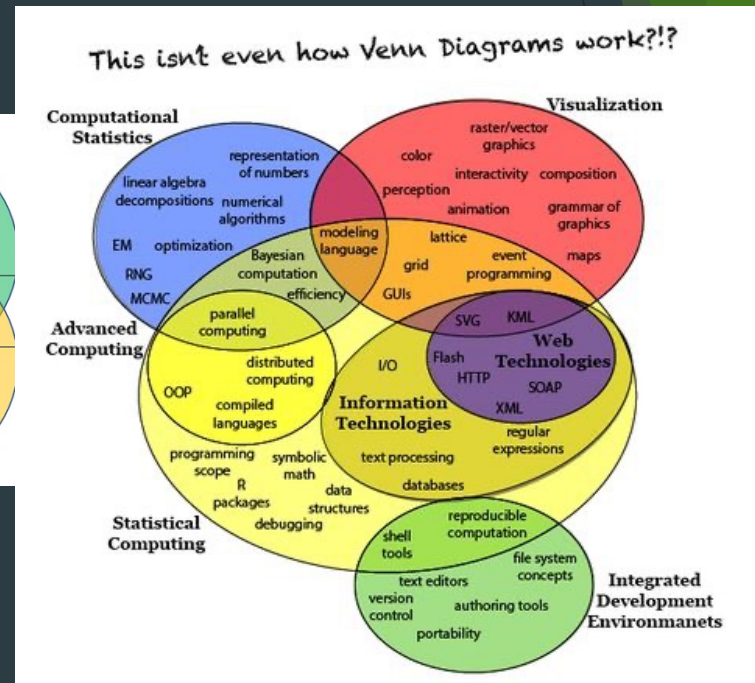
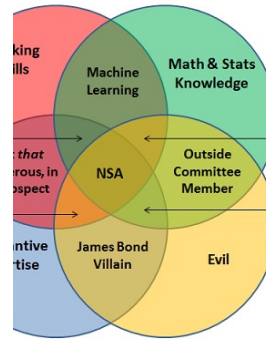
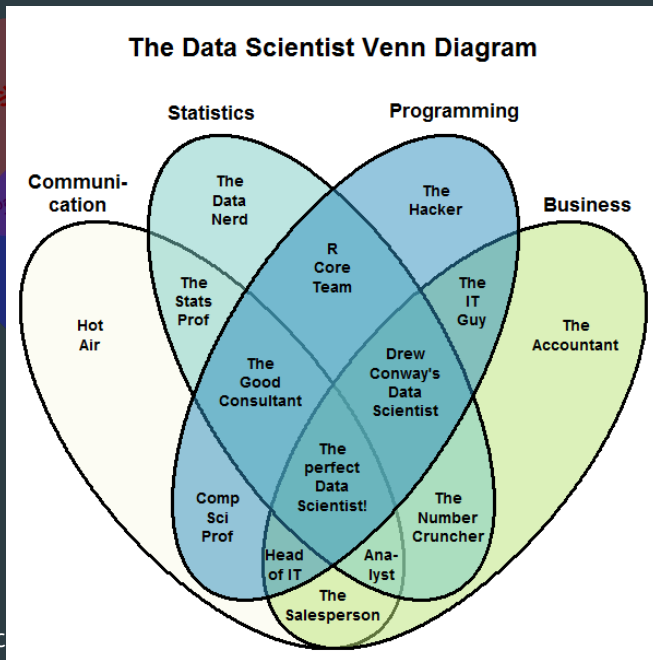


Data Science II: Course Introduction

Spring 2025
Prof. Shannon Quinn

What is “data science”?

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



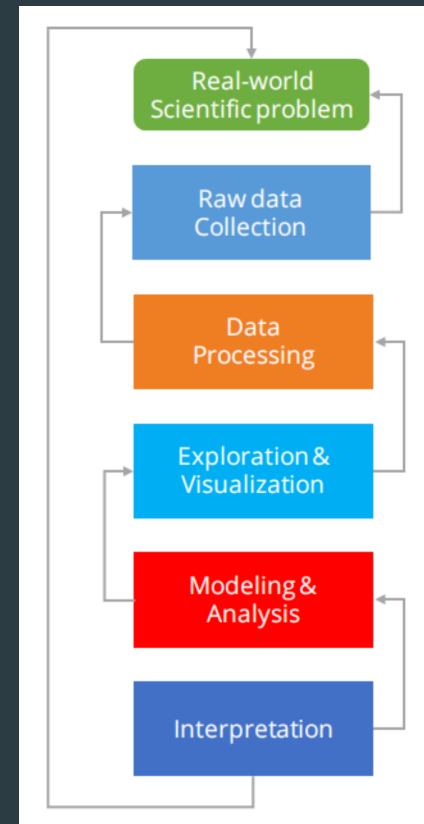
What is “data science”?

