Data Science

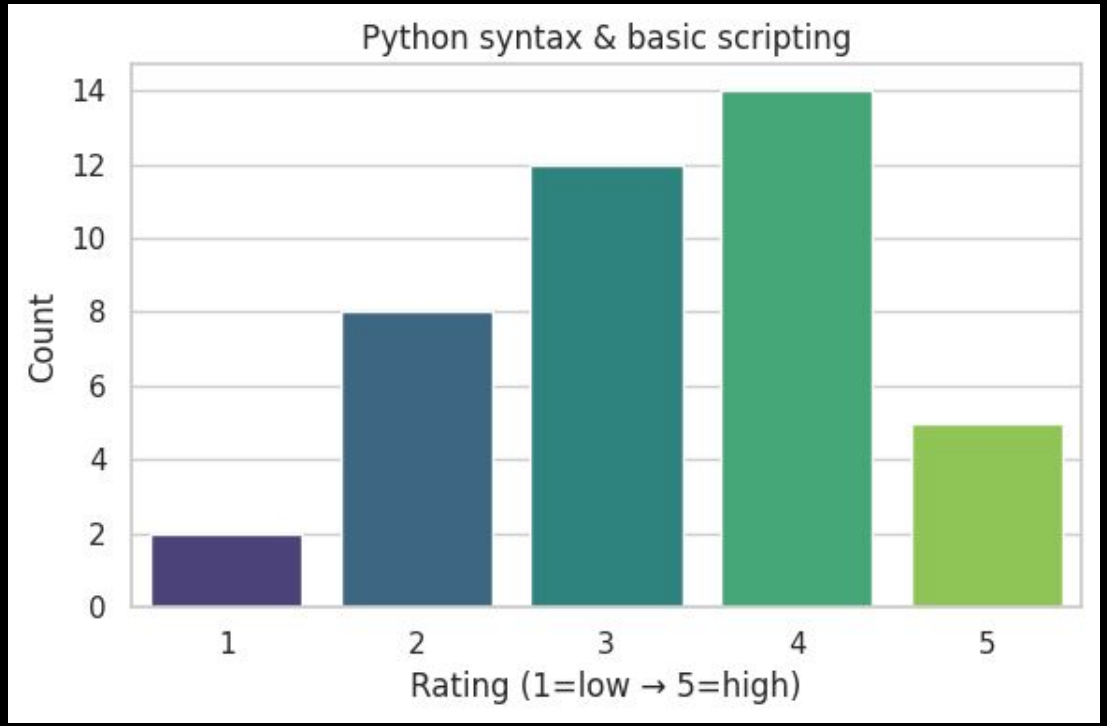
"Computational Statistics Across the Sciences"
CAS PY 191S 1310

