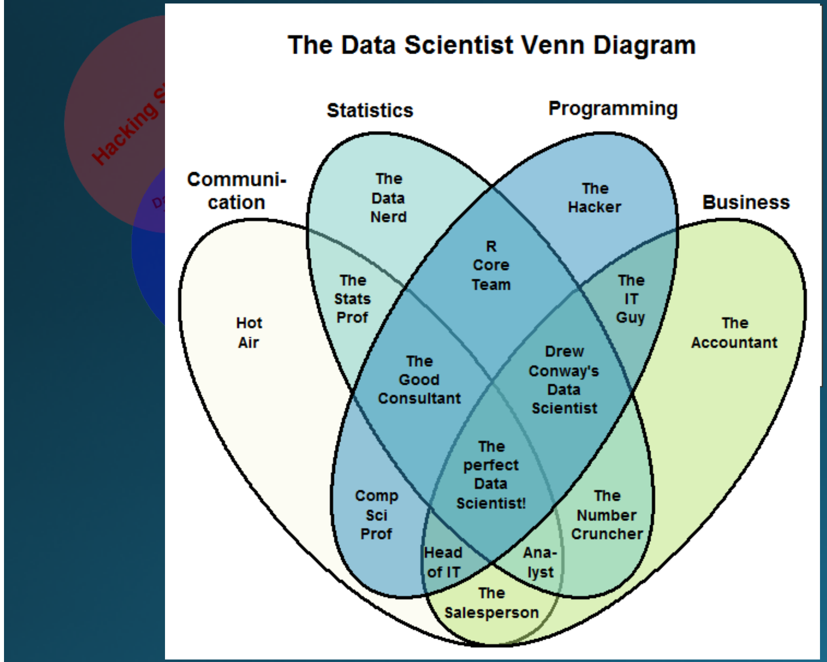
Fall 2019

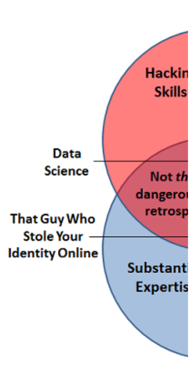
Dr. Shannon Quinn

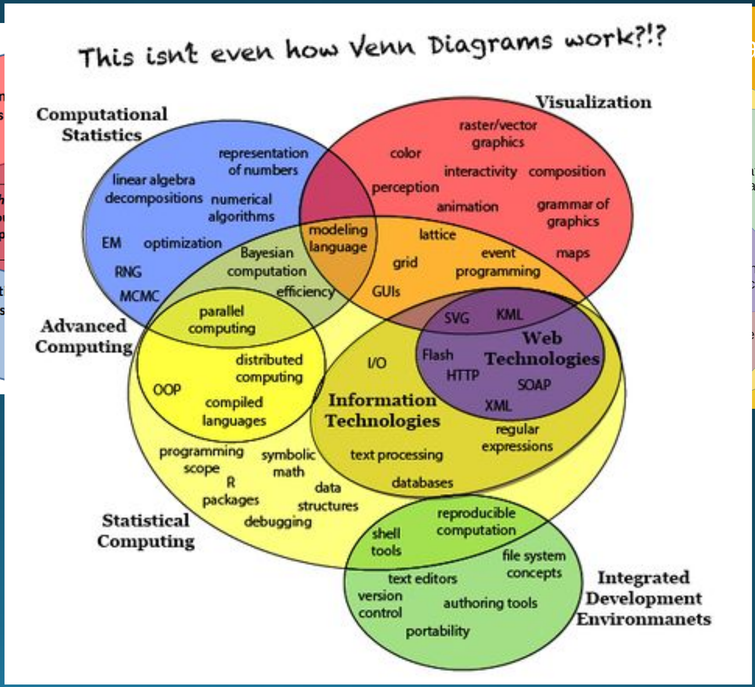
Course Introduction

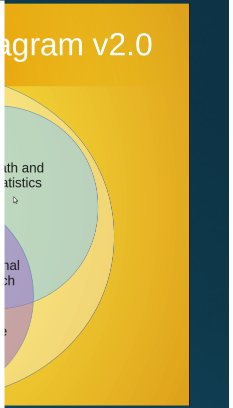
What is "data science"?