- ▶ 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.

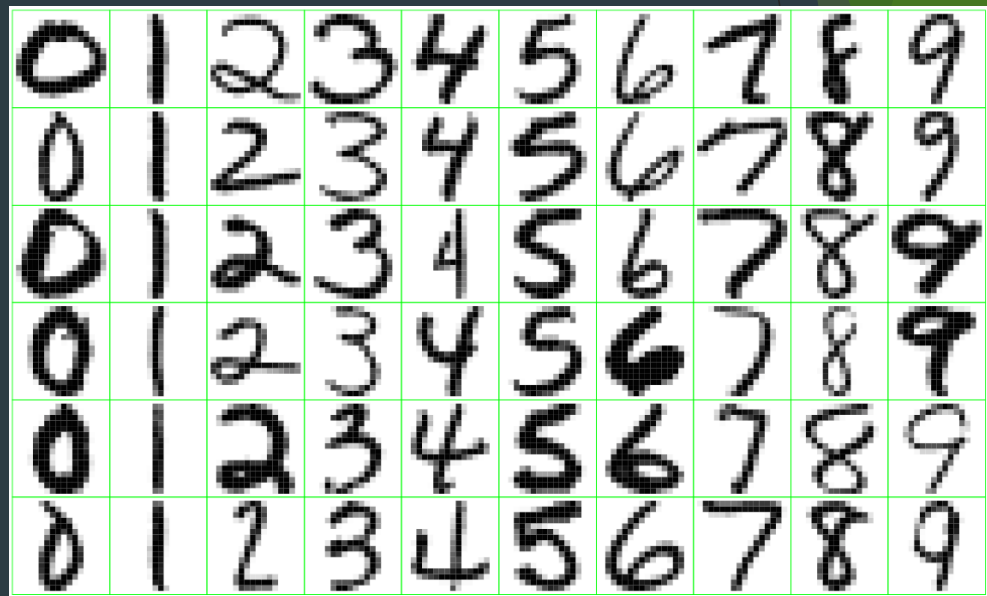
What is “data science”?

- ▶ 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

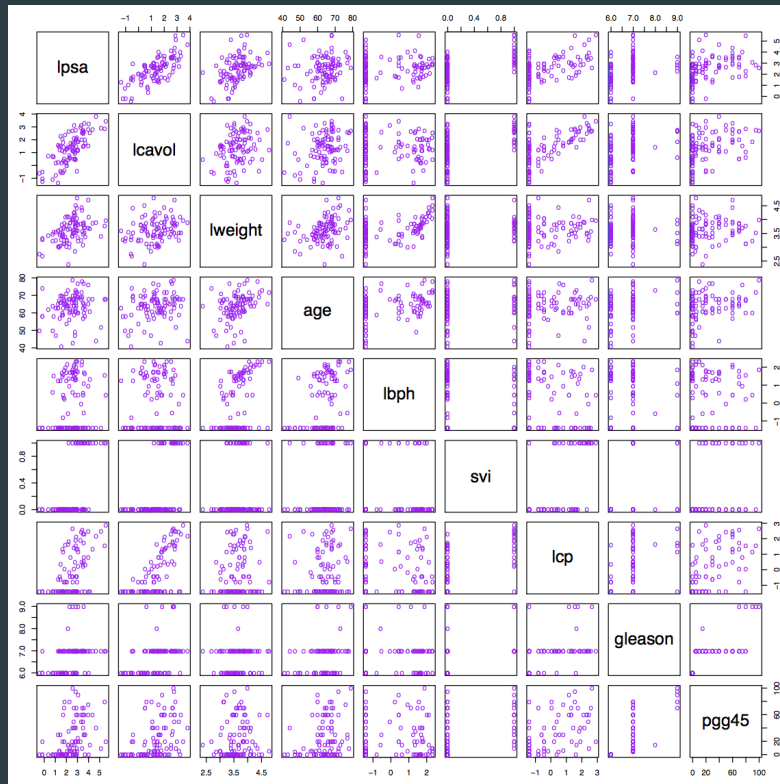


Data Science in practice

- ▶ Can we automatically sort mail based on ZIP code?

