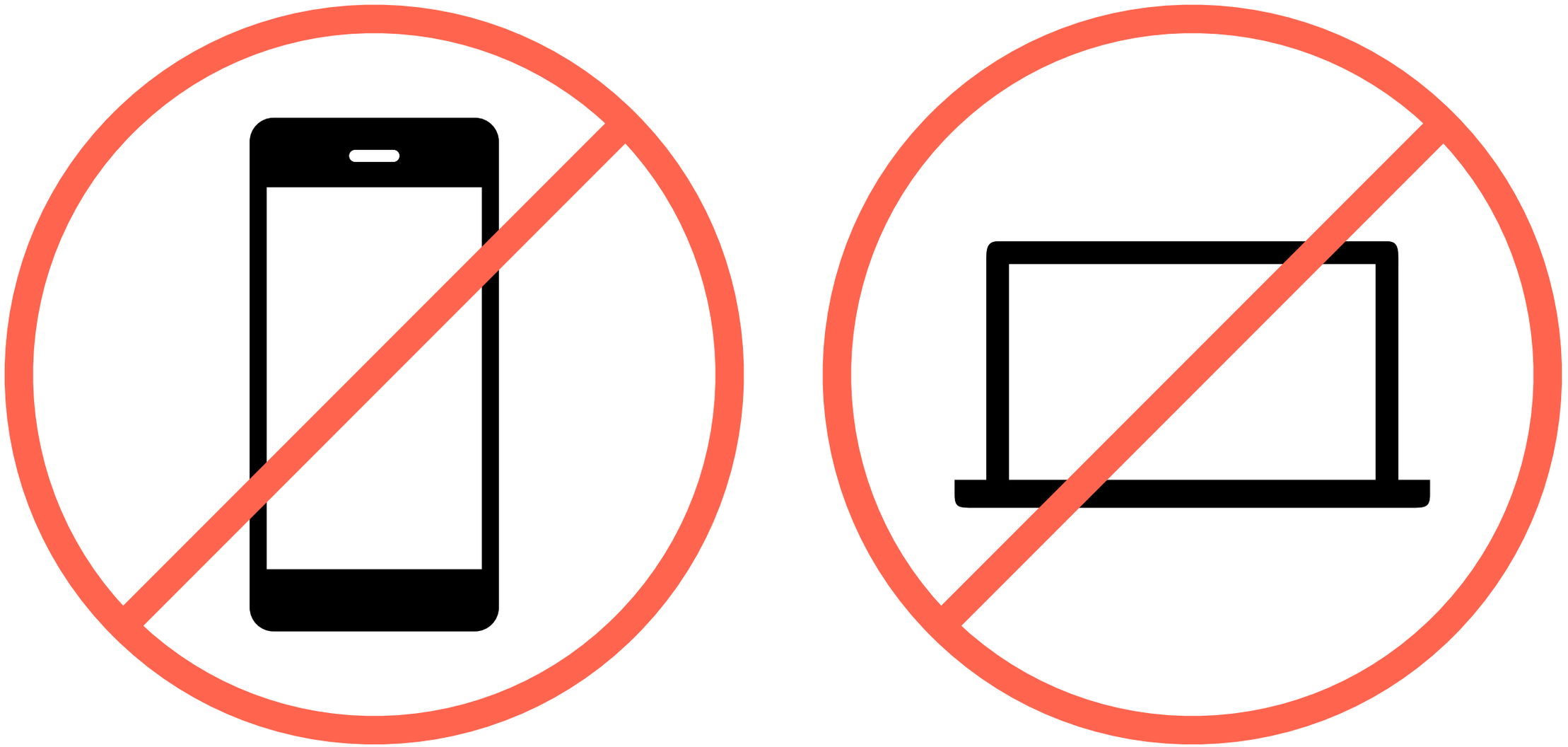
Carnegie Mellon University
Heinz College

