Introduction & Forming Project Groups

CSE481DS Data Science Capstone

Tim Althoff

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Welcome ☺

- Welcome to our Data Science Capstone course!
- We’re excited for each and everyone in this course!

- This is a relatively new course – expect some bumps along the way
- Please give us feedback!
  - We benefit from everyone who did this last year and allowed us to adjust the course to be more helpful in supporting your learning!
  - https://bit.ly/cse481ds-au23-feedback

COVID-19 & Expectations

- We understand that the last 2.5 years have been particularly challenging for everyone.
- We expect all students to participate in all lectures. You cannot appropriately participate in this class through asynchronous recordings.
- The long single course slot is suboptimal, but was the best option available given various scheduling constraints. We will interrupt each class for a short break.
- Please give feedback on your experience.
  - This was the most disagreed on aspect last year.

POPULAR SCIENCE
THE FUTURE NOW
THE CONTROL CENTERS
Using Data to Feed the World, Solve Cold Cases, Battle Malware, Predict Our Fate
OFFICER ALGORITHM
Can a Crime Be Prevented Before It Begins?
NEW WAYS OF SEEING
A Gallery of Extraordinary Infographics
PLUS
Juan Enriquez Reprograms Life
James Gleick Unsplits the Bit
AND
Lawrence Weschler Questions the Cloud
SPECIAL ISSUE
DATA IS POWER
HOW INFORMATION IS DRIVING THE FUTURE

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**

THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

https://xkcd.com/1838/

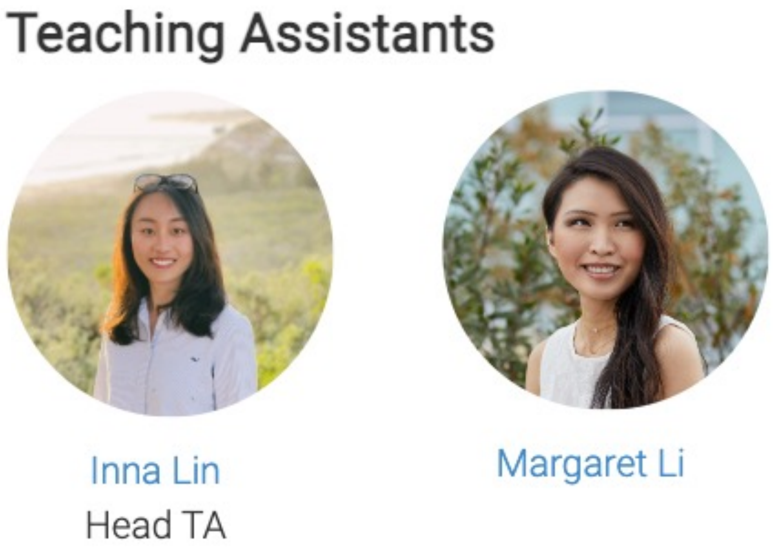
How to turn observational, biased, **scientifically "weak" data** into strong scientific results?

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

Inna Lin
Head TA

Margaret Li

Margaret and Inna are both PhD students in my data science group. Both have significant data science experience in industry as well.
This class is special in that you get lots of access to both of them.
To get the most out this class, interact with them as much as you can.

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 3)**
  - **Tim:** Right after class (starting today)
  - **TA office hours:** TBA, two 1h slots per week
    - We will survey all groups for conflicts to optimize our OH for you
    - We will post a sign-up spreadsheet on Ed
    - 20 min slots to reserve
    - We expect that you make use of these at least every two weeks. Don't wait until your project is on fire!
    - Historically, groups that seek regular feedback, including on assignments before submitting, end up with very good projects.