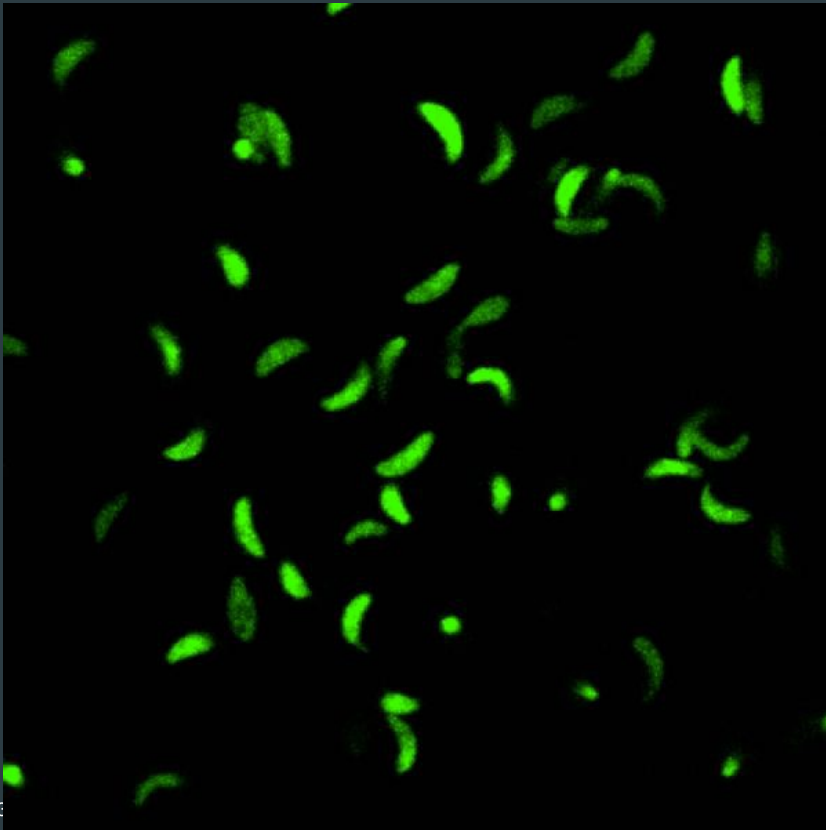


Data Science in practice



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

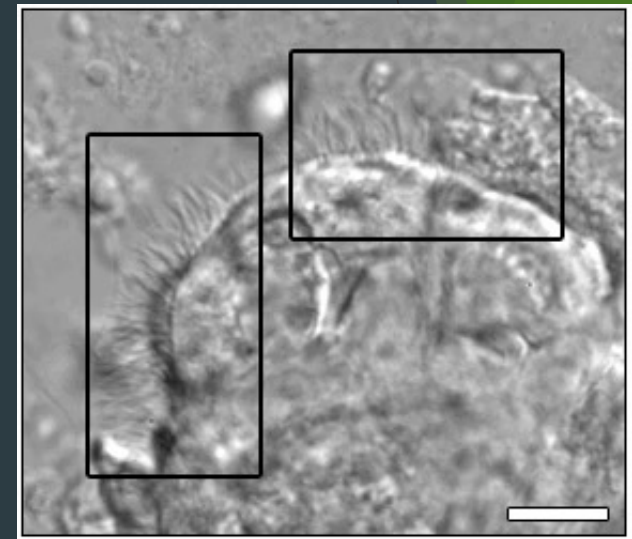
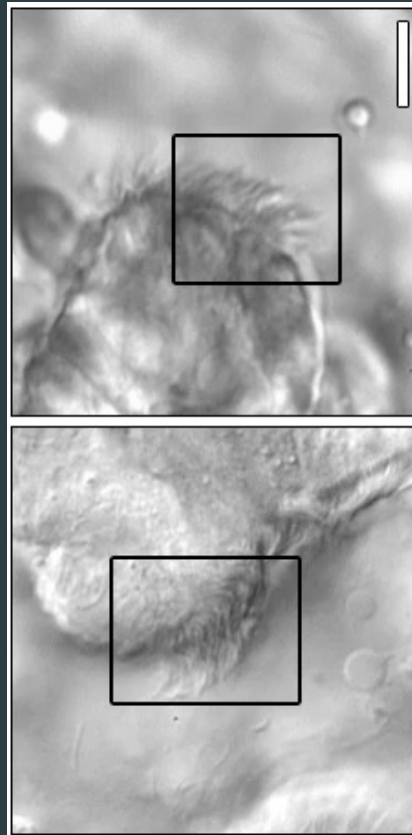
Data Science in practice