Attendance

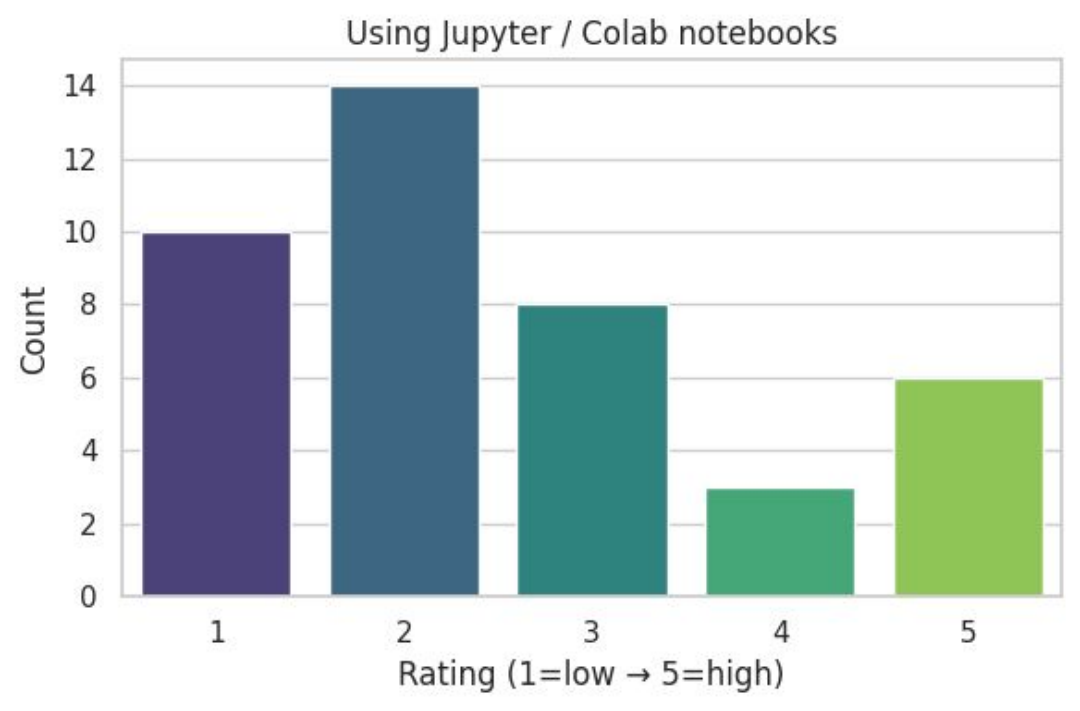
https://forms.gle/iX5EvN6pLrqi2rej8

Introductions

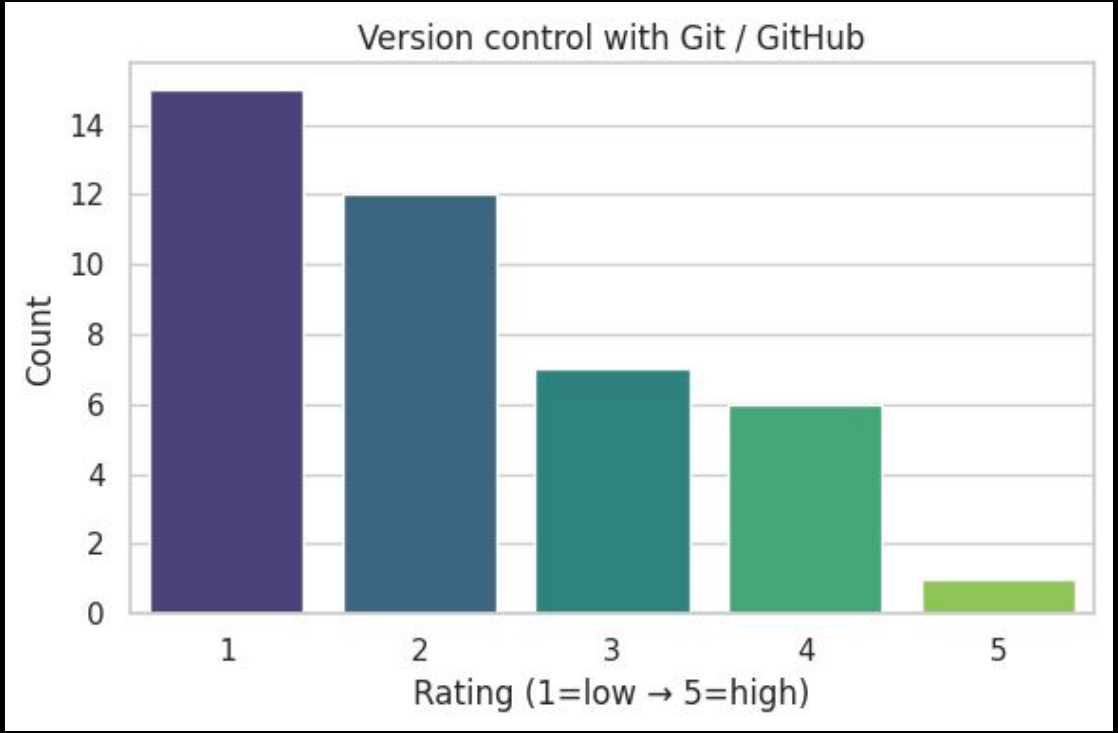
A bit about our class



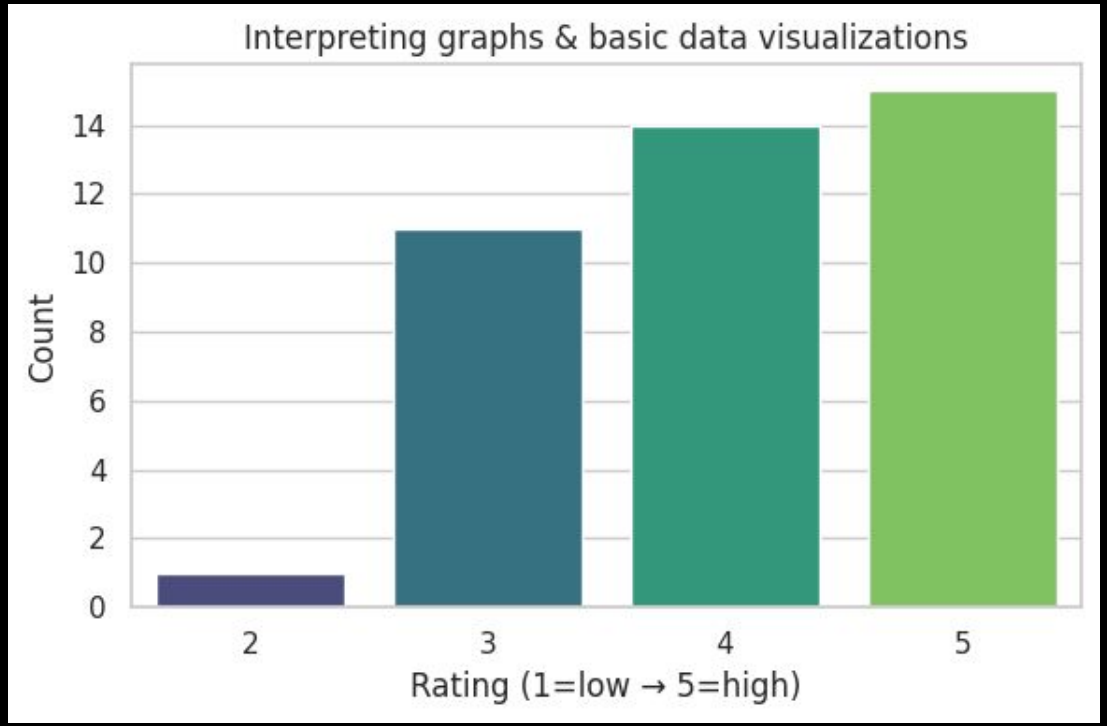
A bit about our class



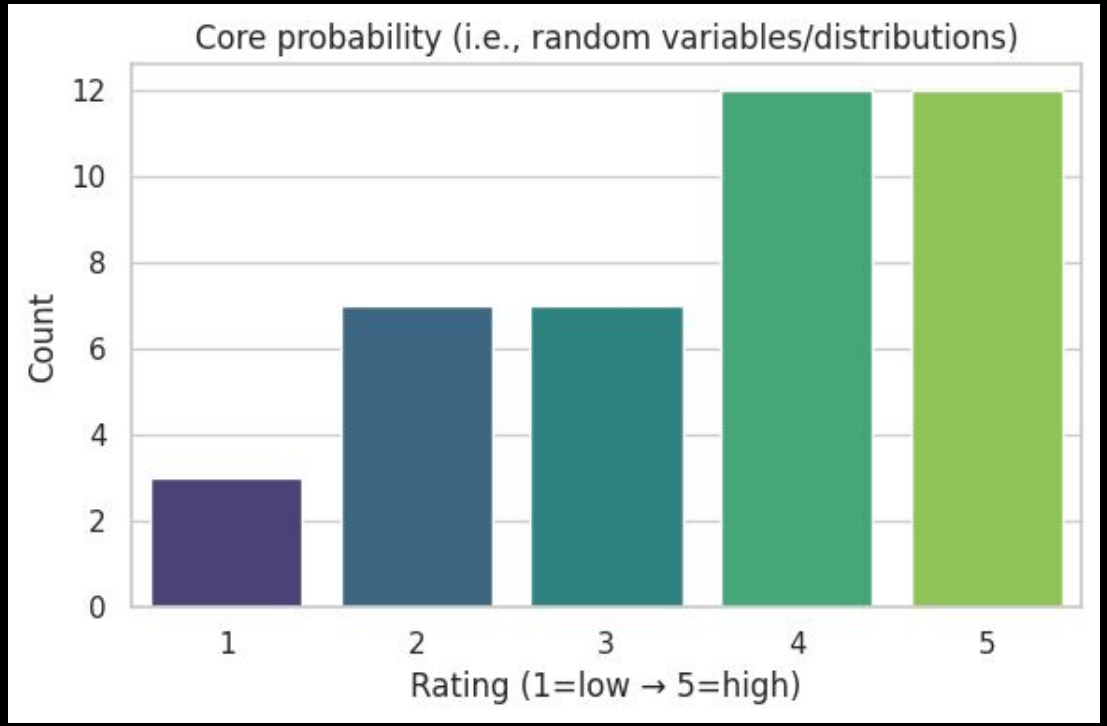
A bit about our class



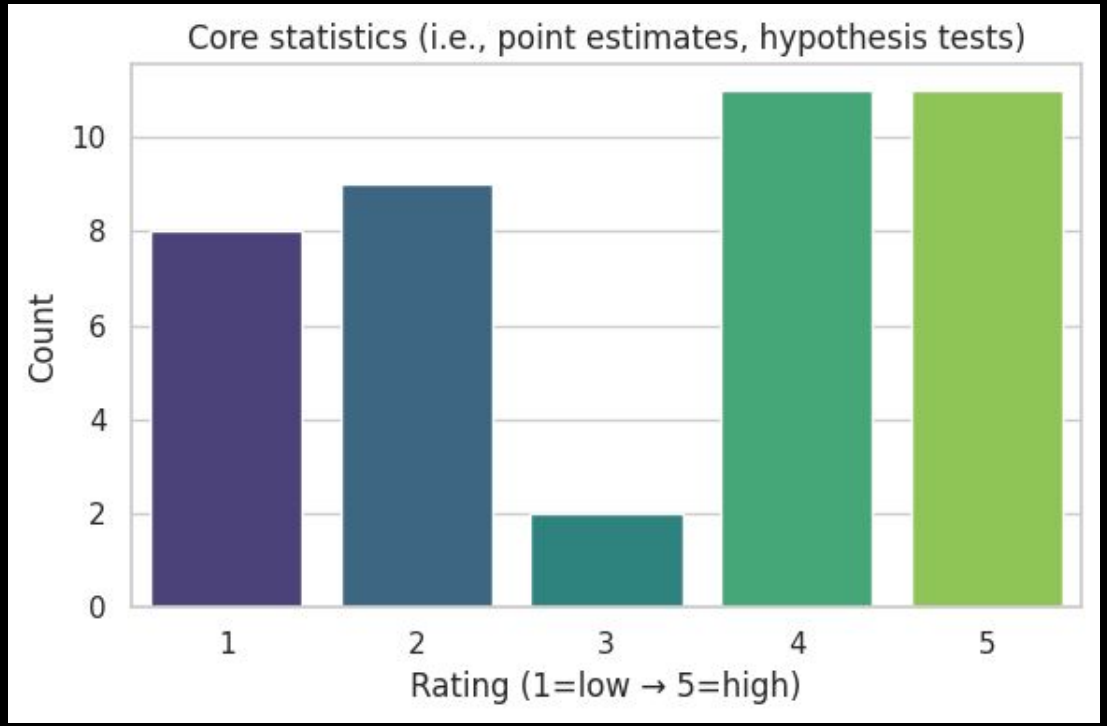
A bit about our class



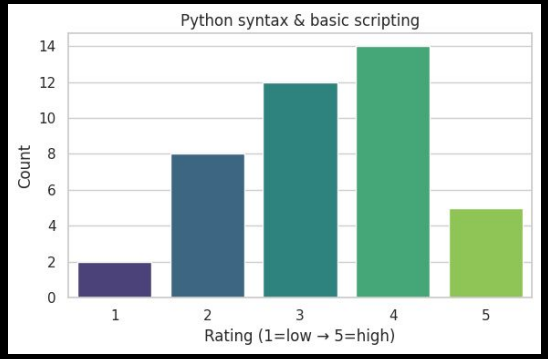
A bit about our class

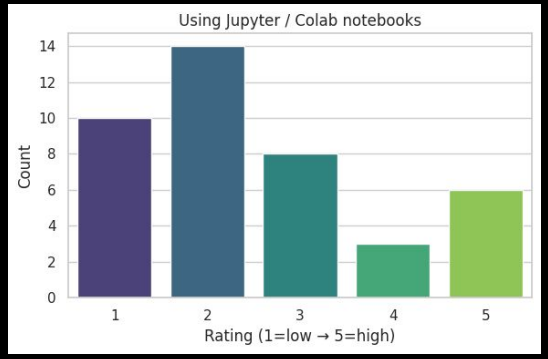


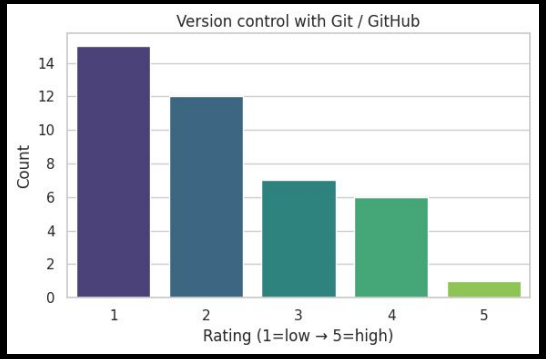
A bit about our class

