Fall 2021

Dr. Shannon Quinn

Course Introduction

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)
  - Workshop 0 (Monday) 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: **3.x**
  - Details to come

Who?

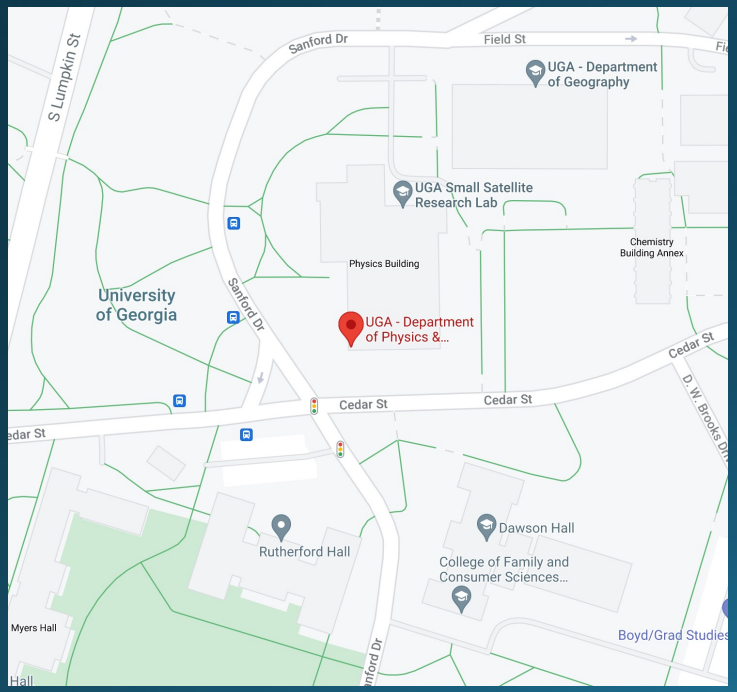
- Dr. Shannon Quinn (that's me)
  - 2008: B.S. in Computer Science from Georgia Tech (go Jackets!)
  - 2010: M.S. in Computational Biology from Carnegie Mellon
  - 2014: Ph.D. in Computational Biology from joint Carnegie Mellon-University of Pittsburgh Ph.D. Program in Computational Biology (CPCB)
-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

When and Where?

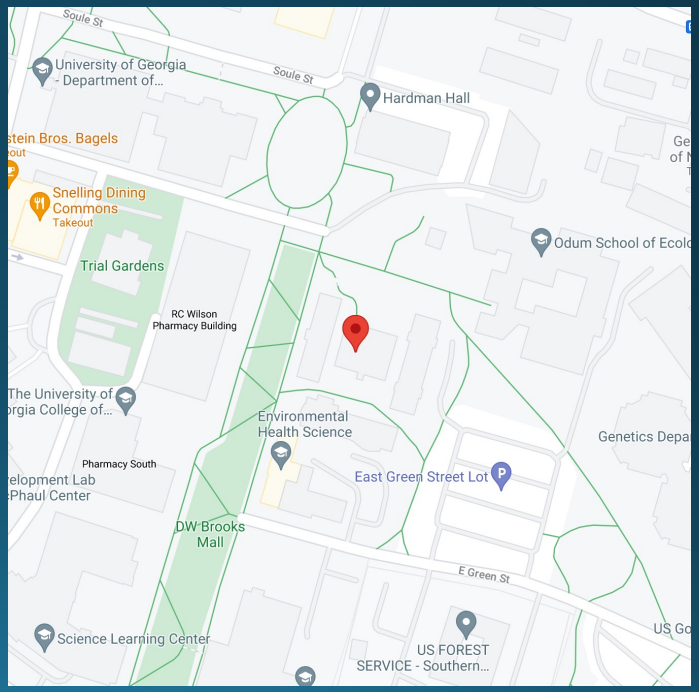
- Lectures
  - Tuesdays and Thursdays: 2:20 – 3:35pm, Physics 0303
  - Mondays: 3:00 – 3:50pm, Forest Resources-4 0517
- Office Hours
  - Boyd 638A
  - TBA / Discord
- TA: TBA
  - Office Hours: TBA

When and Where?