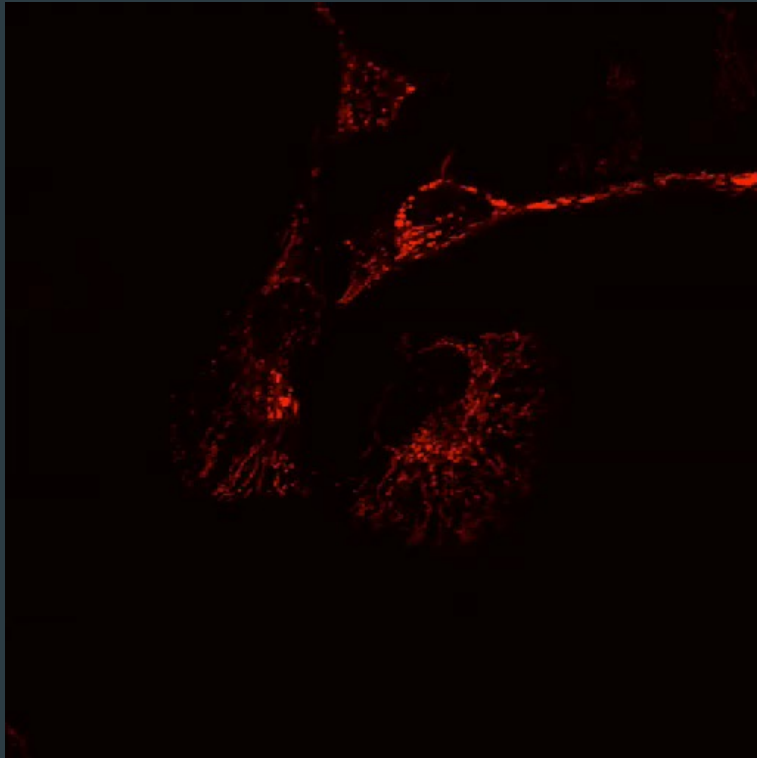
- ▶ What factors determine the movement of the *Toxoplasma gondii* parasite?

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?

Data Science in practice

- ▶ Self-driving cars
- ▶ Chatbots / Image generators
- ▶ Product recommendations
- ▶ Biometrics / facial recognition
- ▶ Precision agriculture
- ▶ Social media feeds
- ▶ Graphics upscaling
- ▶ Security camera intruder detection
- ▶ [[your application here]]

Data Science in practice

- ▶ **It's multi-disciplinary**
 - ▶ At *most* one foot in Computer Science
- ▶ **It's collaborative**
 - ▶ Gone are the halcyon days of the “lone genius”
 - ▶ Teams are the norm
 - ▶ Communication is more important than ever
- ▶ **It's humbling**
 - ▶ There's always something new to learn
 - ▶ There's always someone with more expertise

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

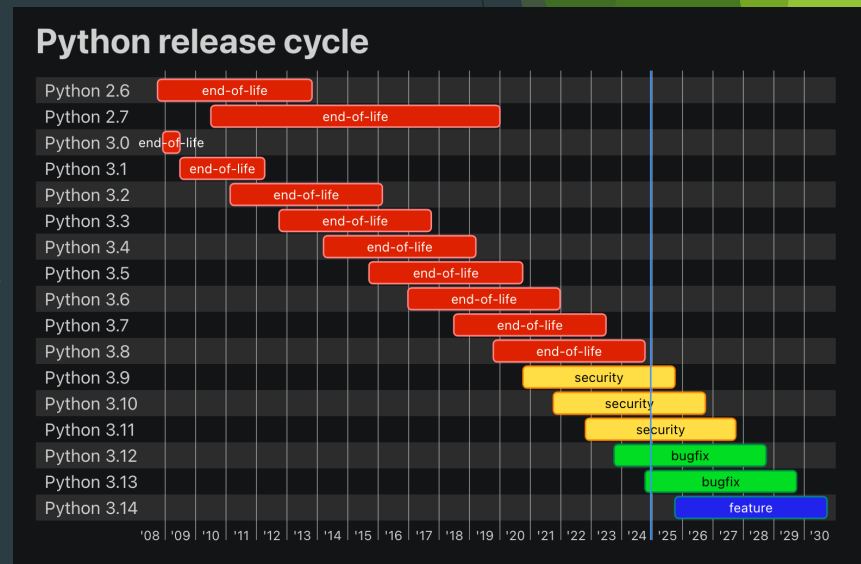
What this course is *not*

- ▶ 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)
 - ▶ Lecture 2 is a Python crash-course, Workshop 0 is hands-on
- ▶ Deep Learning 101
 - ▶ Yes, we will cover deep learning toward the end
 - ▶ There is a LOT more to data science than throwing a ResNet at it

Recent (ie, ~2022) changes to CSCI 3360 have made this point problematic. Needless to say I'm aware DS1 no longer teaches much (if any) prob/stat/linalg, so we'll take this one step at a time.

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: **3.10 - 3.12**
 - ▶ Absolutely no earlier than 3.9
 - ▶ <https://devguide.python.org/versions/>



Who?