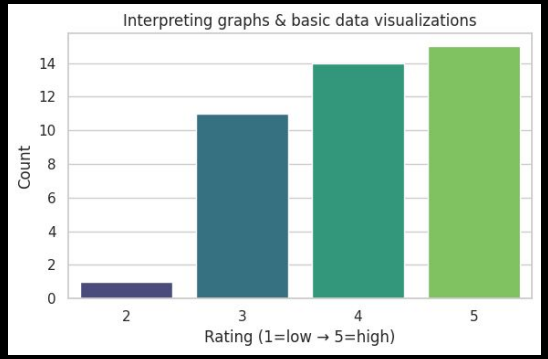


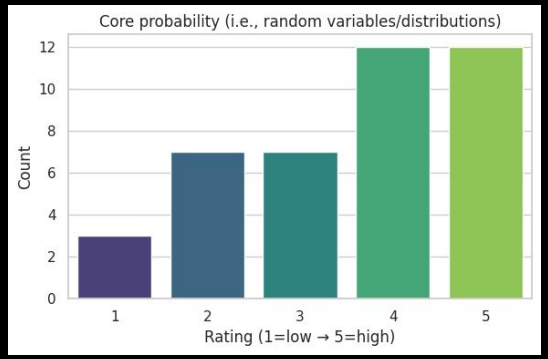
A bit about our class

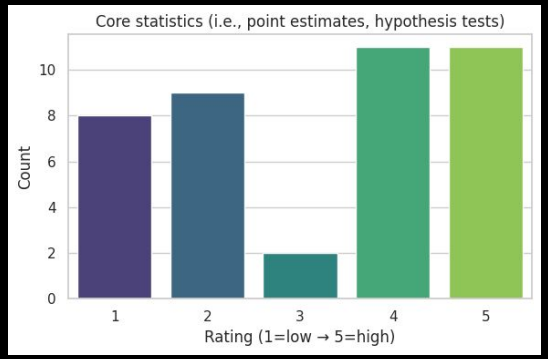




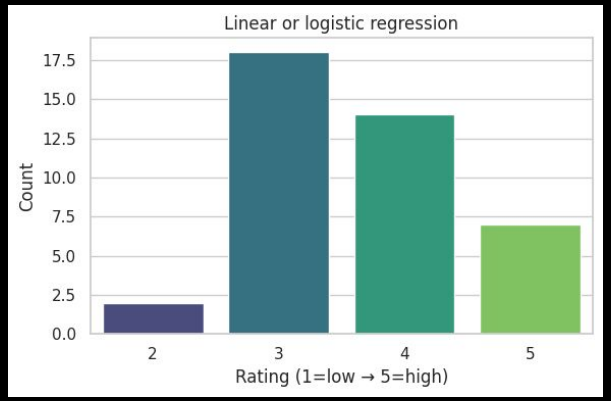


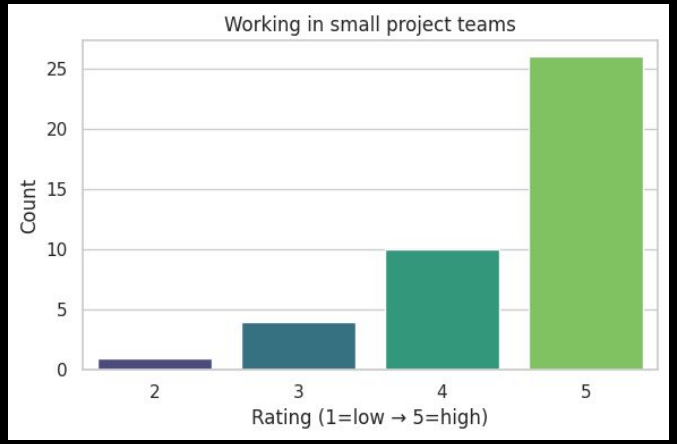


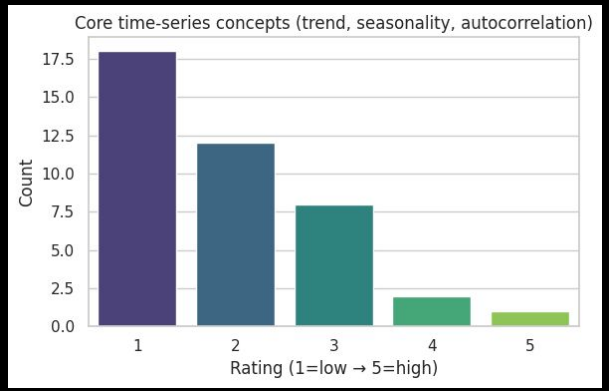




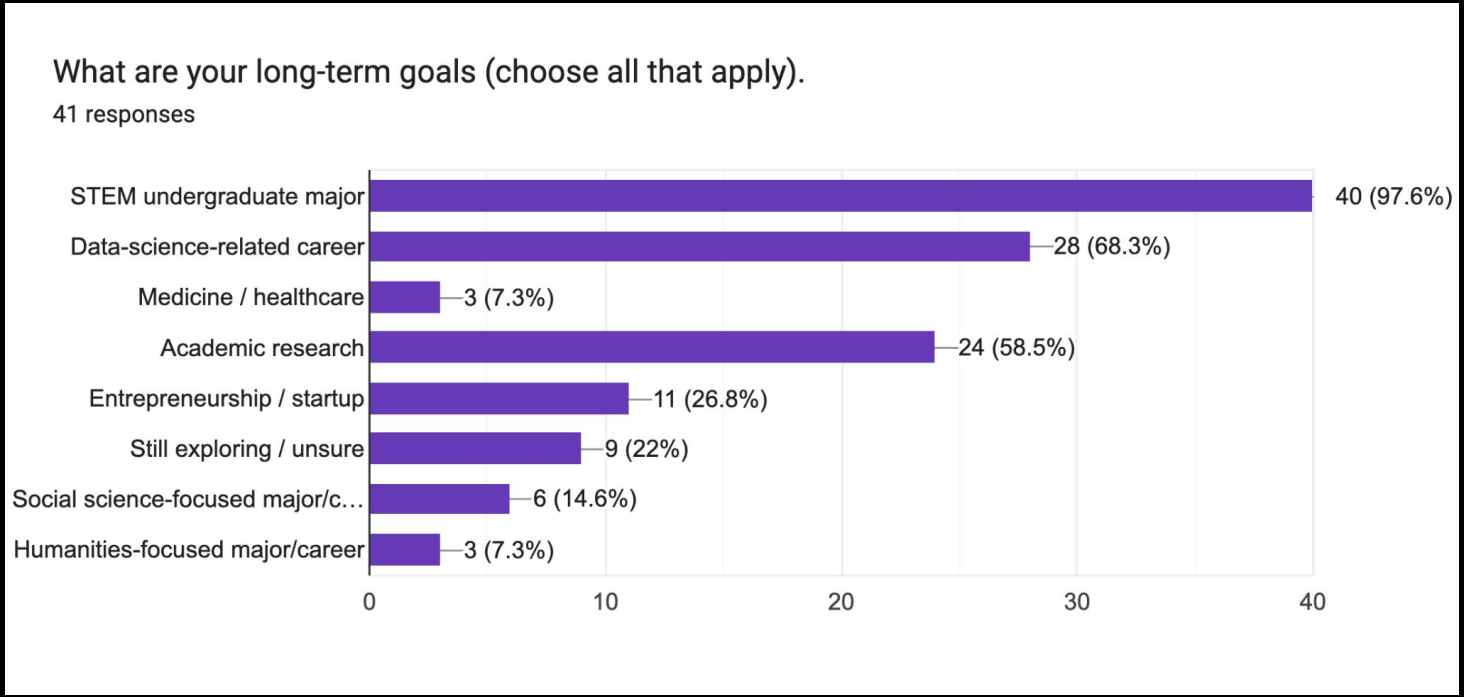
A bit about our class







A bit about our class

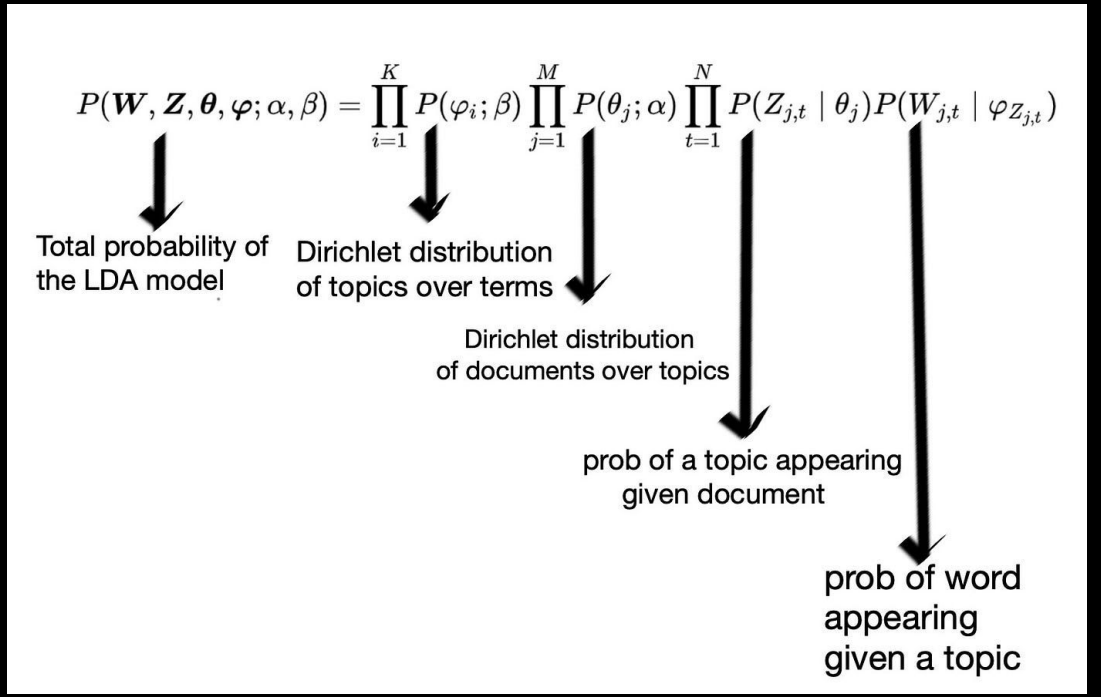


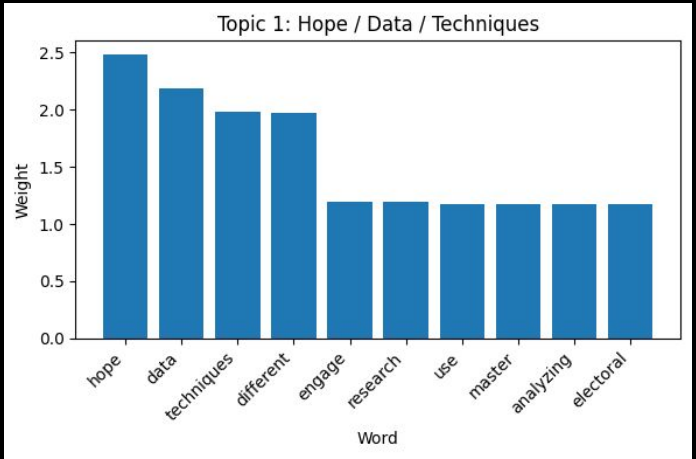
Open response questions:

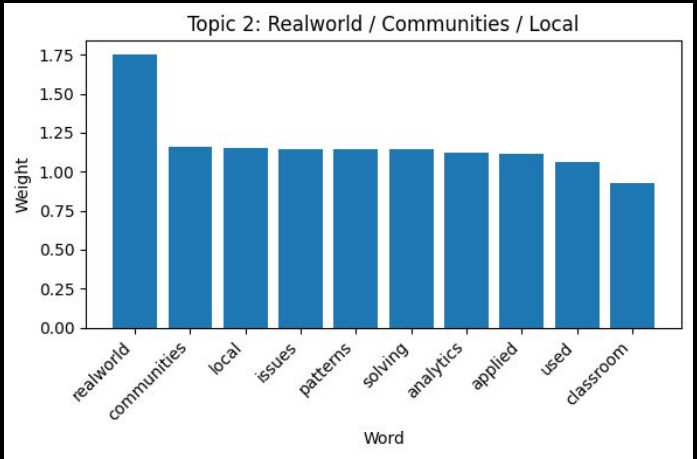