94-775
Unstructured Data Analytics for Policy

George Chen

Cell Phones and Laptops

Just like what you'd expect in a movie theater



We don't want your device screens/sounds distracting classmates

Big Data

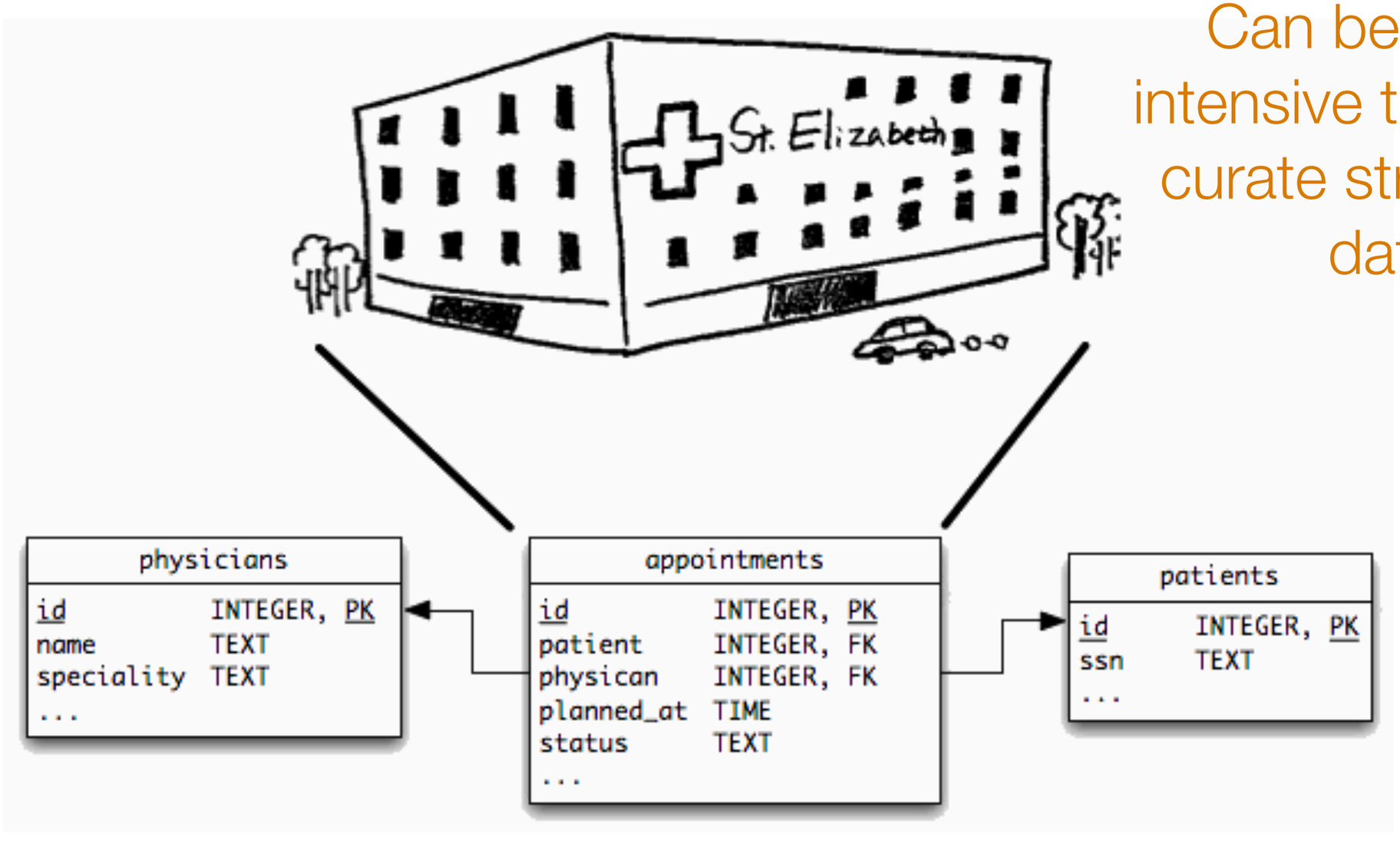
We're now collecting data on virtually every human endeavor

How do we turn these data into actionable insights?

Two Types of Data

Structured Data

Well-defined elements, relationships between elements



Can be labor-intensive to collect/ curate structured data

Image source: http://revision-zero.org/images/logical_data_independence/hospital_appointments.gif

Unstructured Data

No pre-defined model—elements and relationships ambiguous

Examples:

- Text
- Images
- Videos
- Audio
- Numerical measurements

Often: Want to use heterogeneous data to make decisions

Of course, there *is* structure in this data but the structure is not neatly spelled out for us

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

Example 1: Health Care

Forecast whether a patient is at risk for getting a 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

Example 3: Online Education

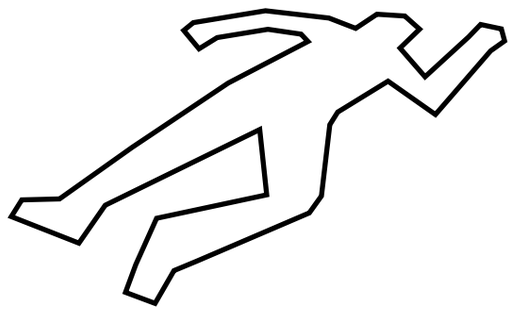
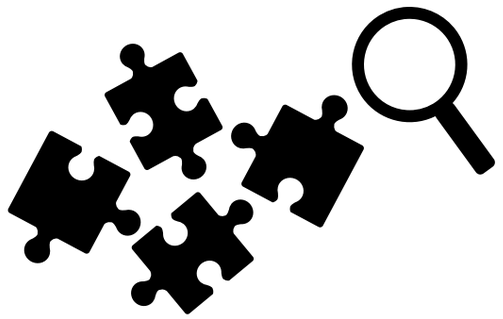
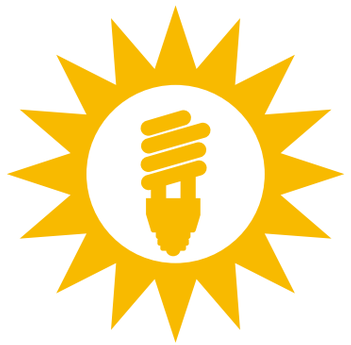
What parts of an online course are most confusing and need refinement?

