95-865
Unstructured Data Analytics

George Chen
Carnegie Mellon University

Fall 2017 Mini-2

Big Data

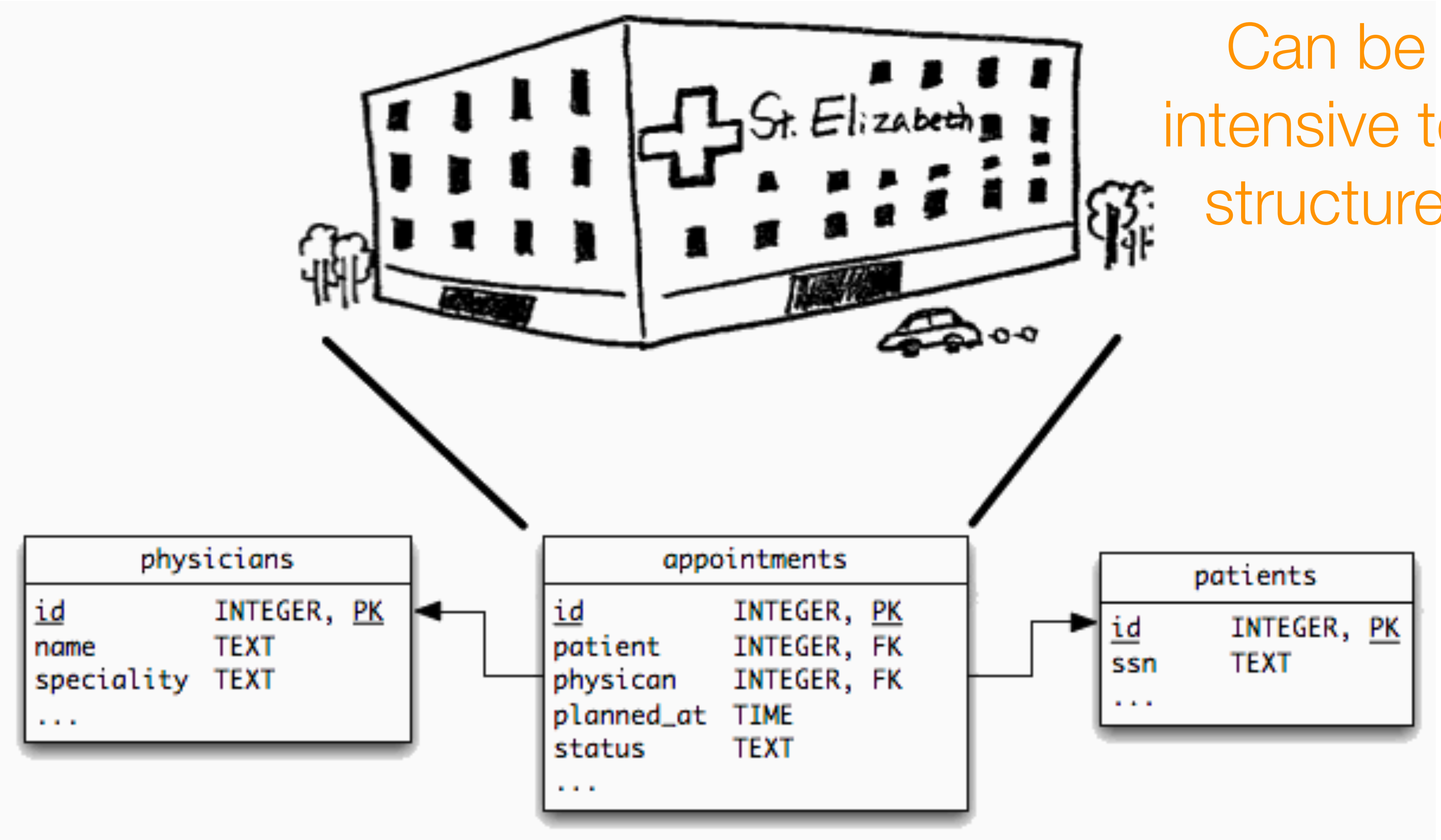
We're now collecting data on virtually every human endeavor

How do we turn these data into actionable insights?

Two Types of Data

Structured Data

Well-defined elements, relationships between elements



physicians

id	INTEGER, PK
name	TEXT
speciality	TEXT
...	

appointments

id	INTEGER, PK
patient	INTEGER, FK
physican	INTEGER, FK
planned_at	TIME
status	TEXT
...	

patients

id	INTEGER, PK
ssn	TEXT
...	