- It's a field singularly devoted to bringing back the Venn Diagram









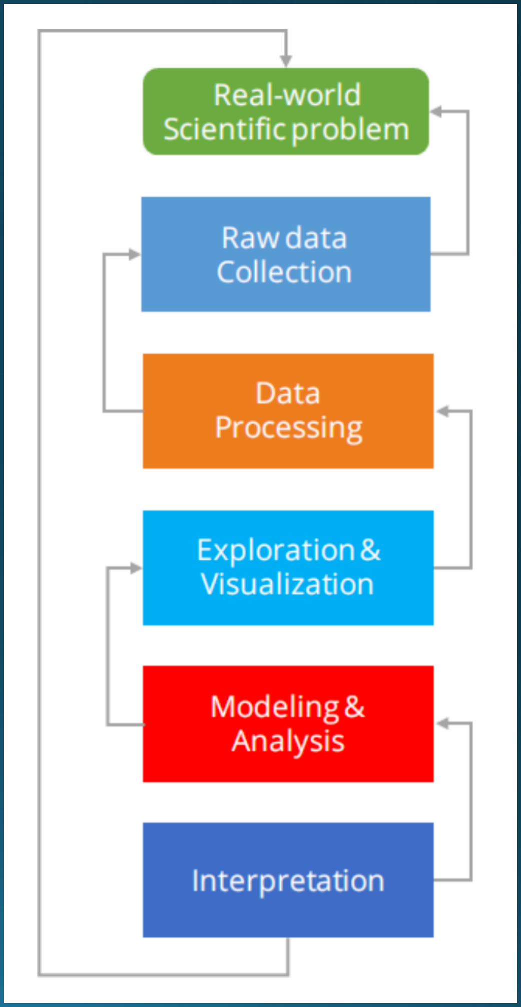
What is “data science”?

- From Wikipedia (emphasis mine):

Data science, also known as data-driven science, is an **interdisciplinary field** about scientific methods, processes, and systems to **extract knowledge or insights from data in various forms, either structured or unstructured**, similar to data mining. Data science is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with data. **It employs techniques and theories drawn from many fields** within the broad areas of mathematics, statistics, information science, and computer science, in particular from the subdomains of machine learning, classification, cluster analysis, data mining, databases, and visualization.

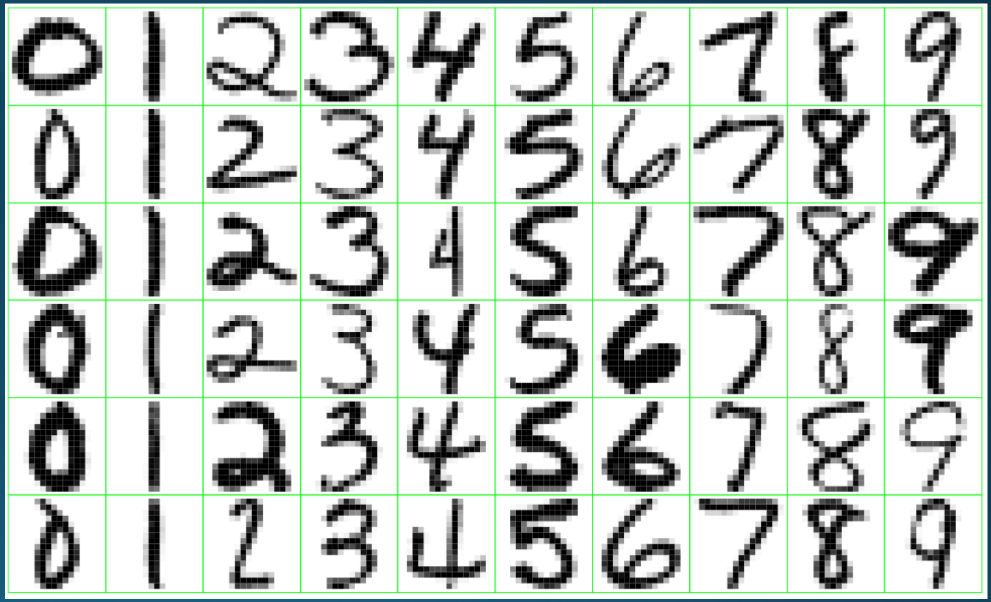
What is "data science"?

