Carnegie Mellon University
HeinzCollege

Practical

94-775 Unstructured Data Analytics

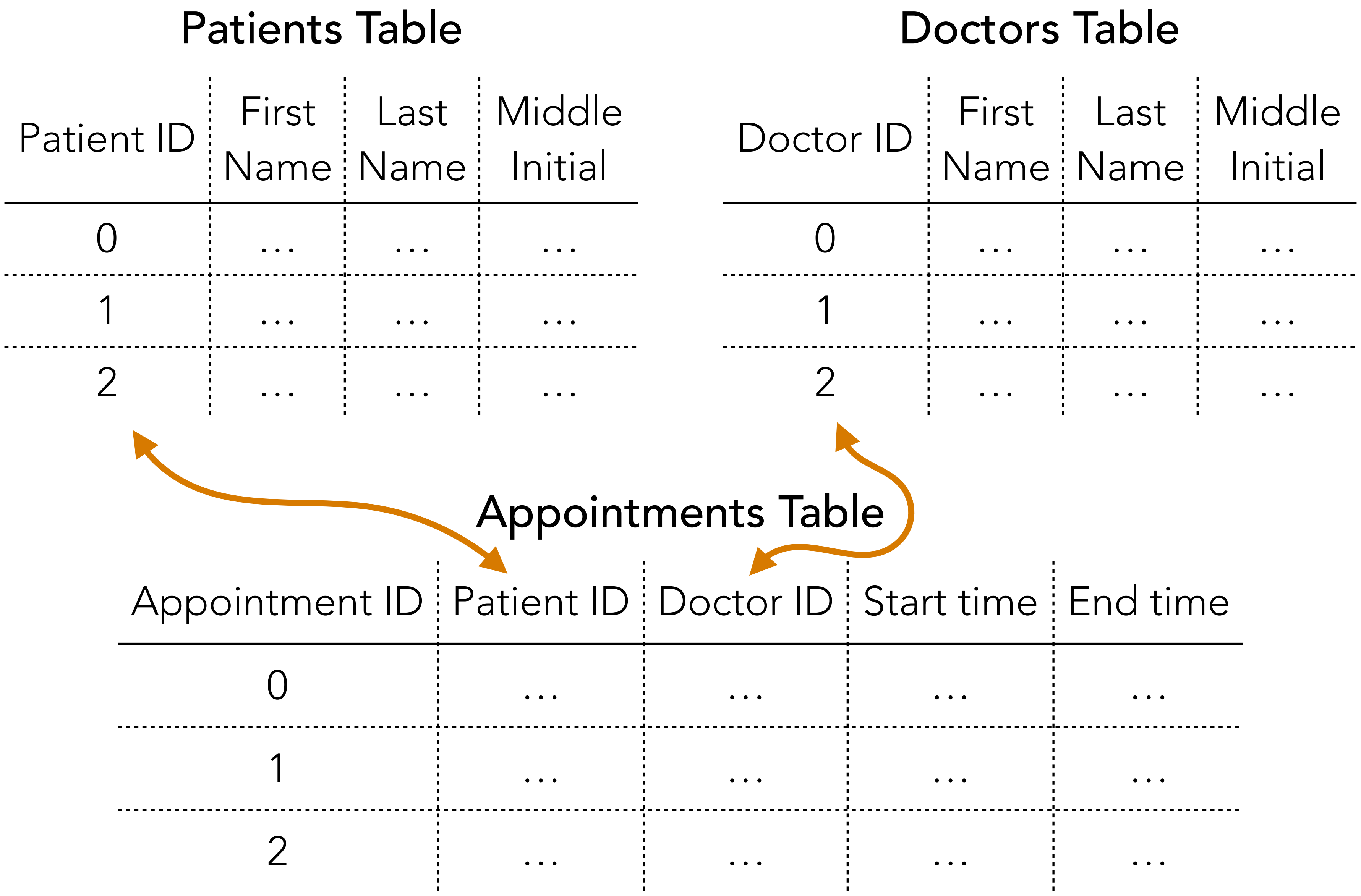
Lecture 1: Course overview

Slides by George H. Chen

What is “unstructured data”?

Structured Data

Well-defined elements, relationships between elements



Patients Table

Patient ID	First Name	Last Name	Middle Initial
0	…	…	…
1	…	…	…
2	…	…	…

Doctors Table

Doctor ID	First Name	Last Name	Middle Initial
0	…	…	…
1	…	…	…
2	…	…	…

Appointments Table

Appointment ID	Patient ID	Doctor ID	Start time	End time
0	…	…	…	…
1	…	…	…	…
2	…	…	…	…

Can be labor-intensive to collect/curate structured data

Unstructured Data

No pre-defined model—elements and relationships ambiguous

Common examples:

- Text
- Images
- Videos
- Audio

Often: Want to make decisions using multiple types of unstructured data, or unstructured + structured data

Of course, there *is* structure in "unstructured" data but it is not neatly spelled out for us

We have to extract what elements matter and figure out how they are related!

