
  - **Managed (databases)**
  - **And ANALYZED ← this class**

What is science?

- From the Latin word scientia, meaning **knowledge**
- A **systematic** enterprise that builds and organizes knowledge in the form of **testable explanations and predictions** about the universe

What this course is about

- Data Science seeks to discover new knowledge by answering questions through data

- It's not all about machine learning
- But some of it is

What data science is **not**

https://xkcd.com/1838/

How to turn observational, biased, **scientifically "weak" data** into strong scientific results?

More next week ☺

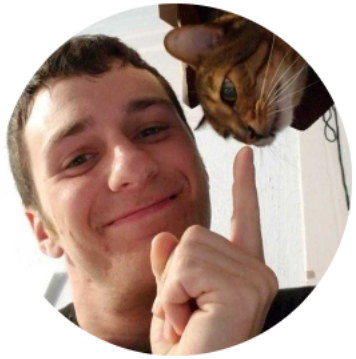
What will we learn?

- End-to-end process of data analysis performed with code
- Not limited to statistical modeling or machine learning, but rather the complete process, including transformation, exploration, modeling, and evaluation choices
  - We focus specifically on all the aspects that are NOT covered in ML / database / data viz courses.
- Hands-on experience on how to work in groups to pursue a complete data science project

Course Staff

Teaching Assistants

Mike Merrill
(Head TA)



Galen Weld

CS481DS Course Staff

- **Office hours:**
  - See course website for TA office hours
    - https://courses.cs.washington.edu/courses/cse481ds/
    - **We start Office Hours this week (Oct 6)**
  - **Tim:** Right after class
  - **TA office hours:** TBA, two 1h slots per week

Logistics: Communication

- **Ed Q&A website:**
  - https://us.edstem.org/courses/2572/discussion/
  - Use Ed for **all questions** and public communication & announcements
    - Search the forum before asking a question
    - Please tag your posts and please no one-liners
- **For emergencies & personal matters, email course staff always at:**
  - *cse481ds-instructors@cs.washington.edu*
- **We will post course announcements to Ed (make sure you check it regularly)**

Work for the course: Group Project

- Project deliverables at different stages

- Project Pitch (Individual): 2%
- Project Plan (Group): 4%
- Validity Reflection Presentation (Group): 4%
- Midpoint Presentation Video (Group): 15%
- Midpoint Feedback Reflection and Action Plan (Group): 4%
- Final Presentation Video (Group): 25%
- Final Project Report (Group): 25%

- We expect all students to participate in all lectures. Participation is required to get credit for your project.

Work for the course: Reflections and Feedback

- Project Selection Reflection (Individual): 1%
- Example Paper Reflection (Individual): 2%
- Spark Colab (Individual): 1%
- Summary of Individual Contribution to Project (Individual): 1%
- Final Reflection (Individual): 2%

- Feedback to other students all throughout quarter (Individual): **14%**
  - We will use a form to keep track of feedback.
  - We will also take note of contributions on Ed.

When to submit?

- **All deliverables are due at midnight PST before class indicated in our course website.**

- **Since the deliverables are interdependent between groups, we will not have a late policy.**

Prerequisites

Students should experience with

- Programming: Python
- Data Structures: CSE 332
- Probability: CSE 312
- and at least one of
  - Machine Learning: CSE 446
  - Data Visualization: CSE 442
  - Data Management: CSE 344
- It can be very helpful to form groups in a way that cover these areas well

Collaboration Policy & Academic Integrity

- **We'll follow the standard CS Dept. approach:** You can get help, but you *MUST* acknowledge the help on the work you hand in
  - www.cs.washington.edu/academics/misconduct
- Failure to acknowledge your sources is a *violation of academic integrity*

Final Thoughts

- **CS481DS is fast paced!**
  - Requires programming maturity
- **Course time commitment:**
  - Very significant capstone project
- **Brand new course – expect this to be bumpy**
  - We are committed to a great learning experience.
  - Kindly let us know where we can improve.

- **It's going to be fun and hard work. ☺**

3 To-do items

- **3 to-do items for you:**
  - **Make sure you can access Canvas & Ed**
  - **Plan your course project (topic, team, dataset)**
  - **Get ready for your project pitch in a few minutes**

- **Additional details/instructions at http://www.cs.washington.edu/cse481ds**

Project Pitches & Forming Groups

Pitches & Process

- We will screencast the project pitches
- Be ready to unmute and give your pitch when it is your turn (in slide order; we shuffled the order)

- Your project pitch is just a starting point – it is to find a team with overlapping interests. Make it your own project. It will naturally evolve throughout the course

- When we are done with pitches, we will give you time to connect with each other and form project groups
  - You need to arrive at six teams or less. Any other team size we will have to split or merge in the end.
  - Instructor and TAs will be available to help/discuss in zoom.
  - We recommend Ed for initial connections and we will create zoom breakout rooms as needed.

Pitches & Process (2)

- To support the process. While others are giving pitches, record your interests in this spreadsheet: http://bit.ly/dscapstoneprojects
  - Now: please add your name and email into a new column
  - While others are pitching: If you may be interested to join their project, put a note into the respective cell (3 super interested, 2 interested, 1 interested if I can't find another project)
- Use Ed Discussion board for matching discussions. We can create zoom breakouts for you.
- **By 6pm PST today** – Fill out spreadsheet with your group (max 6 groups). We will merge / break up groups as needed.