Tuesday / Thursday



Monday

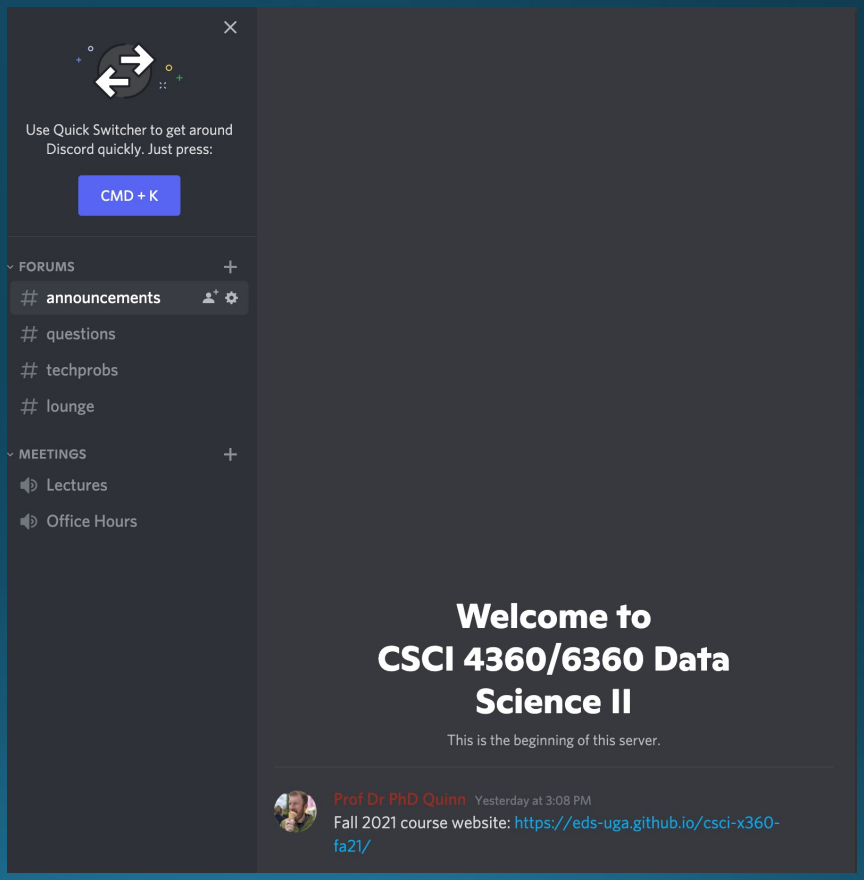


How?

- Course website / syllabus: https://eds-uga.github.io/csci-x360-fa21/

How?

- Discord


Use Quick Switcher to get around Discord quickly. Just press:
CMD + K
FORUMS
announcements
questions
techprobs
lounge
MEETINGS
Lectures
Office Hours
Welcome to CSCI 4360/6360 Data Science II
This is the beginning of this server.
Prof Dr PhD Quinn Yesterday at 3:08 PM
Fall 2021 course website: https://eds-uga.github.io/csci-x360-fa21/

How?

- Discord: post in **`#questions`**
  - If you're having an issue with course infrastructure, use **`#techprobs`**
- Email: squinn@cs.uga.edu
  - I get **tons** of emails every day
  - Discord will not only likely get a faster response, but **the TA and 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 use code-checkers
  - 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).

Grading Breakdown

Assignments	**45%**
Workshop	**15%**
Midterm	**40%**
Final Project	**40%**

Assignments

- There will be **5**; you choose and complete **3**
  - Each will be worth **15%** of your grade
  - Doing more than 3 is extra credit
- Each will be **2.5 weeks long**
  - Released on a Tuesday morning, due on Thursday evenings by **11:59:59pm**
- Will likely entail a written and a programming portion
  - Coding in Python
  - Writing in Word or LaTeX—**nothing handwritten!**
- All deliverables will be submitted through **?????**
  - More details to come!

Workshops

- Most Mondays, we'll have a **workshop**
- This is **student-led and organized**
  - If you're in 4360, you have to do **one**
  - If you're in 6360, you have to do **two**
- The objective of each workshop is to **demo** a proof-of-concept
- 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 linked on the course website!**

Midterm exam

- It's an exam that happens near the mid-term **(Oct 12)**, what more do you want?
- (details will be released later)