Just because something *can* be stored as any of these doesn't mean that it must be unstructured!

Example 1: Health Care

Is a patient at risk of getting a nasty disease?

Data

- Chart measurements (e.g., weight, blood pressure)
- Lab measurements (e.g., draw blood and send to lab)
- Doctor’s notes
- Patient’s medical history
- Family history
- Medical images

Example 2: Electrification

Where should we install cost-effective solar panels in developing countries?

Data

- Power distribution data for existing grid infrastructure
- Survey of electricity needs for different populations
- Labor costs
- Raw materials costs (e.g., solar panels, batteries, inverters)
- Satellite images

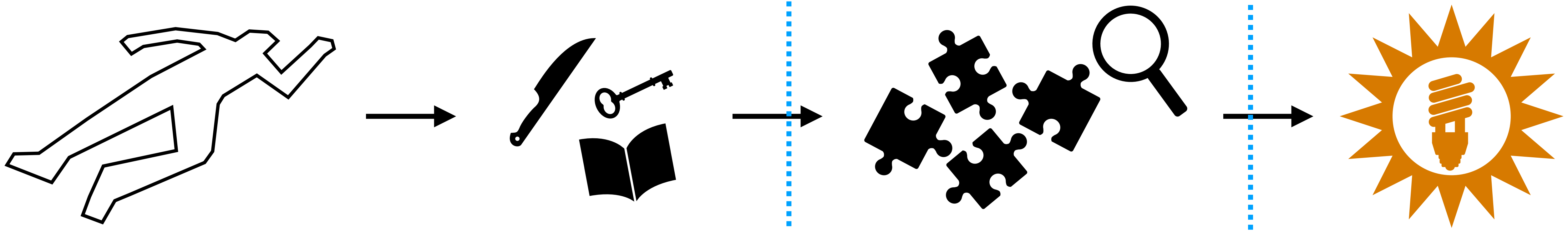
Generative AI Technologies

- AI assistants like (Chat)GPT, Gemini, Claude, Llama, and DeepSeek will continue to get better over time
  - As of April 2023: GPT4 got a B on a quantum computing final exam https://scottaaronson.blog/?p=7209
  - As of July 2024: AlphaProof (built on Gemini) achieved silver for International Math Olympiad problems https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/
- These AI assistants are trained to process unstructured data!
  - We'll gradually build up to how a basic version of GPT works
- For this class, I view whether you get help from AI *to be the same as whether you got help from a human*
  - Regardless of whether you get info from an AI or a human:
    I want *you* to be able to tell whether this info is correct or not
    - Exams will be paper & pencil with no electronics allowed

Image source: African Reporter

Unstructured Data Analytics (UDA)

Much like how many murder mysteries go unsolved,
many data analysis (unstructured or not) problems can be extremely difficult



Question

The dead body

This is provided by a domain expert

Data

The evidence

Some times you have to collect more evidence!

Finding Structure

Puzzle solving, careful analysis

Exploratory data analysis

Insights

When? Where? Why? How? Perpetrator catchable?

Answer original question

Not detailed in lecture but addressed by your final project, which must address a public policy problem

Sometimes: we aim to solve a prediction problem

UDA involves *lots* of data ➔ write computer programs to assist analysis

94-775

Prereq: Python programming

We require 90-803 “Machine Learning Foundations with Python" (which has Python as a prerequisite)

Part I: Exploratory data analysis

Part II: Predictive data analysis

94-775

Part I: Exploratory data analysis

Identify structure present in "unstructured" data

- Frequency and co-occurrence analysis
- Visualizing high-dimensional data/dimensionality reduction
- Clustering
- Topic modeling

Part II: Predictive data analysis

Make predictions using known structure in data

- Basic concepts and how to assess quality of prediction models
- Neural nets and deep learning for analyzing images and text

Course Goals

By the end of this course, you should have:

- Lots of hands-on programming experience with exploratory and predictive data analysis
- A high-level understanding of what methods are out there and which methods are appropriate for different problems
- A *very* high-level understanding of how these methods work *and what their limitations are*
- The ability to apply and interpret the methods taught to solve problems faced by organizations

I want you to leave the course with practically useful skills solving real-world problems with unstructured data analytics!