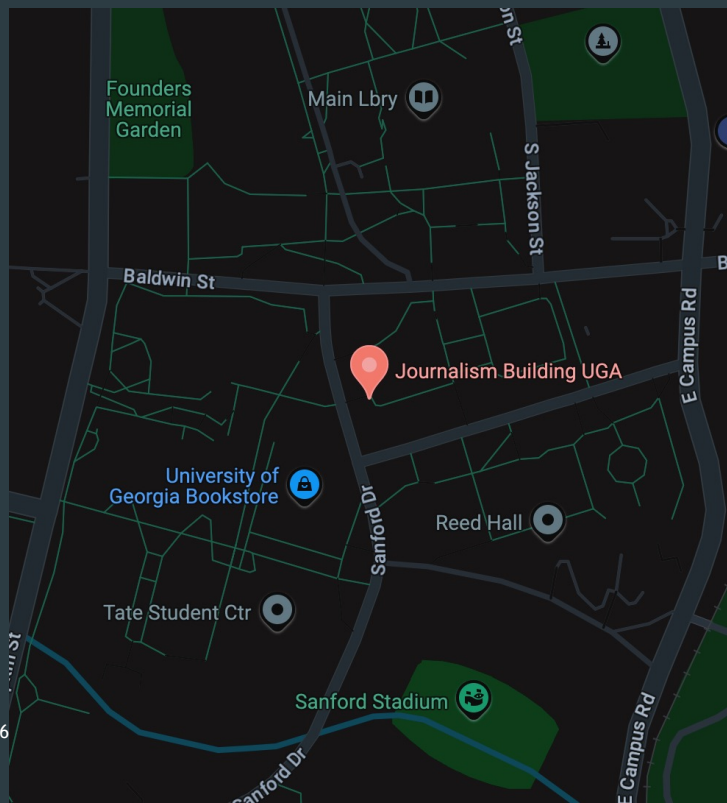
- ▶ Prof. Shannon Quinn (that's me)
 - ▶ 2008: B.S. in Computer Science from Georgia Tech (go Jackets!)
 - ▶ 2010: M.S. in Computational Biology from Carnegie Mellon
 - ▶ 2014: Ph.D. in Computational Biology from joint Carnegie Mellon-University of Pittsburgh Ph.D. Program in Computational Biology
 - ▶ Started at UGA in January 2015
- ▶ Research areas
 - ▶ Biomedical imaging
 - ▶ Computer vision
 - ▶ Distributed computing
 - ▶ Representation learning
 - ▶ **Spatial Biology**

When and Where?

- ▶ Lectures
 - ▶ Tuesdays and Thursdays: 2:20 - 3:35pm, Journalism 0501 (here!)
- ▶ Workshops
 - ▶ Wednesdays: 3:00 - 3:50pm, Dawson Hall 0310
- ▶ Office Hours
 - ▶ Discord (aka *virtual*)
 - ▶ Tuesdays, 1:00 - 2:00pm
 - ▶ Thursdays, 12:00 - 1:00pm
 - ▶ Or by appointment
- ▶ TA: TBD
 - ▶ Will announce when they have been assigned

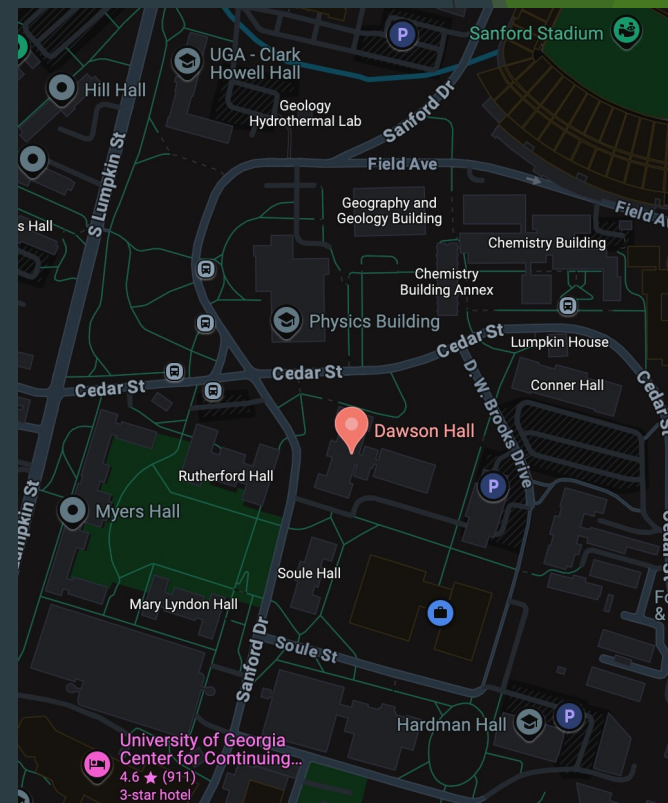
When and Where?