1) In 1–2 sentences, what do you hope to create or learn during this course?
2) In 1–2 sentences, what do you hope to get out of the hands on research/computational project for this course?

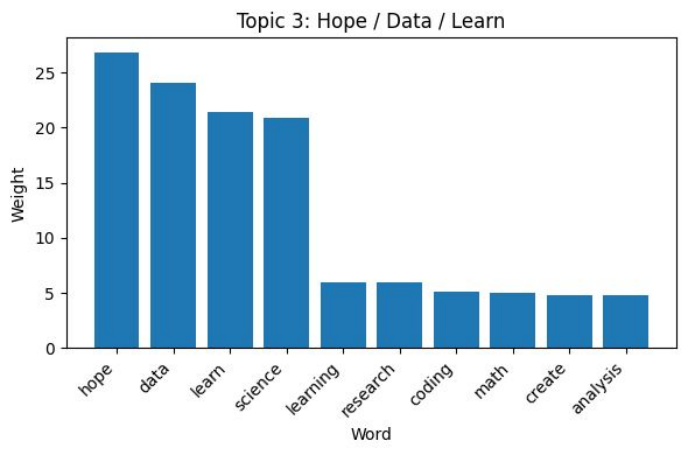
Topic Modeling (Latent Dirichlet Allocation)

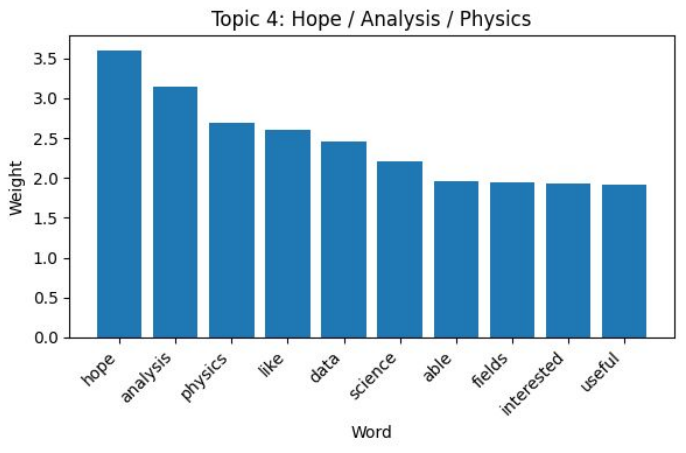
- Bayesian probabilistic model used for topic modeling in natural language processing
- It assumes that each document is a mixture of different topics, and each topic is a mixture of words.
- Helpful for identifying core themes