Logistics: Communication

- **Ed Q&A website:**
  - https://edstem.org/us/courses/4802
  - 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)**

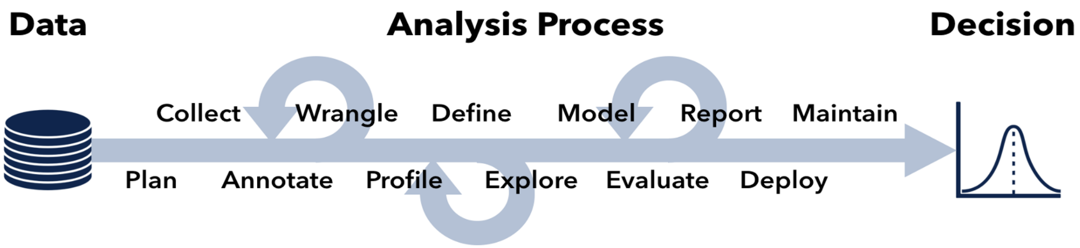
Work for the course: Group Project

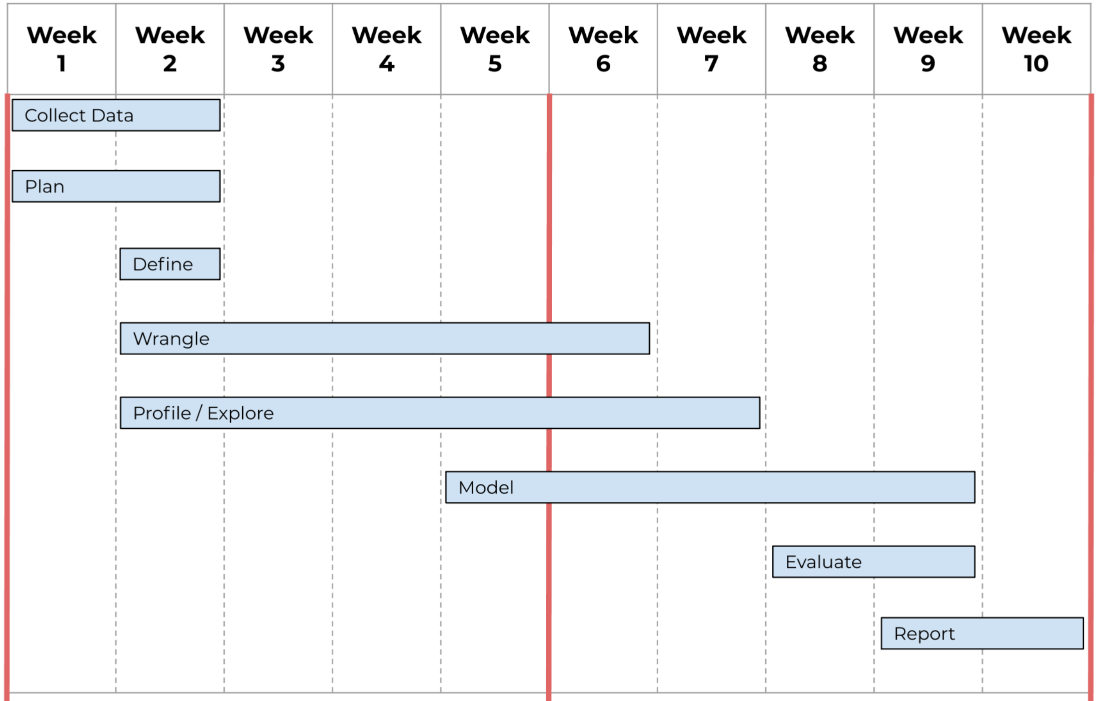
- Project deliverables at different stages

- 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%
  - This is the final product. See website for details!
  - Novelty doesn’t matter (for grade). Thoughtful and appropriate analysis process for the RQ is what we’re looking for.

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

Very Rough Project Timeline





Project Pitch

- What research question do you want to ask?
- What data will you use to answer your question?
- What makes you particularly interested in this question or data?