As we go from covering classical methods to modern ones, it's good to understand *why* newer methods were developed

UDA technologies change very rapidly! GPTs might be hot today but be out of fashion tomorrow!

Course ~~Textbook~~ Materials

No existing textbook matches the course… =(

Main source of material: lectures slides

We'll post supplemental reading as we progress

Check course webpage
http://www.andrew.cmu.edu/user/georgech/94-775/

In general, announcements & links to all course-related things will be in Canvas

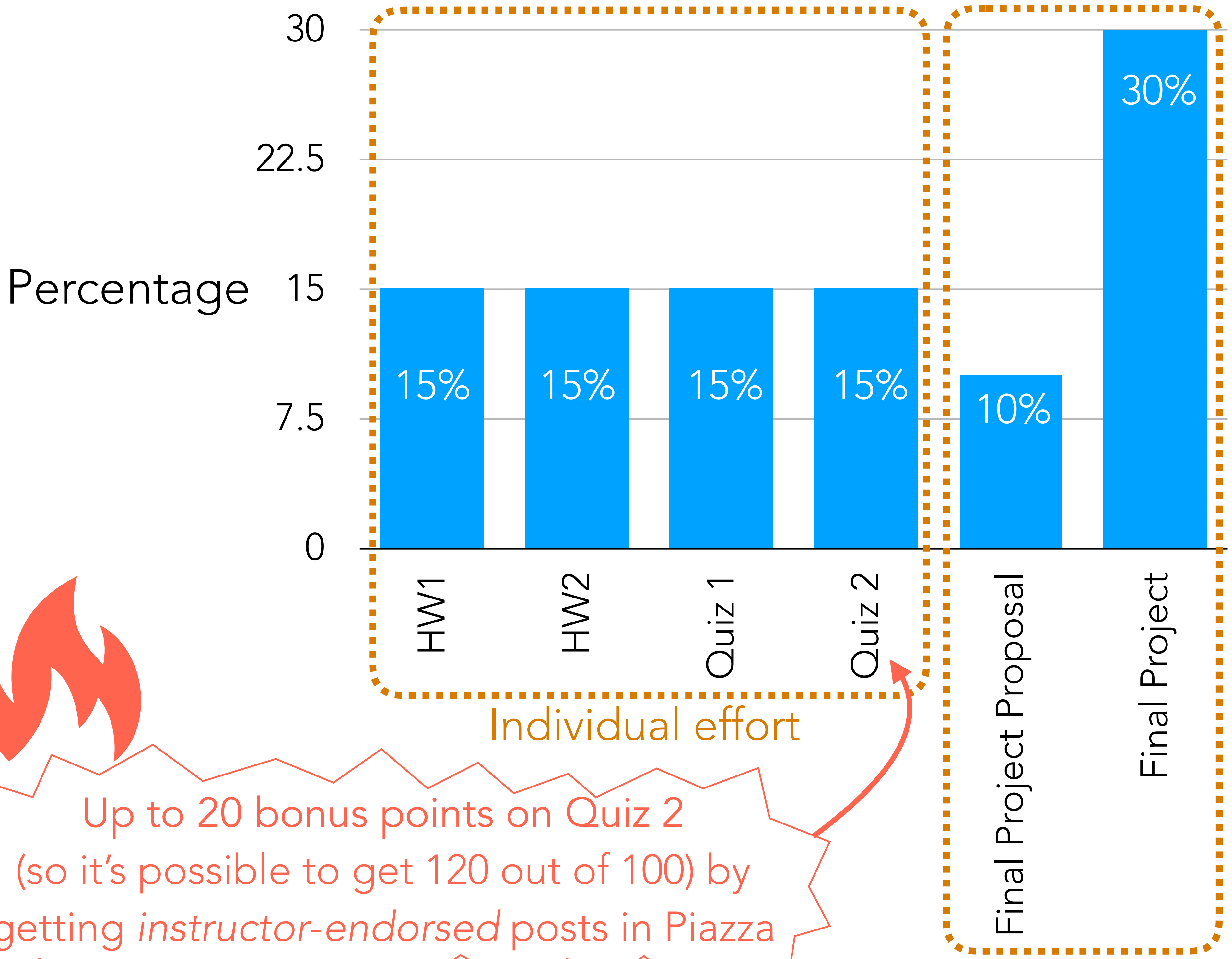
Homework assignments are submitted via Gradescope (link is within Canvas)

Please post questions to Piazza (link is within Canvas)

canvas gradescope® by Turnitin piazza

Deliverables & Grading

Contribution of Different Assignments to Overall Grade



Individual effort

Group effort

Letter grades are assigned based on a curve

Up to 20 bonus points on Quiz 2
(so it's possible to get 120 out of 100) by
getting *instructor-endorsed* posts in Piazza