Can be labor-intensive to 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 we do not know it ahead of time

Example 1: Health Care

Forecast whether a patient is at risk for getting a disease?

Electronic health records

- 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

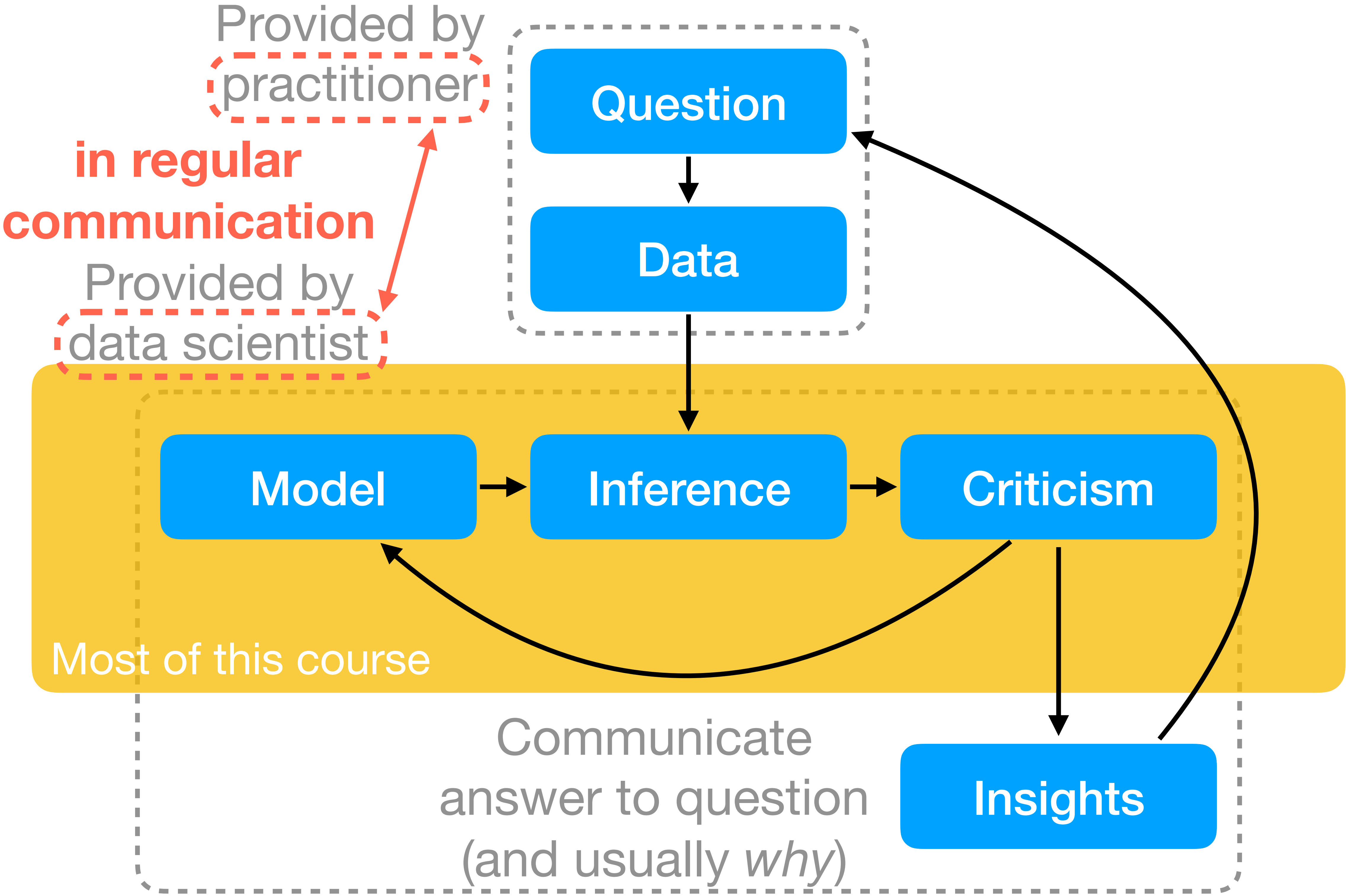
Example 2: Electrification

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

Geographic information system (GIS) & pricing 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

Sketch of Usual Workflow



Course Outline (Tentative)

Part 1: Identify structure present in “unstructured” data

Exploratory data analysis

- Frequency and co-occurrences
- Clustering
- Topic modeling (special kind of clustering)