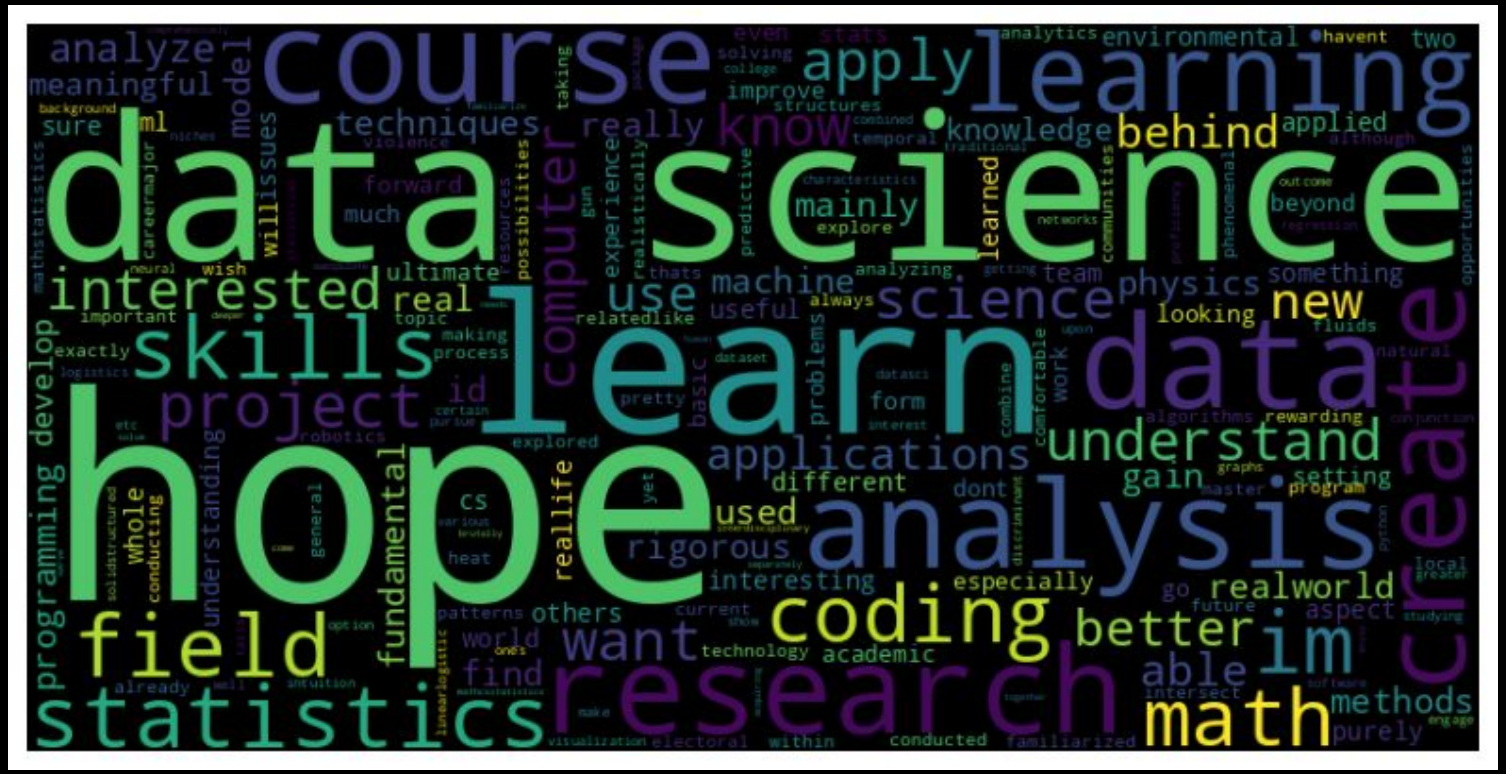

Topic 1: Hope / Data / Techniques
Weight
Word
hope
data
techniques
different
engage
research
use
master
analyzing
electoral


Topic 2: Realworld / Communities / Local
Weight
Word
realworld
communities
local
issues
patterns
solving
analytics
applied
used
classroom


Topic 3: Hope / Data / Learn
Weight
Word
hope
data
learn
science
learning
research
coding
math
create
analysis


Topic 4: Hope / Analysis / Physics
Weight
Word
hope
analysis
physics
like
data
science
able
fields
interested
useful

Most commonly used words in 1-2 sentence reflection



Recall our pre class survey: “What are your goals for this summer?”

- Learning to code from scratch?
- Learning data science best practices?
- Learn about a potential college major or career?
- Learn how it feels to be at an R1 university?
- If you’re a residential student, what it feels like to live away from home?
- …

BOSTON UNIVERSITY

Learning Objectives for the Course

What are my overall goals for you?

- Learn how to be **kind** and **humble** in a technical environment.
  - How do you give good feedback
  - How do we make diverse educational backgrounds and interests a strength
  - How do we decouple knowing a lot/the right answer from being valuable

What are my overall goals for you?

- Learn how to be **kind** and **humble** in a technical environment.
  - How do you give good feedback
  - How do we make diverse educational backgrounds and interests a strength
  - How do we decouple knowing a lot/the right answer from being valuable
- Learn how to be **rigorous** computational statisticians in a world of generative AI.
  - Core skills in coding, testing, computational reproducibility, statistics, and communication

What are my overall goals for you?

- Learn how to be **kind** and **humble** in a technical environment.
  - How do you give good feedback
  - How do we make diverse educational backgrounds and interests a strength
  - How do we decouple knowing a lot/the right answer from being valuable
- Learn how to be **rigorous** computational statisticians in a world of generative AI.
  - Core skills in coding, testing, computational reproducibility, statistics, and communication
- Learn about the value of **Open Access, Open Source** data science projects

Technical Skills

- Strong conception of the data science life cycle
- Strong Python coding skills
  - Basic data types, object oriented programming
  - Unit tests
  - Version control and project maintency through Github
  - Basic code review
- Basic statistical skills in major areas of the field
  - **Learn how to learn probability and statistics**
- Responsible project communication

Core Team – Two instructors

Patrick Bloniasz

Dr. Eugene Pinsky

Core Team – Four Teaching Fellows (Kevin)

Kevin

Tharunya

Tejovan

Zhengyang

About Me

- Undergraduate: Neuroscience, Digital and Computational Studies
- PhD Candidate in Computational Neuroscience (all but dissertation)

About Me

- Undergraduate: Neuroscience, Digital and Computational Studies
- PhD Candidate in Computational Neuroscience (all but dissertation)
- Previous teaching experience:
  - Adjunct Faculty and Researcher at New York University (Cognitive Science)
  - RISE Neurobiology Dry and Wet Lab Teaching Fellow (2 years)
  - Teaching Fellow at Bowdoin (Biomathematics, Biostatistics, Mathematical Statistics, Intro to DCS; 3 years)

About Me

- Undergraduate: Neuroscience, Digital and Computational Studies
- PhD Candidate in Computational Neuroscience (all but dissertation)
- Previous teaching experience:
  - Adjunct Faculty and Researcher at New York University (Cognitive Science)
  - RISE Neurobiology Dry and Wet Lab Teaching Fellow (2 years)
  - Teaching Fellow at Bowdoin (Biomathematics, Biostatistics, Mathematical Statistics, Intro to DCS; 3 years)
- Research Themes:
  - Stochastic Processes (Time Series Analysis); Point Process/Queuing Theory
  - Statistical Signal Processing
  - Electrical Field Potentials, Electroencephalography
  - Anesthesia (Propofol, Ketamine, Sevoflurane)

How to think of me as an academic resource

Even if we don't overlap in career paths, undergraduate styles of education, or research areas, there are still ways to grow:

- Connect you to researchers who **do** have the experiences you're interested in
- Provide you with guidance on how to enter into a field
- Identify skills required to get onto a particular path

Advice: Talk to your afternoon instructor and teaching fellows

Group Discussion: What are data?

Data are ...