Individual Effort Assignments: Collaboration & Academic Integrity

- If you are having trouble, ***please ask for help!***
  - We will answer questions on Piazza and will also expect students to help answer questions!
  - **Do not post your candidate solutions on Piazza**
  - For code: post smallest snippet, how you know it's buggy (error message, etc), & what you've already tried to resolve the issue
- In the real world, you will unlikely be working alone
  - We encourage discussing concepts
  - Please acknowledge classmates you talked to or resources you consulted (e.g., ChatGPT, Gemini, stackoverflow)
- **For individual effort assignments, do not share your code with classmates (instant message, email, Box, Dropbox, AWS, etc)**
- **Do not use solutions from past semesters**

Penalties for cheating are severe: 0 on assignment, possibly fail the course ☹️

The Two Quizzes

This is a brand new exam format for 94-775!

Format:

- **In-person, on paper**
- Each quiz is **40 minutes**
- No electronics may be used during the exam (e.g., do not use a laptop, tablet, phone, calculator)
- Open notes (must be on paper and not electronic)

Back when I taught 94-775 years ago, there was only 1 exam and some students provided feedback saying the 1 exam is too few (and made it too "high-stakes"), so we're going to try this new format out!

Quiz 1: Friday March 28 during recitation slot

Quiz 2: Friday April 11 during recitation slot

There are no real past 94-775 exams that use this format

I will provide problems from real paper and pencil 95-865 exams to help you study (this is the other course that I teach)

Gradescope

- We’re using Gradescope for grading everything (homework & quizzes)
- Your homework will involve coding *but we will ask that you save your code notebook as a PDF and submit only the PDF*
  - **Since we will not be re-running your code, make sure that your PDF includes all the code output!**
- We will scan your quizzes and grade them on Gradescope

Final Project

- Must be done in a group of ~4-5 students
  - You choose your own groups
  - Final project proposals (~2 pages) are due **Tuesday April 1, 11:59pm (by email, 1 email per group)** & must specify who the group members are
- Required components will be stated in 2 slides
- Friday April 24 recitation slot: **final project presentations!**
- Final project reports are due **Monday April 28, 11:59pm** & consist of:
  - Jupyter notebook (edited down to be clean, concise)
  - Slide deck for your final project presentation
- The final slide deck could be modified compared to what is presented the Friday beforehand (e.g., to incorporate feedback)

Final Project Rubric

- **Policy question (15%):** what public policy question are you addressing? Please be clear and concise.
- **Data analysis (30%):** clearly state what part of your data are unstructured (some but not all of the data you are analyzing must be unstructured), and carefully justify every step of your analysis with supporting visualizations/intermediate outputs as needed
- **Code (30%):** your code should actually run!
- **Conclusions (15%):** come up with insights that are based on your quantitative data analysis and that address your original policy question
- **Presentation (10%):** how polished is your final project presentation? — this is based on the live presentation your group makes (changes made to the slides after the presentation don’t affect this score)

Final Project *Proposal*

- **Policy question:** what public policy question are you addressing? Please be clear and concise.
- **Data:** what data have you found that you want to analyze, and why is at least some portion of it unstructured?
- **Proposed analysis:** what specific methods do you want to try and why? In what way would these address your proposed policy question? Are there specific obstacles you think you will have to address? What would a "successful" analysis look like?

Important: you should have already downloaded & looked at the data

Common problem in previous years: some groups didn't actually look at the data, and then after they looked at the data, they had to completely change their proposal

Some final projects from the past years have already been posted on Canvas

Note: I have omitted Jupyter notebooks from these past projects though, and only included the proposals & final slide decks

python™

- The data science/machine learning tools available have changed *drastically* over the last few years
    - Working with most of the latest innovations from computer scientists requires some programming (at this point, Python is standard for machine learning research)
    - Also good to solidify your programming background by learning more languages
- We will be using Anaconda Python 3
  https://www.anaconda.com/

Late Homework Policy

- You are allotted 2 late days
    - If you use up a late day on a HW, you can submit it up to 24 hours late with no penalty
    - If you use up both late days on the same HW, you can submit up to 48 hours late with no penalty
- Late days are *not* fractional
- This policy is in place precisely to account for various emergencies (health issues, etc) and you will not be given additional late days
- There is no need to tell us if you're using a late day or not (we'll figure it out from submission timestamps)

Course Staff

TA: Johnna Sundberg

Instructor: George Chen

Office hours start next week (we’re still sorting out the schedule): details will be posted on Canvas