- If you want my opinion: Dr. Lee nailed it in CSCI 3360
- Data Science encompasses **the entire problem stack**
  - Problem definition
  - Data collection & cleaning
  - Exploration
  - Modeling
  - Interpretation & insights

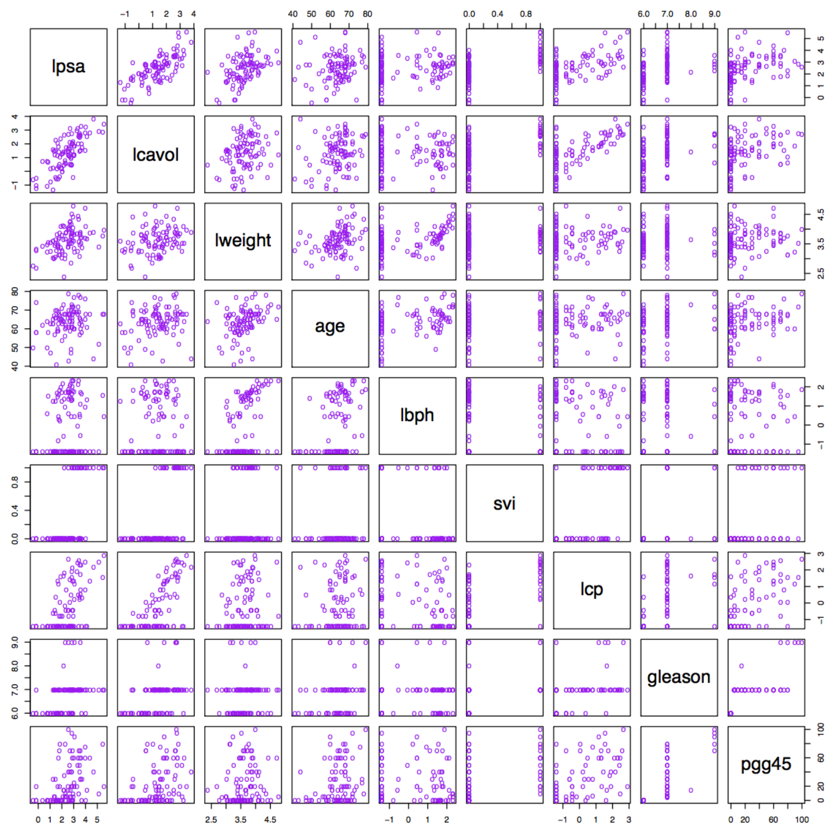


Data Science in practice

- Can we automatically sort mail based on ZIP code?