Unsupervised learning

Part 2: Make predictions using structure found in part 1

Predictive data analysis

Supervised learning

- Basic classification and regression models
- Adaptive nearest neighbor methods
- Deep learning models for classification

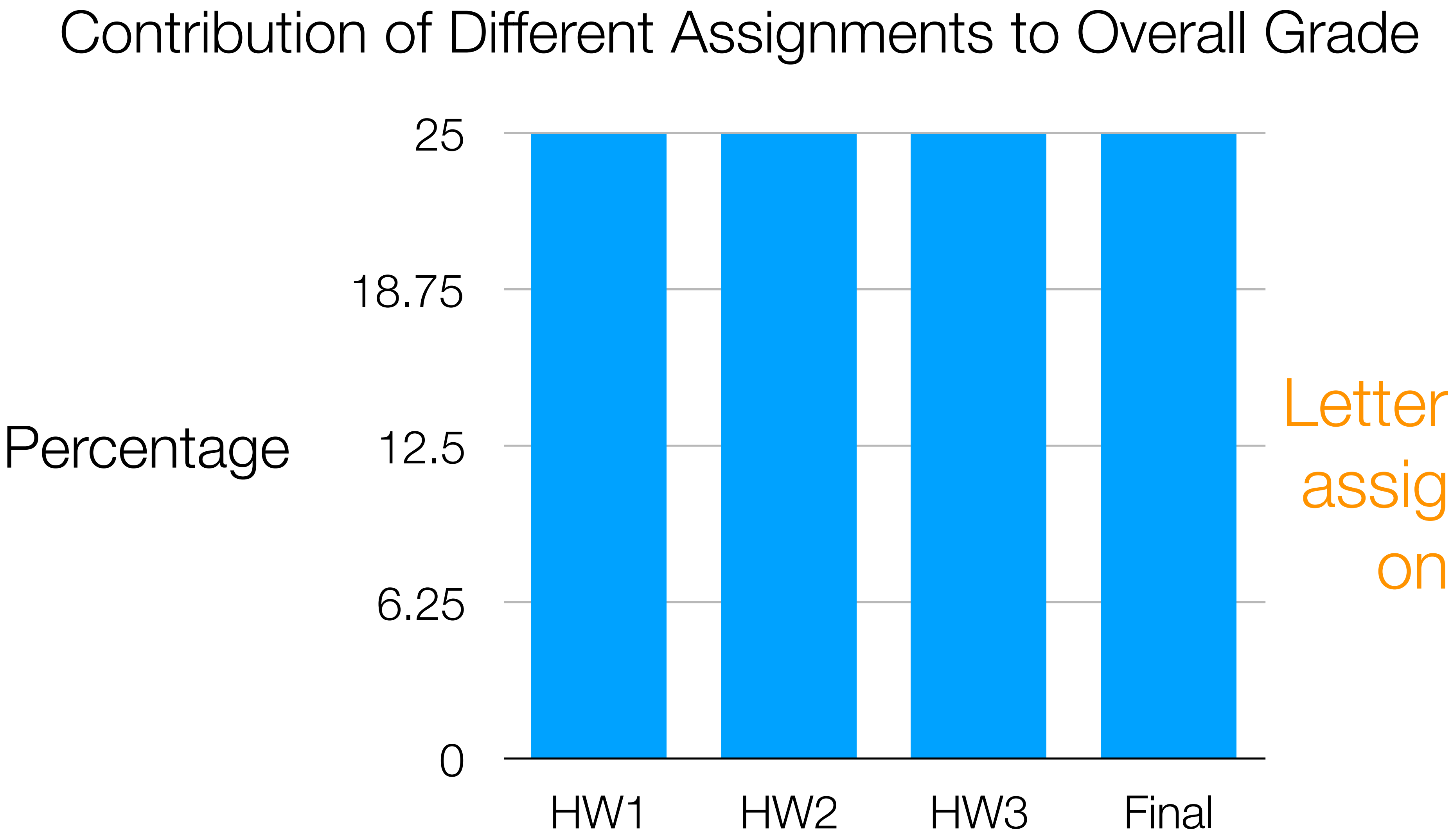
Course Goals

By the end of this course, you should have:

- Lots of hands-on 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 problems faced by organizations

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

Deliverables & Grading



Letter grades are assigned based on a curve

Assignments will involve coding in Python
(we will use popular packages such as scikit-learn **and** tensorflow**)**