Data are individual facts, statistics, or items of information—usually numeric or categorical—that represent measurements or observations about the world.

Data serve as the **foundational input for analysis, modeling, and decision-making**

Data are …

Data are individual facts, statistics, or items of information—usually numeric or categorical—that represent measurements or observations about the world.

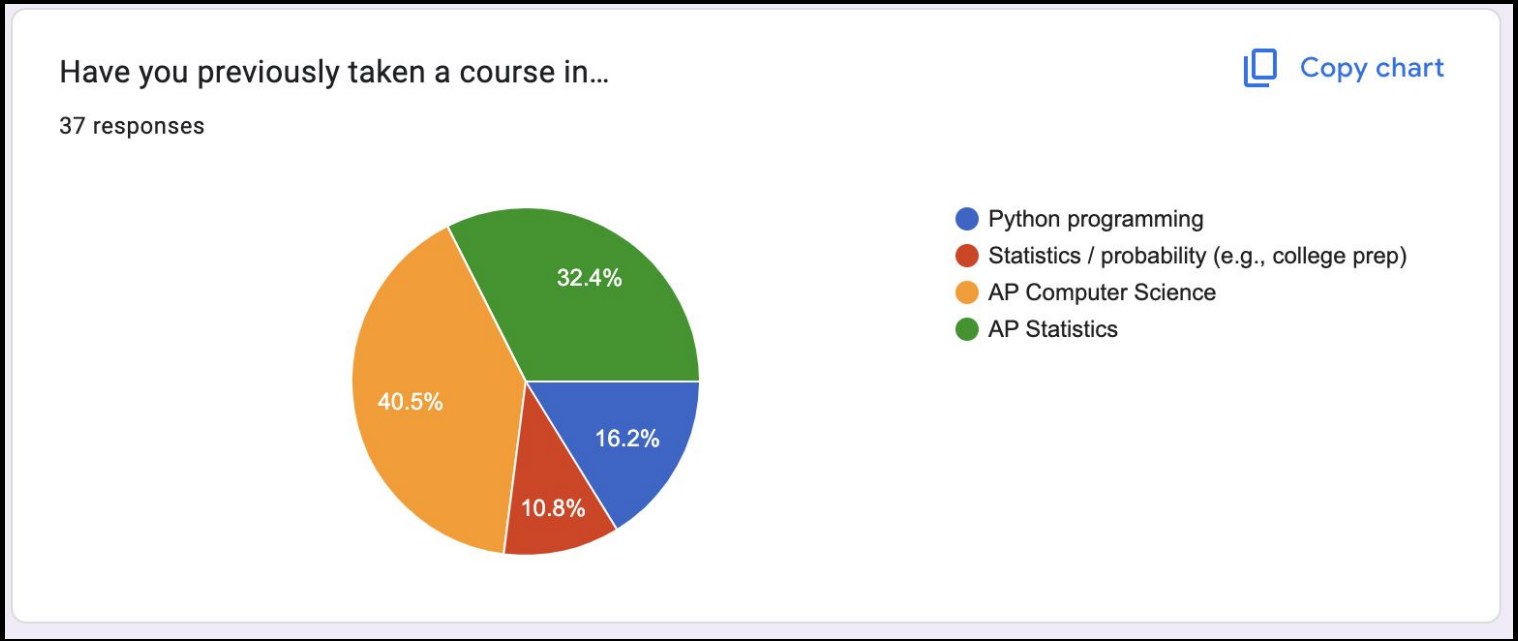
My working definition:

1. Data is any representation of a feature or aspect of a system (e.g., community, human body, physical structure) expressed in abstract terms (e.g., in terms of classes, numbers, vectors etc.) where we have a model of the **generating process.**

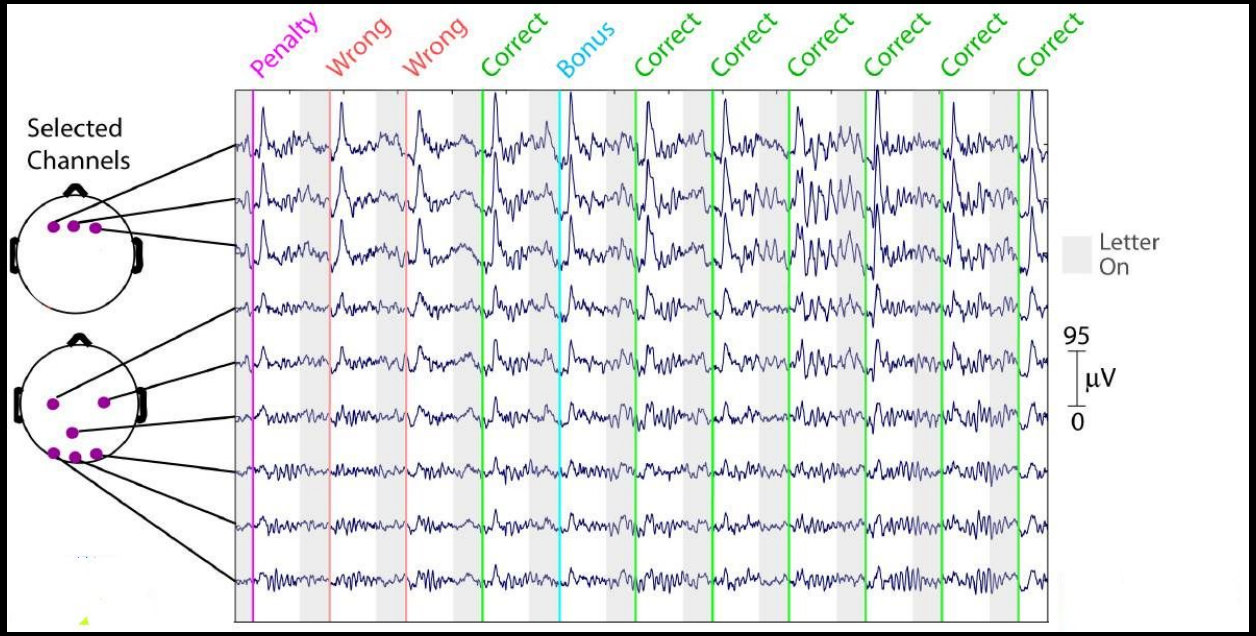
If you don’t know any of the following we should be wary of working with our data:

1. The system feature being represented
2. The measurement tool used to generate the data
3. The limitations or omissions of the measurement tool

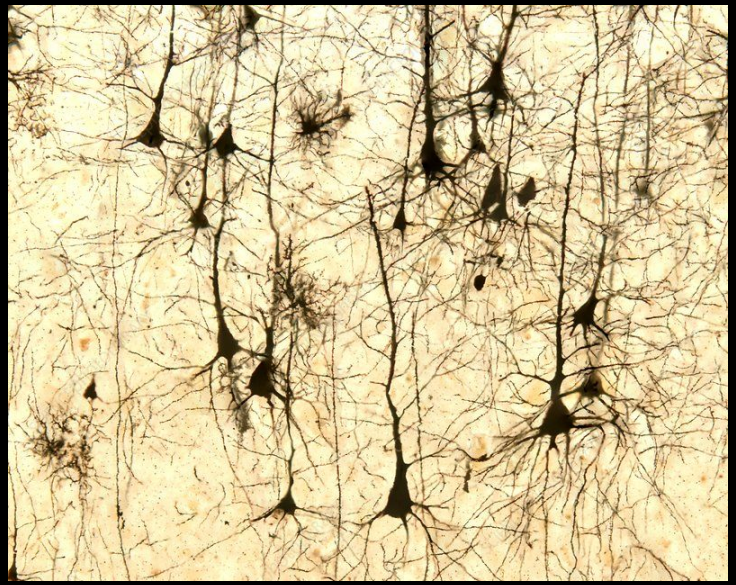
Example: “Have you previously taken a course in…”