Final Project

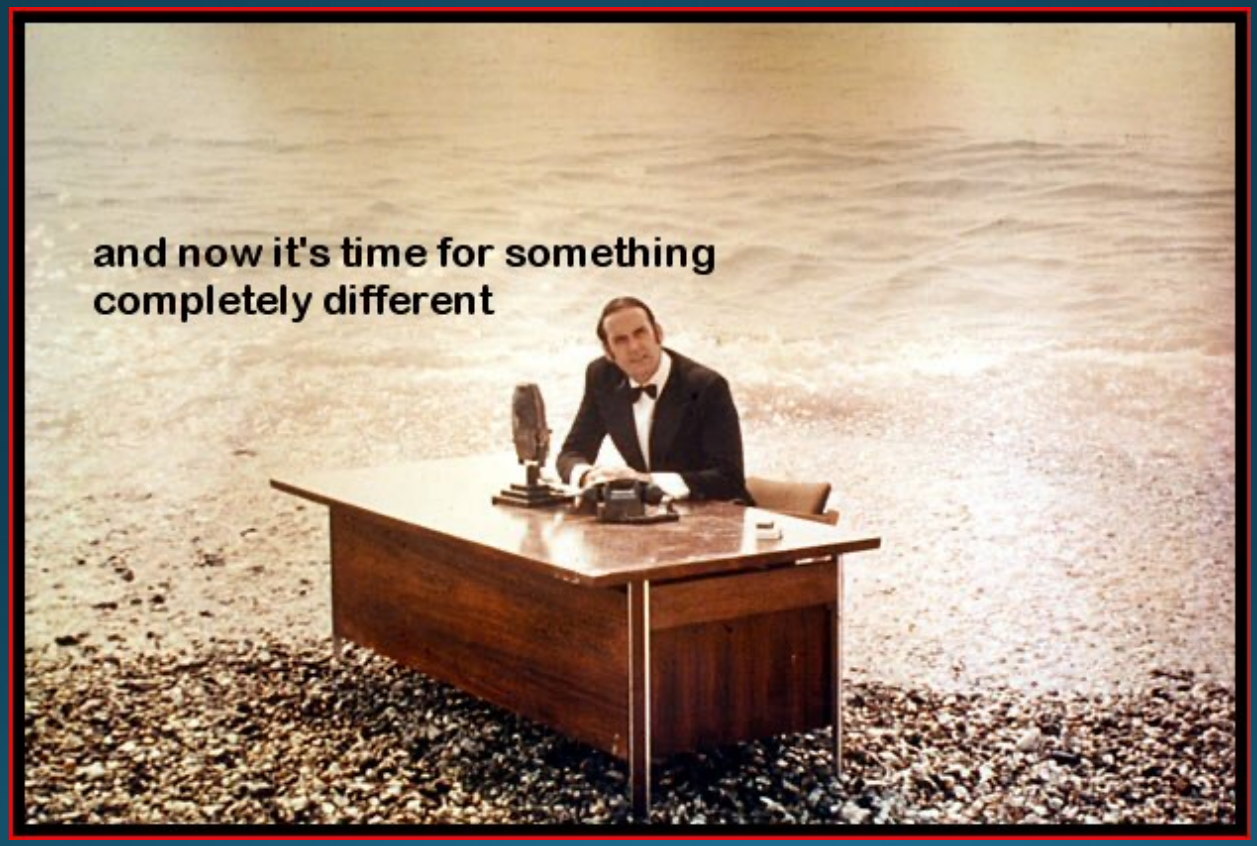
- **Teams** (of 2-3 students, ideally) will work on a specific data science question
- Three components:
  - The **proposal**, which outlines the team you'll work with, the question you'll address, and the methods + tools you'll use to address it
  - Project **updates**, at least two of them a few weeks apart, which formally lay out your progress, any roadblocks and workarounds, and changes to your original goals
  - 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)
- More details to come!

Midterm or Final

- Choose **one**
  - Midterm exam (Oct 12)
  - Final project
- Doing both is extra credit