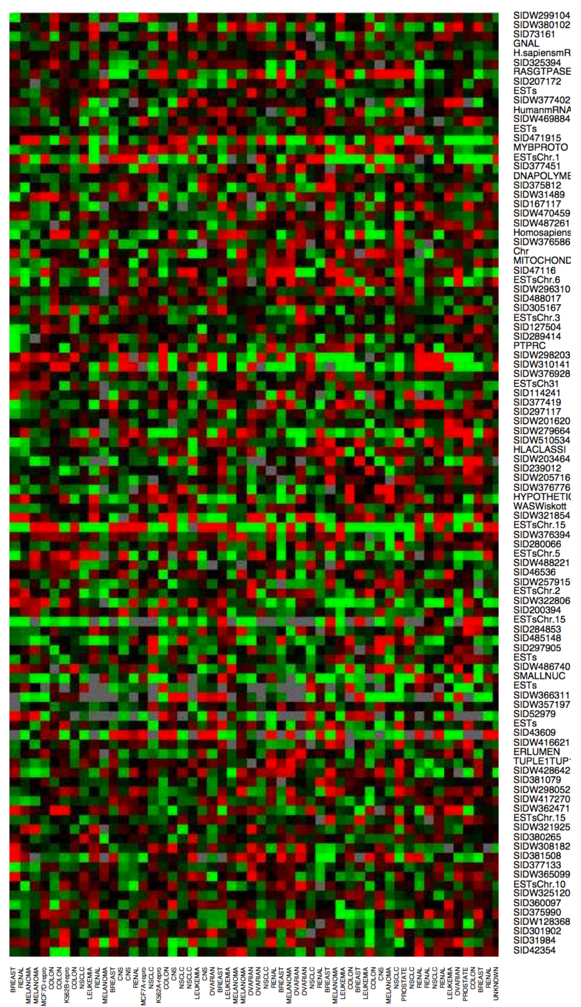
Data Science in practice



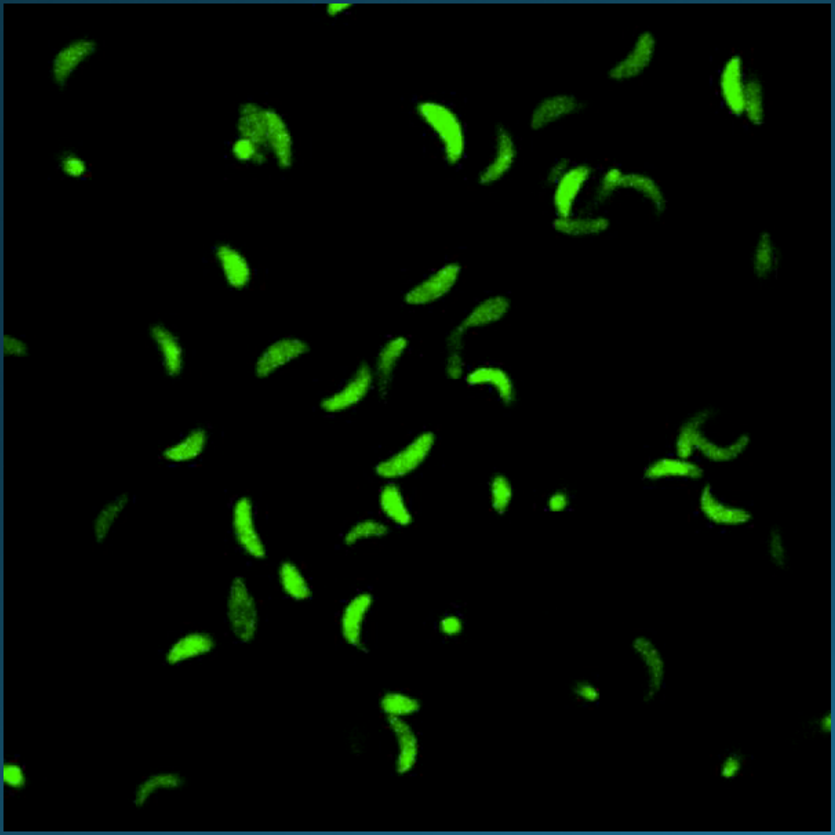
- What features of prostate cancer are indicative of production of specific antigens?

Data Science in practice

- Which genes are overactive or underactive in cancer patients?