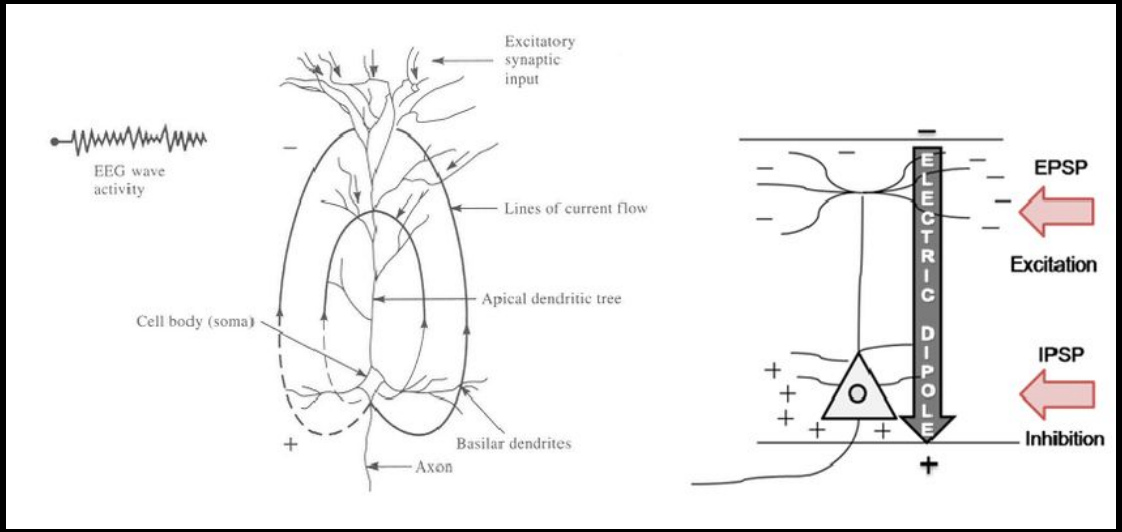


Data Example 1: Brain Data (Signal processing)

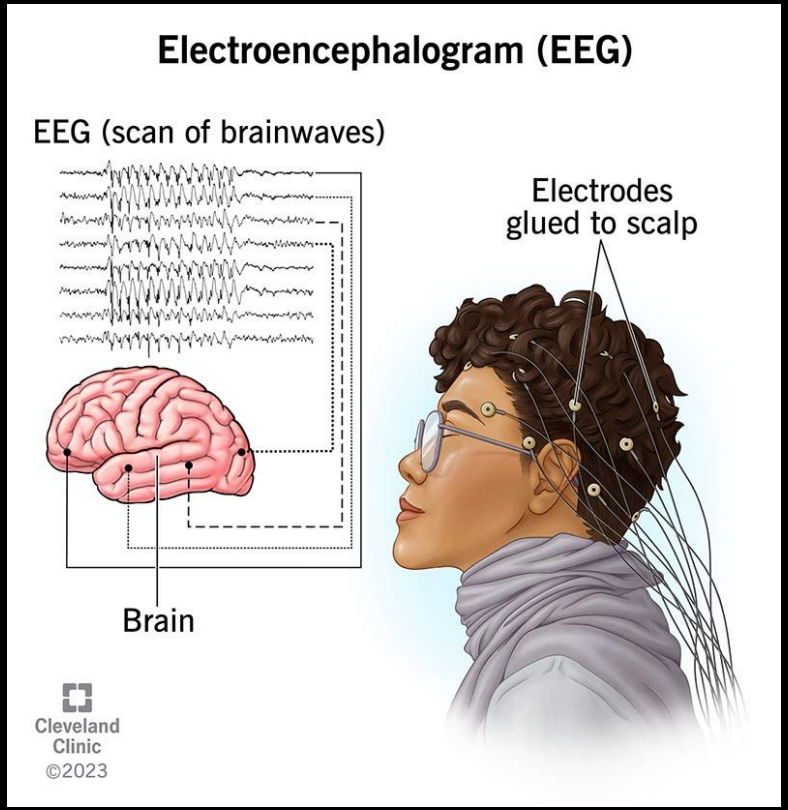


System feature being represented: Cortical neurons in the brain (layer 5)