Tuesday / Thursday



CSCI 4360/636

Wednesday



University of Georgia
Center for Continuing...
4.6 ★ (911)
3-star hotel

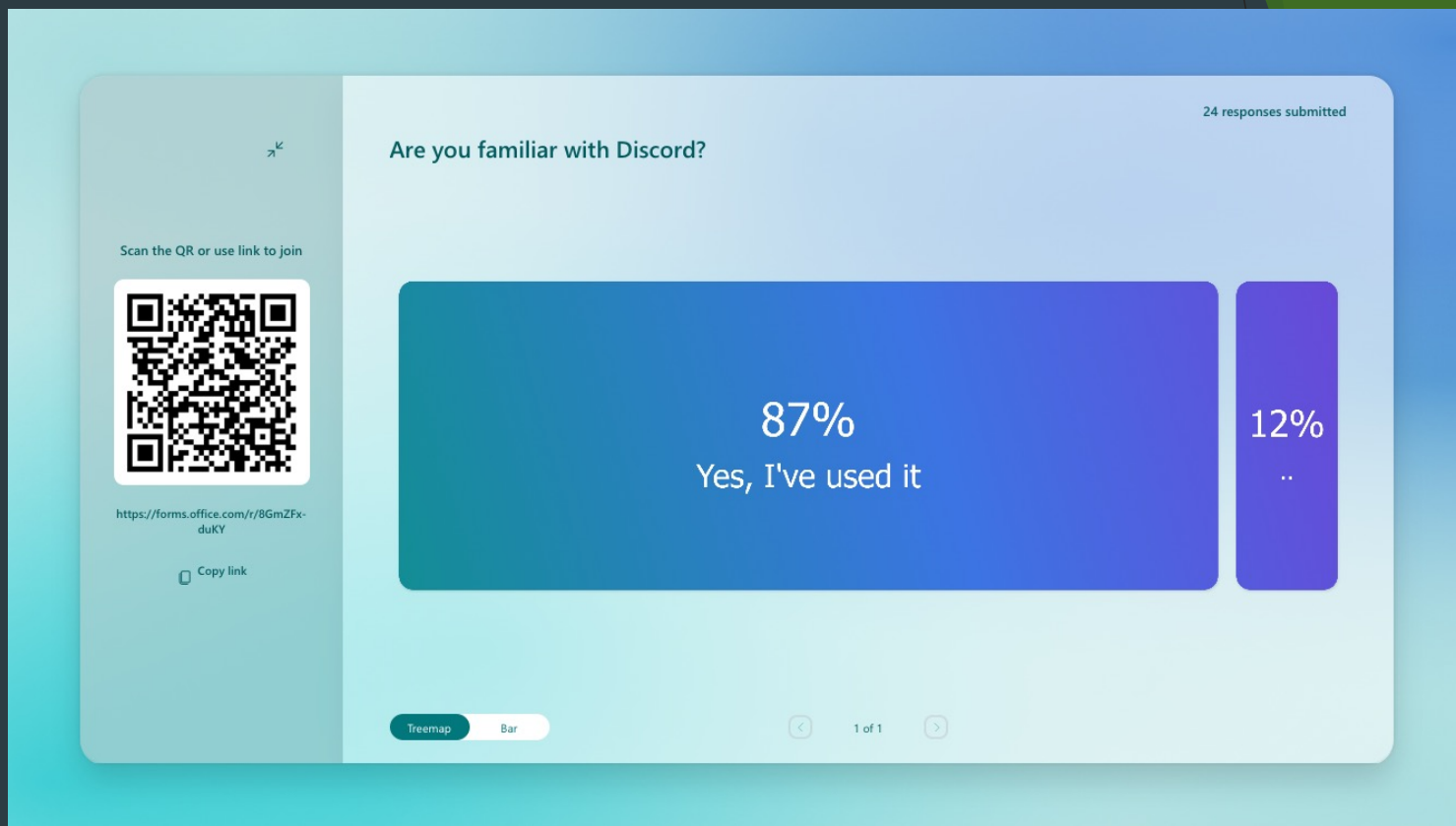
How?

- ▶ Course website / syllabus: <https://eds-uga.github.io/csci-x360-sp25/>

WELCOME TO DATA SCIENCE II

SPRING 2025 EDITION

How?



How?

Welcome to CSCI 4360/6360 Data Science II

This is the beginning of this server.

► Discord server

A screenshot of a Discord server's channel list. The list is organized into two main sections: FORUMS and MEETINGS. The FORUMS section includes channels for announcements, questions, techprobs, and lounge. The MEETINGS section includes voice channels for Lectures, Office Hours, and Voice Lounge, each with a timer set to 00/40. The interface is dark-themed with green accents.

- FORUMS
 - # announcements
 - # questions
 - # techprobs
 - # lounge
- MEETINGS
 - Lectures (00/40)
 - Office Hours (00/40)
 - Voice Lounge

How?

- ▶ Discord: post in `#questions`
- ▶ Email: squinn@cs.uga.edu
 - ▶ I get tons of emails every day
 - ▶ Discord 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>
 - ▶ AutoLab has a code-checker built in
 - ▶ 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).