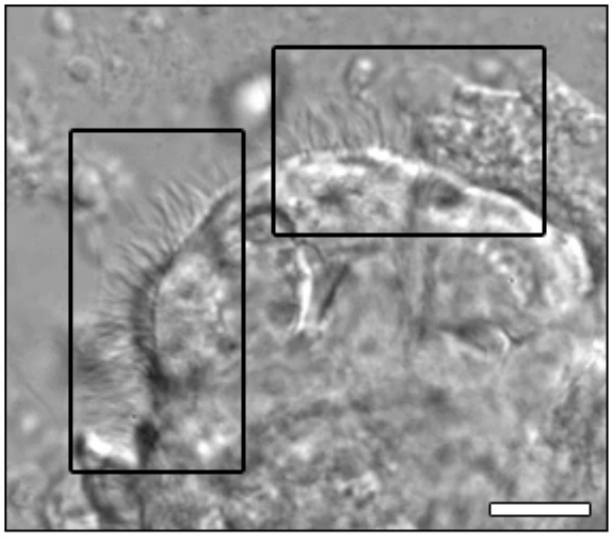
Data Science in practice

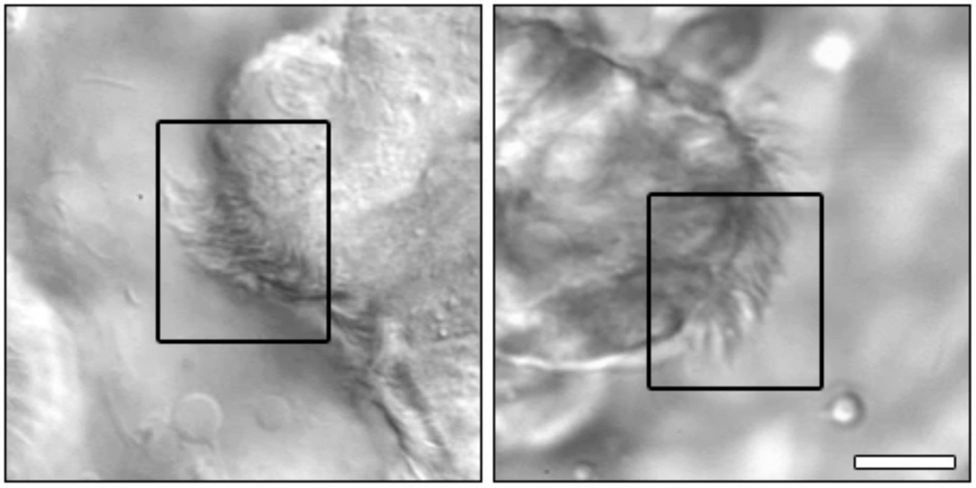


- What factors determine the movement of the *Toxoplasma gondii* parasite?

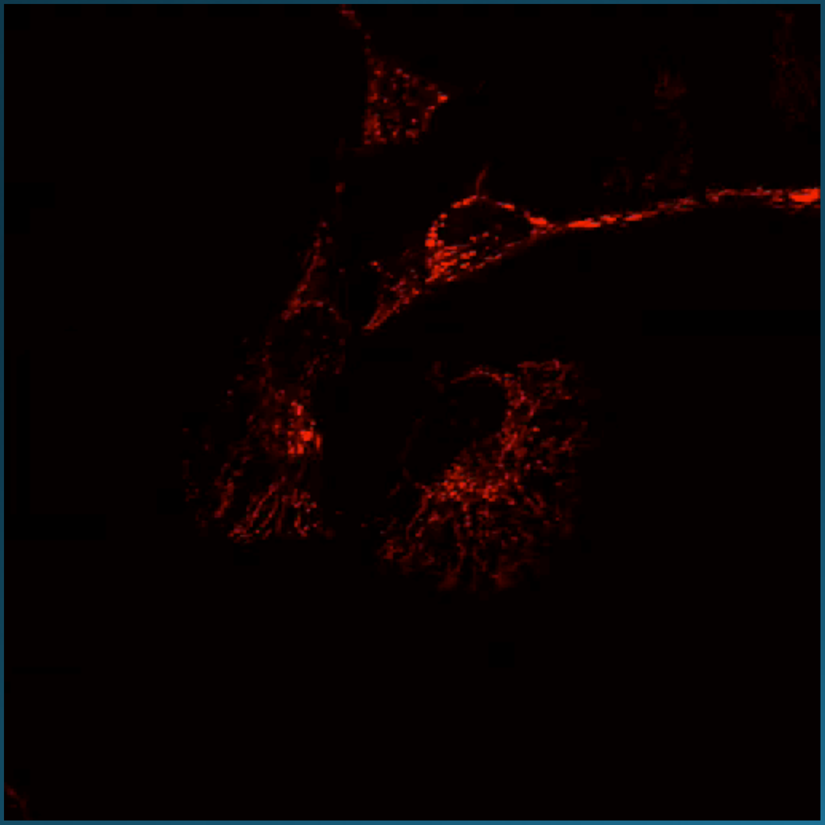
Data Science in practice

- How is the motion of cilia associated with and indicative of specific pathologies?





Data Science in practice



- What are the protein patterns of mitochondria under different conditions, and how do these changes take place?

What this course *is*

- Picks up where 3360 Data Science I left off
- A deeper dive into modeling and quantitative analysis methods
- “What to do when you’ve already tried Option A”
- Combination of theory and practice using latest data science tools and techniques

What this course is *not*

- Introduction to probability and statistics
  - Should be able to derive Bayes' Theorem from law of conditional probability, no sweat
  - Don't need to prove the SVM dual, but should be aware of it and its function
- Introduction to programming
  - No Python experience required, but are expected to pick it up **FAST** (i.e., you've programmed before, just not in Python)
  - **Next week** is a Python crash-course

What?

- Course title: **Data Science II**
    - CSCI 4360 (for undergraduates)
    - CSCI 6360 (for graduates)
- Course textbooks: **none required**
    - Lots of recommended books—check out the course website
    - Will continue to update with more references
- Python: not required but *strongly recommended*
    - **3.x** (2.x is hitting EOL in a few weeks)
    - Details to come

Who?