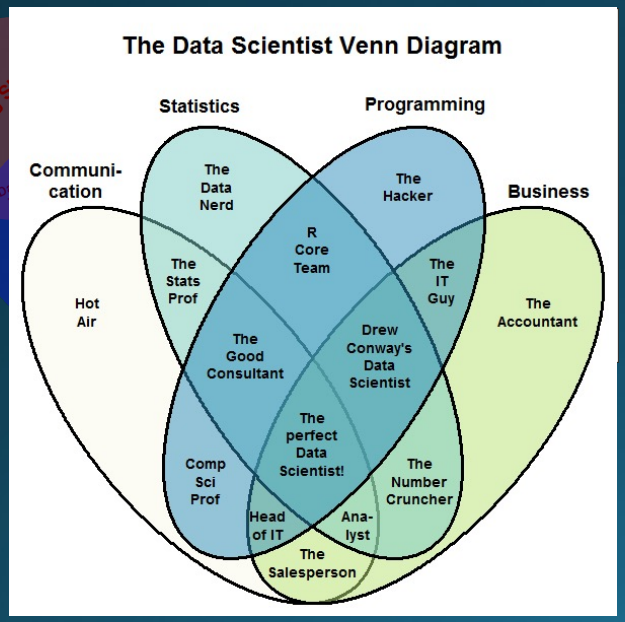
"Homework 0"

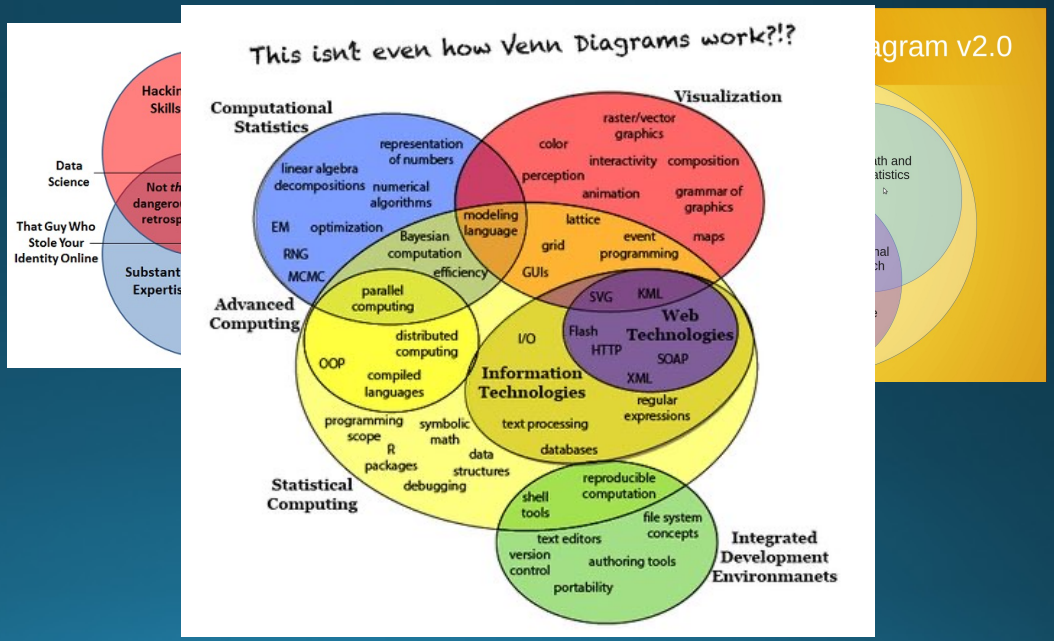
- Accept the invitation to Discord if you haven't already
  - … o_o
- Put together groups of 2-3 students and **pick a date + topic for a workshop** (see the course website for available dates)
  - Try to tie in with the lecture material surrounding that date, if possible
  - First [student-run] workshop is **Monday, Aug 30!**
  - **Sign up here:** https://docs.google.com/spreadsheets/d/1S9fuFx47iEiB5zogrZURiRu7aBYdiB3bGUPmob25zyw/edit?usp=sharing
- Homework 1 comes out next **Tuesday!**


and now it's time for something
completely different

What is "data science"?

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





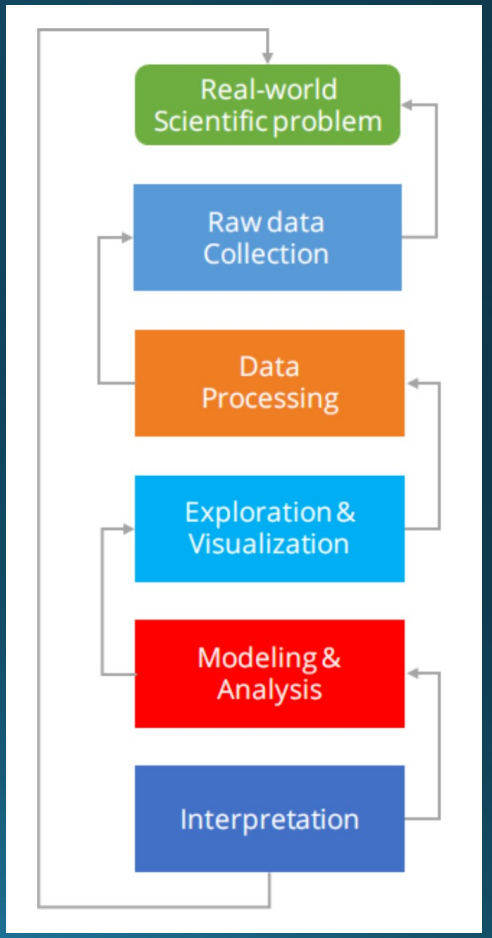
What is "data science"?