Data

- Clickstream info through course website
- Video statistics
- Course forum posts
- Assignment submissions

Image source: African Reporter

Unstructured Data Analysis

Question	Data	Finding Structure	Insights
			
The dead body	*The evidence*	*Puzzle solving, careful analysis*	*When? Where? Why? How? Perpetrator catchable?*
This is provided by a practitioner	Some times you have to collect more evidence!	Exploratory data analysis	Answer original question

There isn't always a follow-up prediction problem to solve!

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

94-775 Course Outline (Tentative)

Part I: Python programming for data analysis

Fast crash course highlighting some key Python concepts

- Built-in data structures
- Control flow
- Numpy

Part II: Exploratory data analysis

Identify structure present in "unstructured" data

- Frequency analysis
- Visualization
- Clustering

Part III: Predictive data analysis

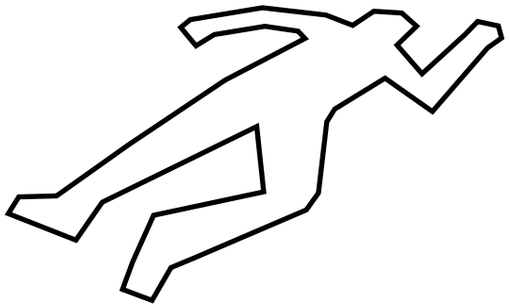
Make predictions using structure found in Part II

- Classical classification
- Neural nets and deep learning

Unstructured Data Analysis

Not detailed in lecture but addressed by final project

Question



The dead body

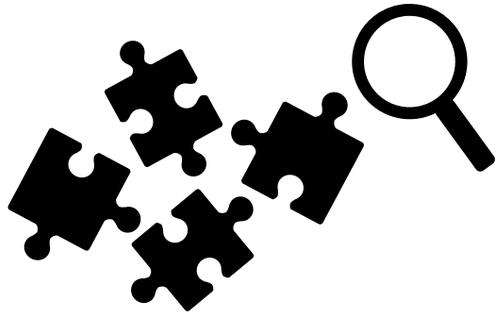
This is provided by a practitioner

Data

The evidence

Some times you have to collect more evidence!

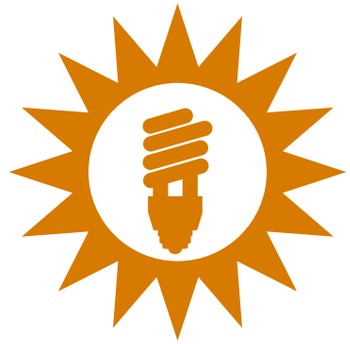
Finding Structure



Puzzle solving, careful analysis

Exploratory data analysis

Insights



When? Where? Why? How? Perpetrator catchable?

Answer original question

There isn't always a follow-up prediction problem to solve!

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

Course Goals

By the end of this course, you should have:

- 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

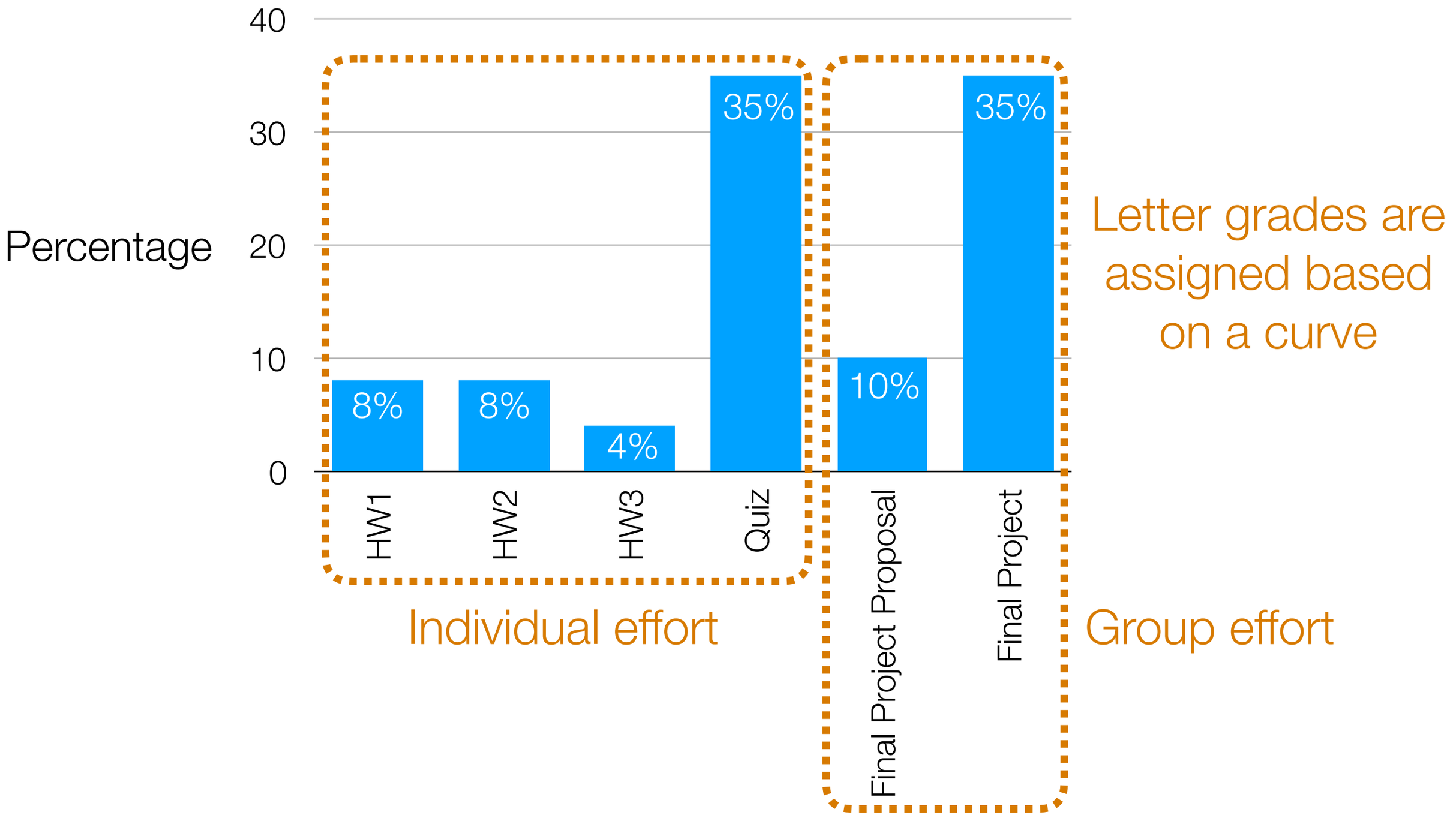
- A *very* high-level understanding of how these methods work

- The ability to apply and interpret the methods taught to solve a policy question

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

Deliverables & Grading

Contribution of Different Assignments to Overall Grade



Letter grades are assigned based on a curve

Individual Effort Assignments

- If you are having trouble, **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**
- In the real world, you will unlikely be working alone
  - We encourage you to discuss concepts/how to approach problems
  - Please acknowledge classmates you talked to or resources you consulted (e.g., stackoverflow)
- **For individual effort assignments, do not share your code with classmates (instant message, email, Box, Dropbox, AWS, etc)**

Penalties for cheating are severe: 0 on assignment, F in course =(

Final Project

- Must be done in a group of ~4 students
    - You can choose your own groups
    - Final project proposals (4 pages) are due April 10 at 3pm & must specify who the group members are
- Must address a policy question and involve UDA
- Must have a coding component written up as a Jupyter notebook that summarizes the key findings
    - This notebook serves as the final project report and is due **May 3, 3pm**
- Last week of lecture: final project presentations!

Course ~~Textbook~~ Materials

No existing textbook matches the course… =(

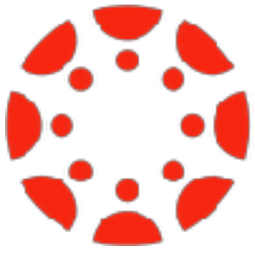
Main source of material: lectures slides

We'll post complimentary reading as we progress

Check **course webpage**

Assignments will be posted and submitted on **canvas**

Please post questions to **piazza (link is within canvas)**



canvas

piazza

Computing Environment

- We will be using **Anaconda (Python 3.6 version)**
  https://www.anaconda.com/what-is-anaconda/

- You will be submitting assignments in the form of **Jupyter notebooks**

Late Homework

- You are allotted 2 late days
    - If you use up a late day on an assignment, you can submit up to 24 hours late with no penalty
    - If you use up both late days on the same assignment, 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

Course Staff

Dylan Fitzpatrick

Runshan Fu

Teaching Assistants

George Chen

Instructor

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