Some problems will require cloud computing
(we will use Amazon AWS)

Programming and Cloud Computing

- The data science/machine learning tools available have changed *drastically* over the last few years
    - Working with most of the latest innovations requires some programming (Python is common)
- Datasets encountered by many organizations are now often *massive*
    - Datasets often either won't fit or won't be processed fast enough on your personal machine but renting compute resources is now cheap (e.g., Amazon AWS, Google Compute)

Course Prerequisites

What you should already have:

- Python coding experience (if you don’t know Python we’ll assume you can pick it up rapidly on your own)
    - Review session: Thursday Oct 26, 12pm-1:30pm, HBH 1007
- Ability to follow basic math derivations (largely similar to calculations with tables in Google Spreadsheet/ Excel)
    - I am for the most part *not* going to go into derivations for algorithms encountered
    - However, I will be going over what structure algorithms *assume* to be in data

Course ~~Textbook~~ Materials

No existing textbook matches the course… =(

Main source of material: lectures slides

We'll post complimentary reading as we progress

Everything can be found on:

canvas piazza

(Piazza link is within Canvas)

Heads up: Within Piazza you get to interact with students from both Pittsburgh and Adelaide!

Computing Environment

- We will be using **Anaconda (Python 3.6 version)** https://www.anaconda.com/what-is-anaconda/
- We will give instructions for any third party packages to install and how to set up **Amazon AWS** for cloud compute
- You will be submitting assignments in the form of **Jupyter notebooks**

Final Exam

Time and place (according to current schedule):

- Pittsburgh students:
  Friday December 15, 2017, 1pm HBH 1202
- Adelaide students:
  Friday December 8, 2017, time location TBD

Format (tentative):

- In-class final **where you have to bring a laptop computer and produce a Jupyter notebook** that answers a series of questions
- No collaboration (obviously)

Course Policies

- Please do not use cell phones and laptops during class

- All homework submissions are online in Canvas
(you submit your Jupyter notebook and any accompanying files, all zipped up) — late homework will not be accepted

- Homework should reflect your individual understanding and the code you submit should be code you wrote yourself

Collaboration & Academic Integrity

- 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 and how to approach problems
  - Please acknowledge classmates you talked to or resources you consulted (e.g., stackoverflow)
  - **Do not share your code with classmates**

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

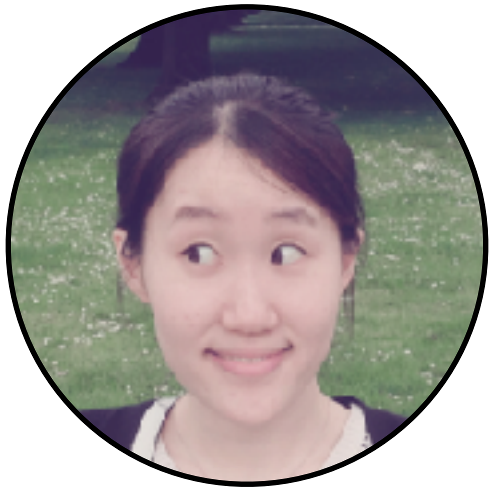
Course Staff

Email course staff: uda-course-f17@lists.andrew.cmu.edu

Emaad
Manzoor

Rashmi
Raghunandan

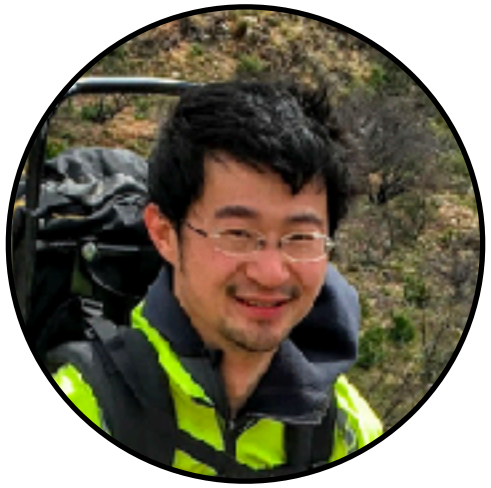
Runshan
Fu



Yoonjung
Kim

TA's

Instructor



George Chen

Office hours:
Check Canvas

You are a beta tester

This is the first offering of this course!

Please report bugs!

I will be soliciting feedback regularly.

We will adjust homework difficulty to try to make the workload reasonable.