- 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.

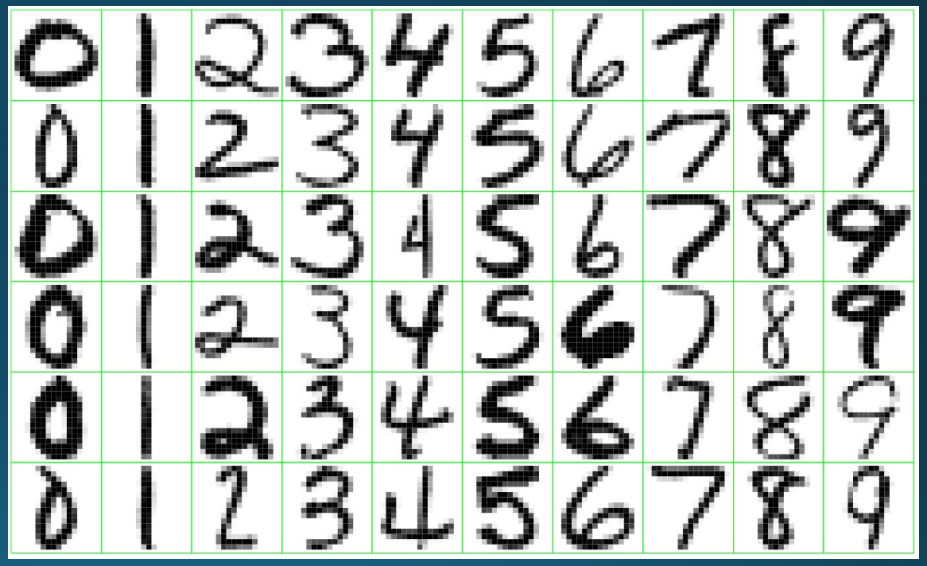
What is "data science"?

- 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

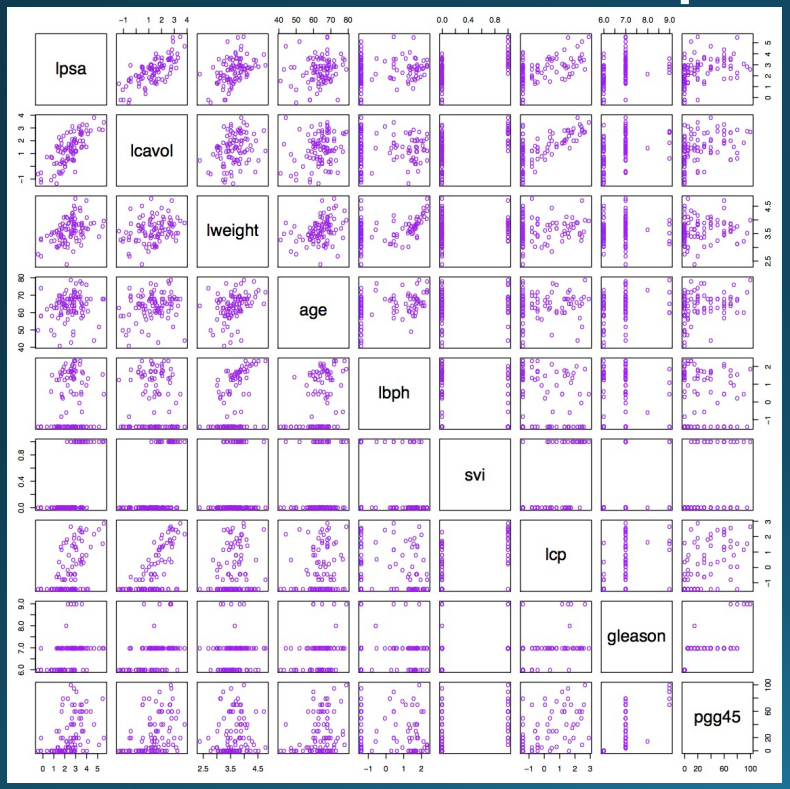


Data Science in practice

- Can we automatically sort mail based on ZIP code?