Honesty at UGA

A note on chatbots

- ▶ We'll cover the technical aspects of chatbots toward the end of the course, so stay tuned
- ▶ Use of chatbots to help in your assignments is not prohibited... per se
 - ▶ I can't really stop you
 - ▶ But these aren't exactly a panacea if you don't understand something
- ▶ Like any other external source, **cite your use of chatbots** (or collaboration with anyone else) **in your assignment write-ups**
- ▶ I've worked with generative text models for 15+ years (before they were cool)
 - ▶ They're pretty cool... but they have distinctive "tells"

Grading Breakdown

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

Assignments

- ▶ There will be **5**. You are required to complete **3**.
 - ▶ So yes, each is worth **15%** of your grade
 - ▶ Any more than 3 will be extra credit
- ▶ Each will be **2.5 weeks long**
 - ▶ Released on a Thursday afternoon
 - ▶ Due on a Tuesday 2.5 weeks later by **11:59:59pm**
- ▶ Will entail a written and a programming portion
 - ▶ Coding in Python
 - ▶ Writing in Word or LaTeX—**nothing handwritten!**
- ▶ All deliverables will be submitted through **AutoLab**

Workshops

- ▶ Most Wednesdays, we'll have a **workshop**
- ▶ This is **student-led and organized**
 - ▶ Everyone is required to do **one** (teams of 2-3 is fine, even necessary)
 - ▶ Any more is **extra credit** (only available once all required slots have been taken)
- ▶ The objective of each workshop is to **demo** a proof-of-concept for your student colleagues
 - ▶ Should be in the 20-35 minute range
- ▶ 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 (**Feb 27**), what more do you want?
- ▶ (details will be released later)
- ▶ Biggest take-away: you can do **EITHER** the midterm **OR** the final project
 - ▶ Doing both is extra credit

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
 - ▶ **Two updates**, one-page deliverables that lay out 1) what your team has accomplished so far, 2) obstacles you have encountered and how you plan to work around them, and 3) any deviations you anticipate from your proposal
 - ▶ 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!**

“Assignment 0”

- ▶ Accept the invitation to Discord if you haven't already
 - ▶ Or ping me if you haven't received any such invitation
 - ▶ **The Discord is only for registered students**
- ▶ Put together groups of 2-3 students and **pick a date + topic for a workshop** (see the course website for available dates)
 - ▶ First [student-run] workshop is **Wednesday, Jan 15!**
 - ▶ **Sign up here:**
<https://docs.google.com/spreadsheets/d/1S9fuFx47iEiB5z0grZURiRu7aBYdiB3bGUPmob25zyw/edit#gid=0>
- ▶ Register an account on AutoLab
 - ▶ <https://autolab.cs.uga.edu>
 - ▶ Then let me know (e.g., ping me in Discord) so I can add you to the course

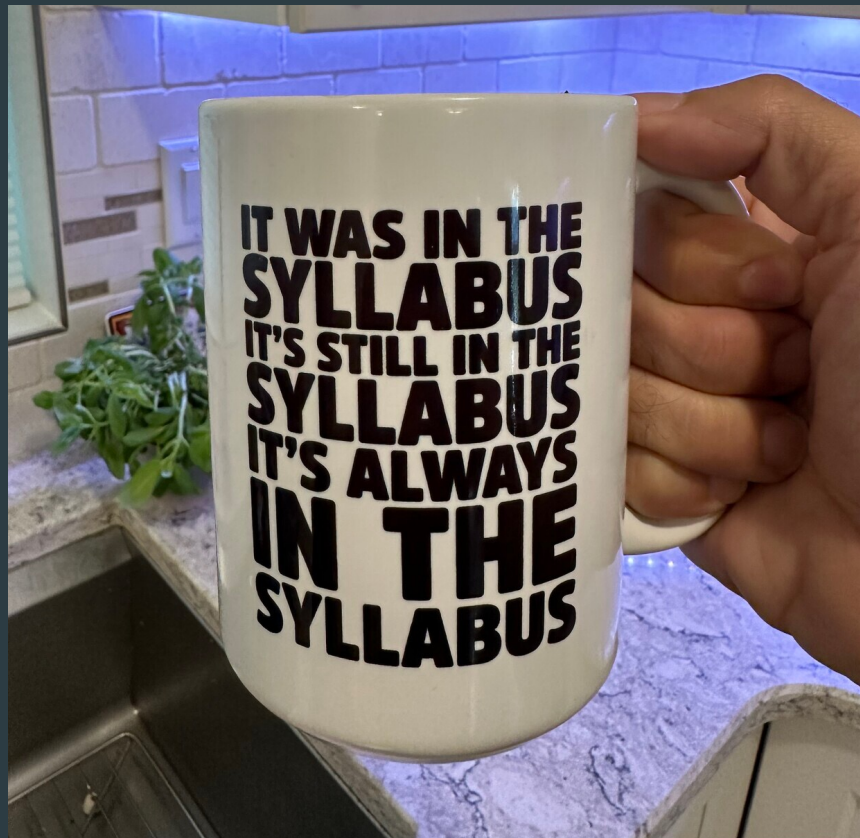
Assignment 1

- ▶ Coming out Thursday (Jan 9)
- ▶ Machine Learning review
 - ▶ Do you remember probability and statistics?
 - ▶ Do you remember linear algebra?
 - ▶ Do some coding, interact with AutoLab, submit a write-up
- ▶ **Due Tuesday, January 28**