- Dr. Shannon Quinn (that’s me)
  - 2008: B.S. in Computer Science from Georgia Tech (go Jackets!)
  - 2010: M.S. in Computational Biology from Carnegie Mellon
  - 2014: Ph.D. in Computational Biology from joint Carnegie Mellon-University of Pittsburgh Ph.D. Program in Computational Biology (CPCB)
- Research areas
  - Biomedical imaging
  - Representation learning
  - Computer vision
  - Distributed computing
  - Biosurveillance
  - **Data Science + Public Health**

When and Where?

- Lectures
  - Tuesdays and Thursdays: 2:00 – 3:15pm, Geography/Geology 200C
  - Mondays: 2:30 – 3:20pm, Chemistry 453
- Office Hours
  - Boyd 638A
  - TBA (things are a bit crazy right now)
- TA: TBA!
  - Office Hours: **<TBA>** (seeing a pattern? let's train a classifier!)

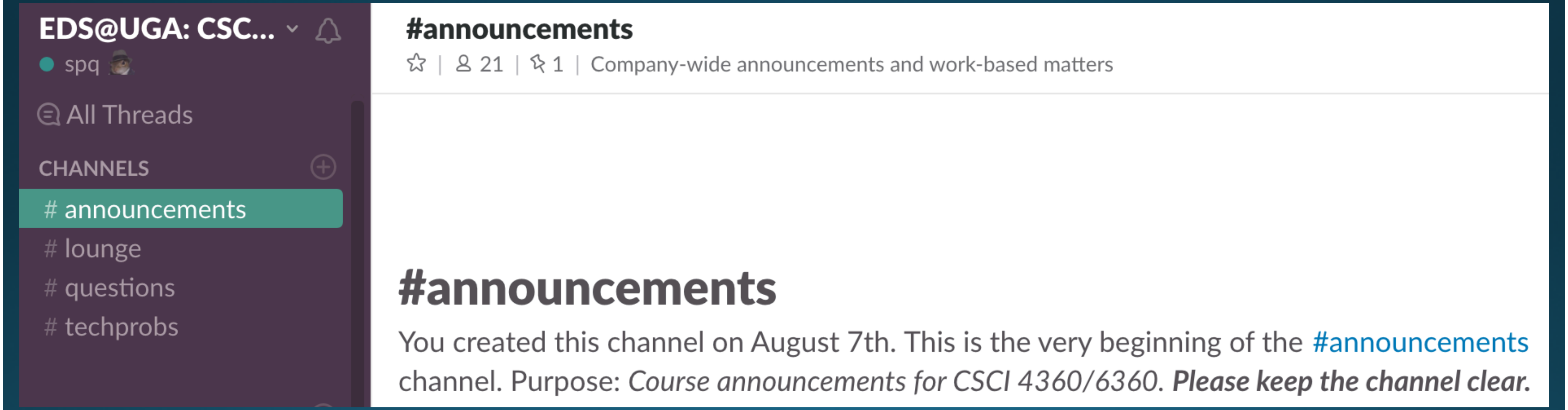
How?

- Course website / syllabus: https://eds-uga.github.io/csci4360-fa19

Welcome to Data Science II

How?

- Slack chat: https://eds-uga-csci4360.slack.com/



#announcements

Company-wide announcements and work-based matters

#announcements

You created this channel on August 7th. This is the very beginning of the #announcements channel. Purpose: *Course announcements for CSCI 4360/6360.* ***Please keep the channel clear.***

How?

- Slack: post in **`#questions`**
- Email: squinn@cs.uga.edu
  - I get **tons** of emails every day
  - Slack will not only likely get a faster response, but **your fellow students could probably also answer even more quickly**

Honesty at UGA

- I'd like to think I don't have to justify this
  - There's an official UGA Honesty Policy https://ovpi.uga.edu/academic-honesty/academic-honesty-policy
  - I have automated code-checkers (thank you, data science)
  - It's **way easier than you think** to spot copied code
- The official policy in this class:

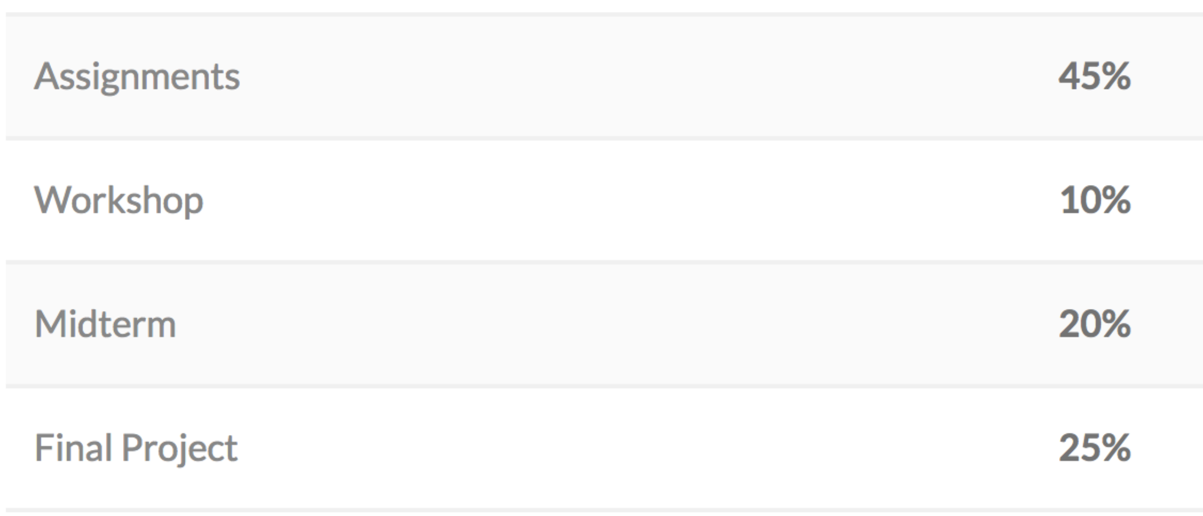
Discuss ideas and concepts with your classmates (or anyone!).

Write the code yourself (unless you're on a team).

Attendance Policy

- When *I'm* speaking—
  - Meh. You're all adults
- When *anyone other than me* is speaking—
  - **Strict attendance requirement**
  - Guest lecturers (will be at least 2 this semester), classmates (workshops, final projects)
- If you need to miss class, that's fine!
  - Don't need to ask my permission, either
  - **But you do need to tell me you won't be attending, or couldn't attend**

Grading Breakdown



Assignments	45%
Workshop	10%
Midterm	20%
Final Project	25%

Assignments

- There will be **5**
  - So yes, each is worth **9%** of your grade
- Each will be **two weeks long**
  - Released on a Tuesday morning, due two weeks later by **11:59:59pm**
- Will likely entail a written and a programming portion
  - Coding in Python
  - Writing in Word or LaTeX—**nothing handwritten!**
- Details to come about how to submit the assignments

Workshops

- Most Mondays, we'll have a **workshop**
- This is **student-led and organized** (you have to do **one!**)
- The objective of each workshop is to **demo** a proof-of-concept for your student colleagues
- This can be
  - implementing a topic we covered in class
  - demonstrating how to use a tool that would help with the topics we're covering
  - some other neat course-related use-case
- **Recommended topics are on the course website!**

Midterm exam

- It’s an exam that happens near the mid-term **(Oct 3)**, what more do you want?
- (details will be released later)

Final Project

- **Teams** (of 3-4 students, ideally) will work on a specific data science question
- Three components:
  - The **proposal**, which outlines the team you'll work with, the question you'll address, and the methods + tools you'll use to address it
  - The **presentation**, where you talk about how awesome your problem is and how you and your team killed it dead (or have almost done so)
  - The **deliverables**, consisting of 1) the *complete* code, and 2) a 6-10 page NIPS-style write-up of your project
- More details to come!

”Assignment 0”

- Email me with your preferred email for me to send you a Slack invite
- Put together groups of 2-3 students and **pick a date + topic for a workshop** (see the course website for available dates)
  - Try to tie in with the lecture material surrounding that date, if possible
  - First [student-run] workshop is **Monday, Aug 26!**
  - **Sign-up link will be posted in the Slack chat**
- Assignment 1 comes out **Tuesday**, so do this ASAP!

Next week

- Workshop 0, on using the Anaconda distribution for installing and configuring your own Python environment! (far and away the easiest way to get up and running with Python)

Questions?