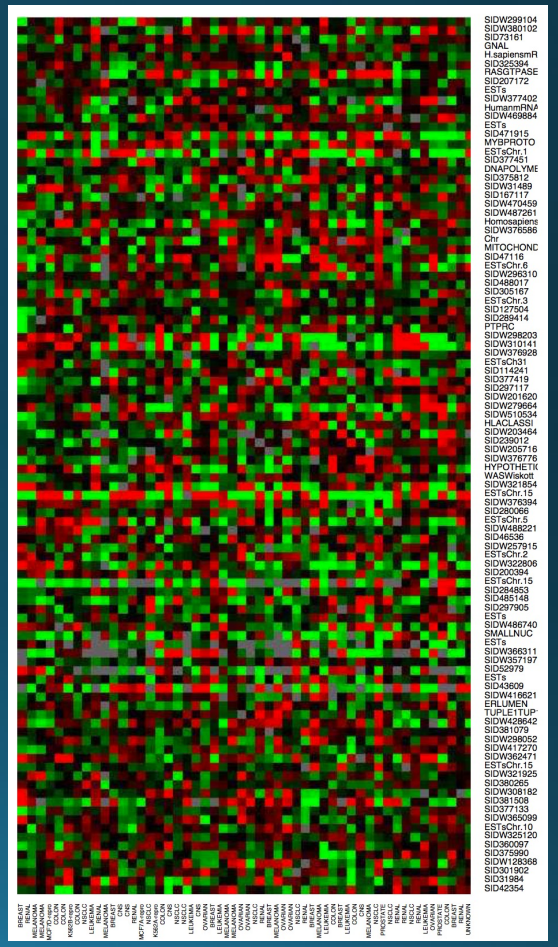
Data Science in practice



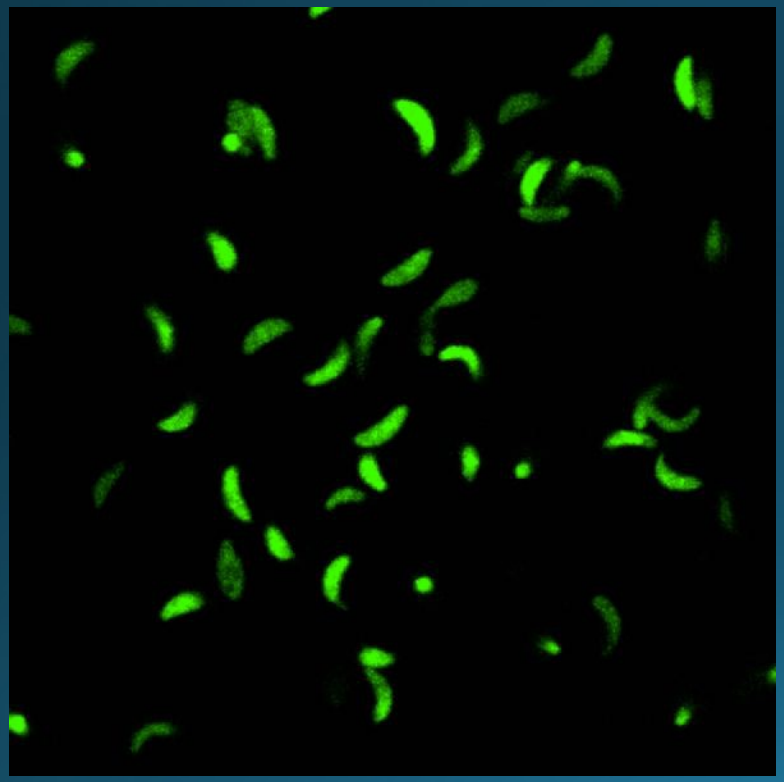
- What features of prostate cancer are indicative of production of specific antigens?

Data Science in practice

- Which genes are overactive or underactive in cancer patients?