Midpoint Presentation

By this point you should have:

- Collected **ALL** of your data
- Cleaned + Wrangled your data
- Understand whether your major risks could sink your project and how to address risks
- **At least** initial steps towards answering your question and initial results to show

Final Report

Submit a detailed final report which includes (**among others**):

- A description of why your question is important to answer
- Your main findings
- An evaluation of whether you were able to satisfactorily answer your question

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%

- Participation & feedback to other students all throughout quarter (Individual): **14%**
  - We will use a form to keep track each class
    - Say your name before asking questions, making comments
  - We will also take note of contributions on Ed

NOT SURE IF HARSH GRADING
OR IF I JUST DID REALLY BADLY

Grading Philosophy

- **Goal of grading:** clearly, accurately, consistently, and fairly communicate learning progress and achievement to you
- We want to help you learn and give you **optimal feedback**
  - In our grading getting **50% means meeting expectations.** Above 50% means you're doing a particularly good job!
  - We do this to give you early warning signs when we believe you need to address something. We don't believe it's fair for you to get surprised by a low grade late in the quarter. We want to give you chances to listen to feedback and improve.
  - We don't do this between 90 and 100% because this adds more noise and makes it harder for you to make up and lower early grades.
  - Communicating your idea, data and results way are important skills. That's why we pay attention to timing and allow large deviations to influence the final grade.
  - We will post grading statistics on Ed for transparency
- In the end, we will curve all grades generously
  - If you do an honest attempt at answering the questions, listen to our feedback, and try to implement it, you will not have to worry about your grade.
- Make sure you **distribute project work fairly** among your group. If we need to we will give non-uniform grades to group members.

Time Commitment

- 5 credits ~ 15h/week total time commitment
- We have designed this course to stay within these bounds.
- However, work load varies if
  - Your project has fewer than expected members.
  - You're missing relevant prerequisites.
  - You do not consistently invest in your project each week as recommended.

When to submit?

- All deliverables are due at midnight PST before class indicated in our course website.
- Assignments are uploaded to Canvas.
- Since the deliverables are interdependent between groups (e.g. for feedback in class), we will not have a late policy.
  - Communicate and rely on your group when you need some slack

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*

So what about Language Models? ☺

- It is hard to find someone who has not heard about ChatGPT and related tools, and these tools are undeniably useful for generating ideas, providing suggestions, editing, and more. In this class, we will ask you to follow specific, ethical guidelines when using generative AI such as ChatGPT.
  - See course website for details
- For example, you are responsible for any assignment, whether created by yourself or generative AI. You need to acknowledge AI use and share prompts alongside your assignments.
- Best outcomes are likely achieved through generative AI use in moderation.

Final Thoughts

- **CS481DS is fast paced!**
  - Requires programming maturity
- **Course time commitment:**
  - Very significant capstone project
- **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 put your project pitches on the screen
- Be ready to give your pitch when it is your turn (in slide order; we shuffled the order)
- Listen carefully to others and think about who you could team up with. You likely want to talk to them in a few minutes.

- 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.
  - In addition to overlapping interests, being aligned on expectations on course performance can help. Do you care deeply about very good grades? Are you still thinking about dropping this course?

- When we are done with pitches, we will give you time to connect with each other and form project groups. Use this room.
  - 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.

- **By 5pm PST tomorrow (10/4)** – Fill out spreadsheet with your group (max 6 groups). If we have to, we will merge / break up groups as needed.

Due next week:

- Project Plan & Project Selection Reflection
- Details on Project Plan (see template on website)
  - Six slides, one per question in template.
  - Pre-recorded video presentation six minutes max.
    - We mean this! To be fair across groups we will stop you and include this in our grading.
  - Upload the recording on Canvas.
  - Be ready in class to receive feedback from others.
  - Be ready to give feedback to the other groups after their presentation.

- We do these steps early to help you formulate a successful and fun project. You have lots of freedom in choosing your project topic. Use it ☺

Thank you for sharing your feedback with us!

https://bit.ly/cse481ds-au23-feedback