Tomorrow

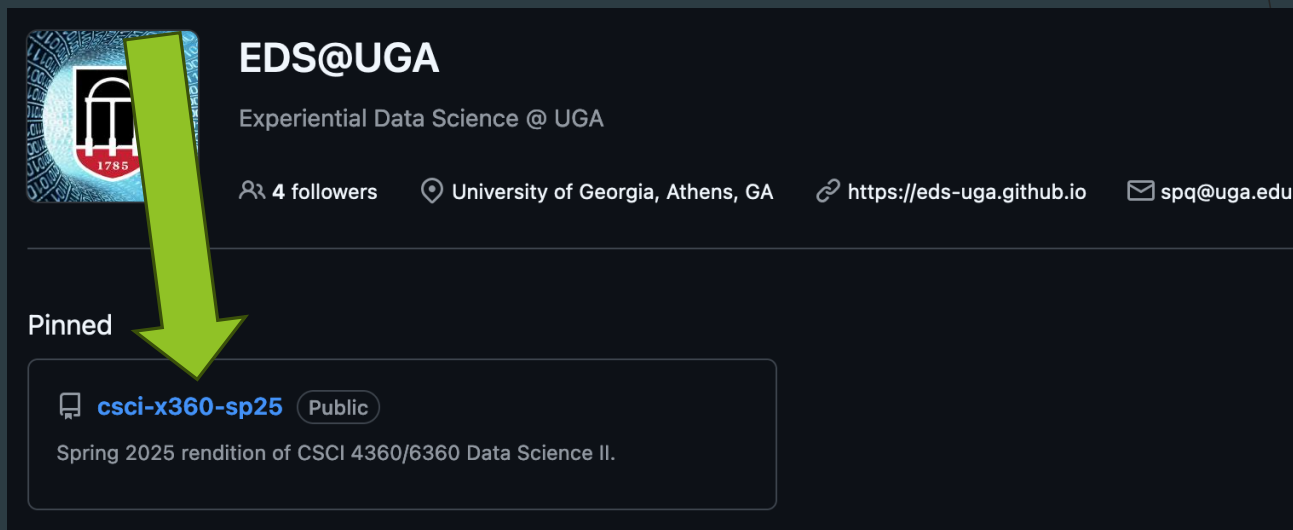
- ▶ Workshop 0, on using the Anaconda distribution for installing and configuring your own Python environment! (far and away the easiest way to get up and running with Python)

Final thoughts



Final thoughts

- ▶ If in doubt, go to: github.com/eds-uga
- ▶ Click the pinned repo



The screenshot shows a GitHub profile for 'EDS@UGA'. The profile name is 'EDS@UGA' and the bio is 'Experiential Data Science @ UGA'. It has 4 followers and is located at the University of Georgia, Athens, GA. The website is <https://eds-uga.github.io> and the email is spq@uga.edu. Under the 'Pinned' section, there is a repository named 'csci-x360-sp25' which is public. The description of the repository is 'Spring 2025 rendition of CSCI 4360/6360 Data Science II.'. A large green arrow points from the top of the repository card to the profile picture.

Questions?



BinderHub



Build and launch a repository

GitHub repository name or URL

GitHub repository name or URL

Git ref (branch, tag, or commit) Path to a notebook file (optional)

HEAD Path to a notebook file (optional) launch

Copy the URL below and share your Binder with others:

Fill in the fields to see a URL for sharing your Binder.

Expand to see the text below, paste it into your README to show a binder badge: launch binder

- ▶ Set up a prototype BinderHub instance for stress testing
 - ▶ The technology behind mybinder.org
 - 1. Put your Jupyter notebook[s] in a [public] GitHub repo
 - 2. Specify the environment (environment.yml, requirements.txt)
 - 3. Submit the link to your GitHub repo
- ▶ Learn more here <https://mybinder.readthedocs.io/en/latest/>

BinderHub

<https://hub.binder.itg.cs.uga.edu/>



- ▶ Great for testing your programs in an isolated and replicable environment
- ▶ Even better if your laptop isn't up to running Python
- ▶ Takes a very long time to build environments (~30 minutes)
- ▶ No GPU access

- ▶ **Send me feedback on your experience with it! We're wondering if this would be something useful for the whole campus.**