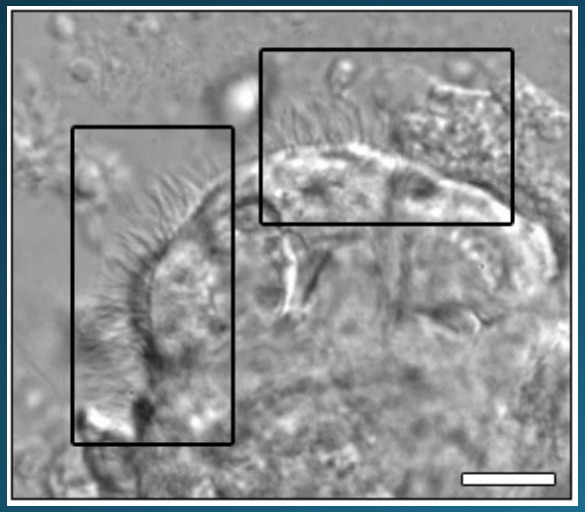
Data Science in practice

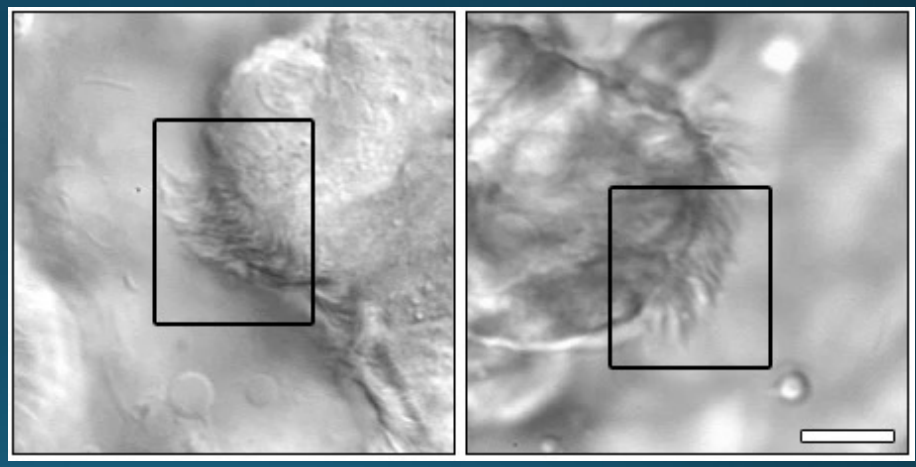


- What factors determine the movement of the *Toxoplasma gondii* parasite?

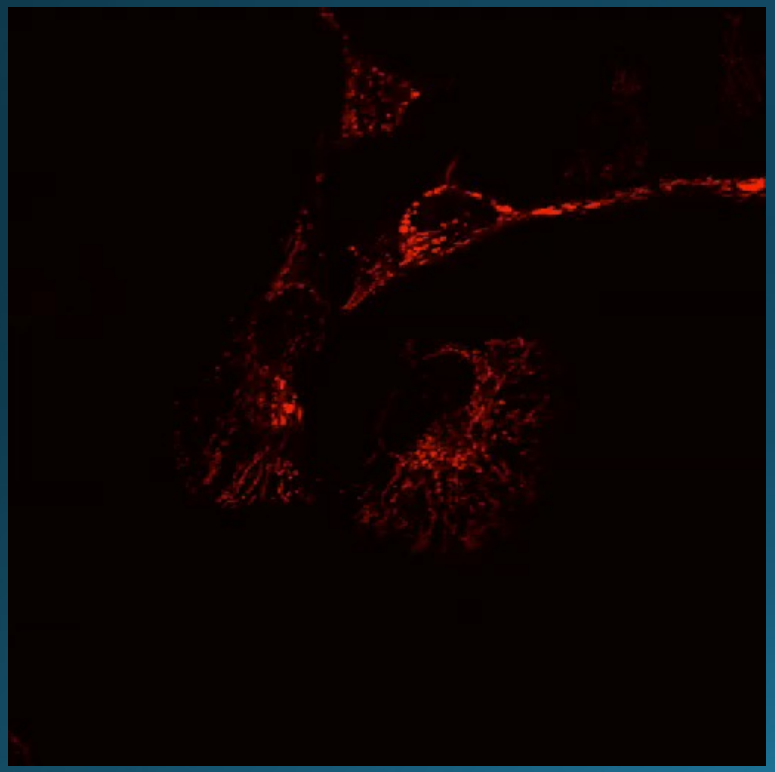
Data Science in practice

- 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?

Questions?

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)
- Python crash course