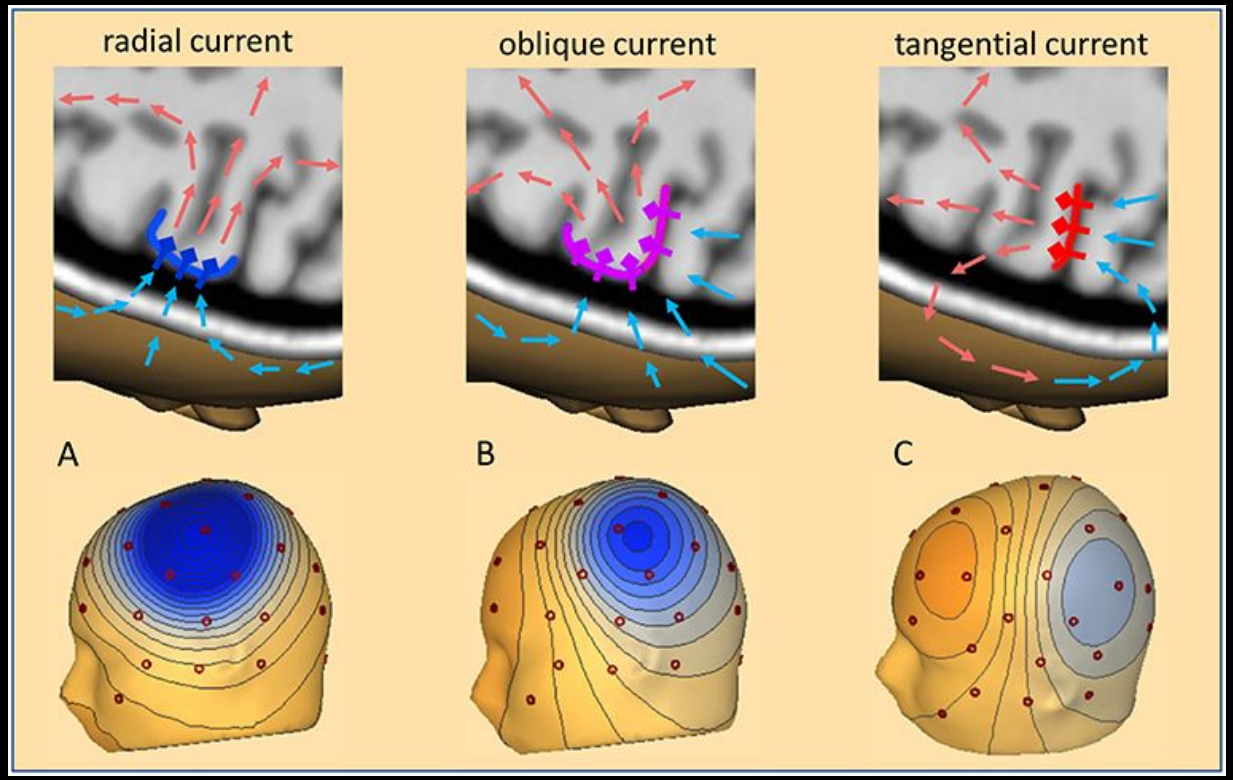




The measurement tool used to generate the data

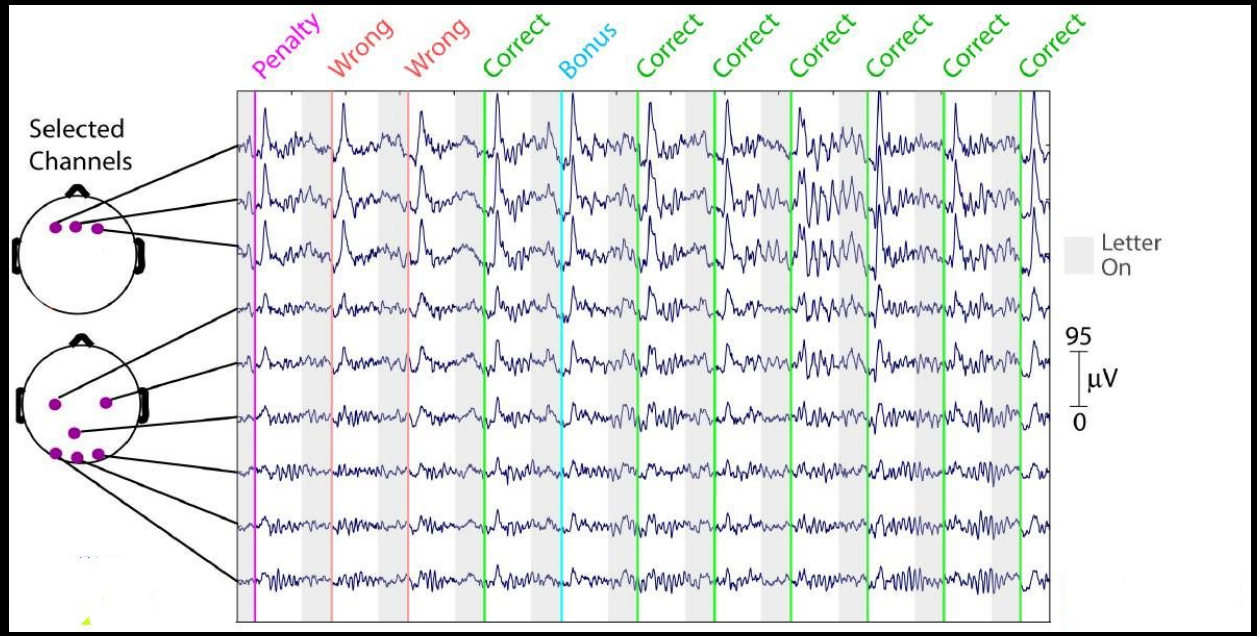


Limitations: Doesn't record the majority of activity



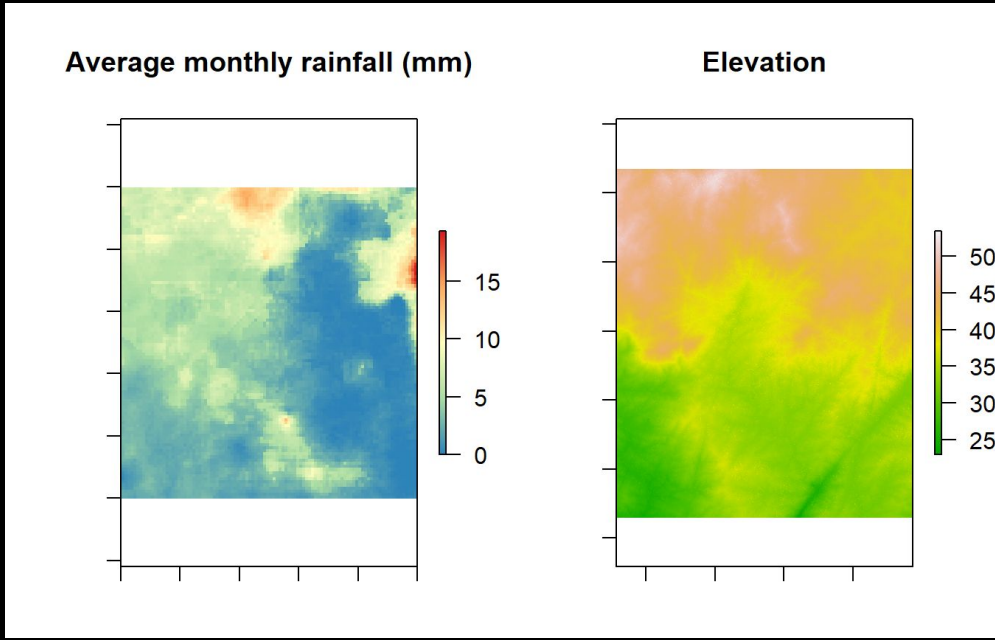
Data Example 1: Brain Data (Signal processing)

1. Cortical neurons in the brain (layer 5)
2. Extracellular electrical recordings (EEG)
3. Ignores all non-electrical signals and electrical signals that don't contribute to "dipoles"



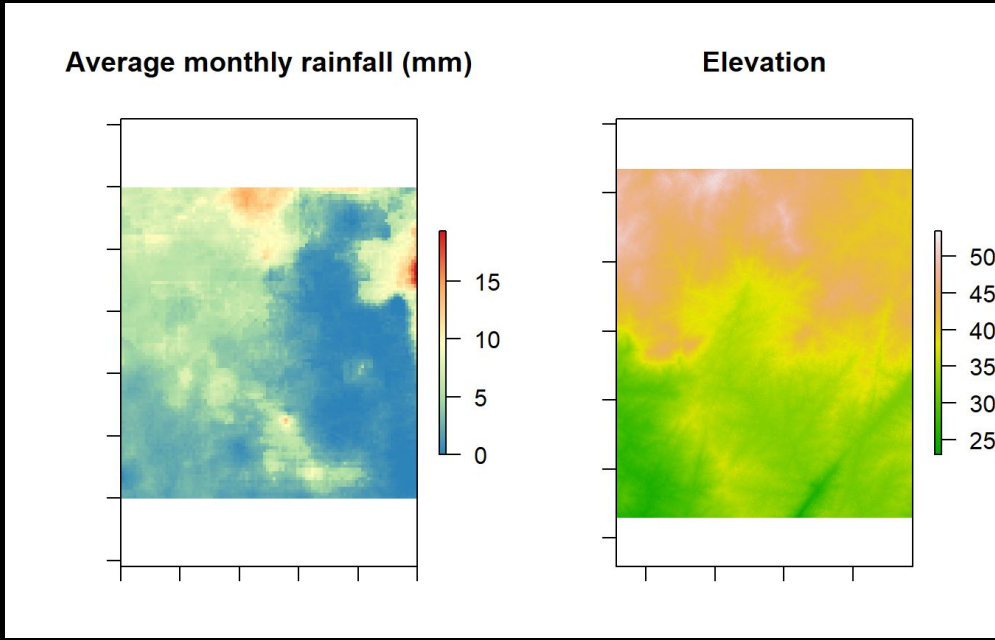
Data Example 2: Environmental Science (Point Processes)

1. The process by which precipitation is deposited in a geographic region, influenced by elevation, topography, and atmospheric circulation.



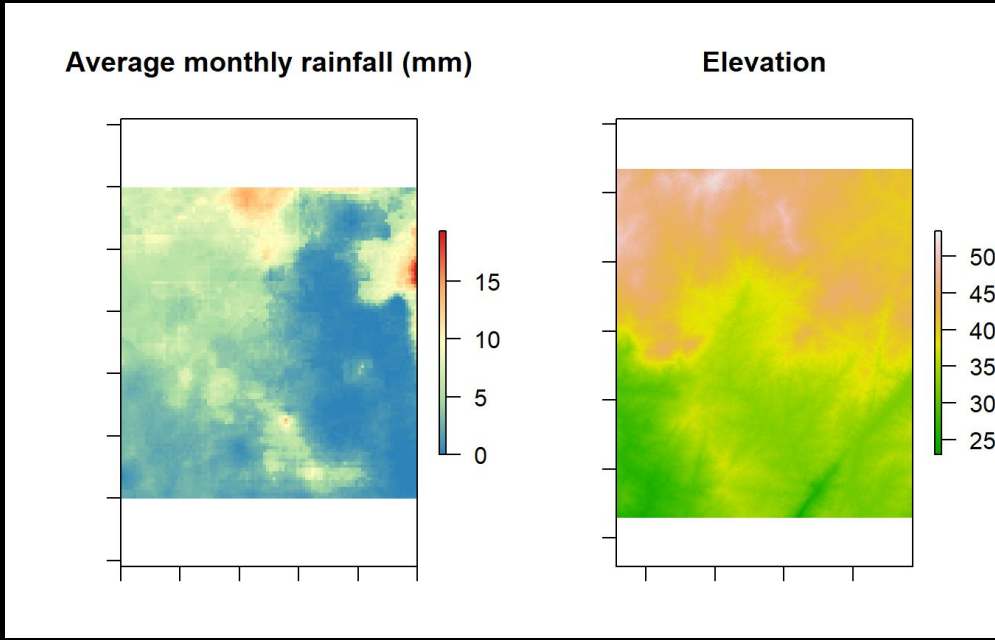
Data Example 2: Environmental Science (Point Processes)

2. Rain gauges at discrete elevation levels that record cumulative precipitation each month.



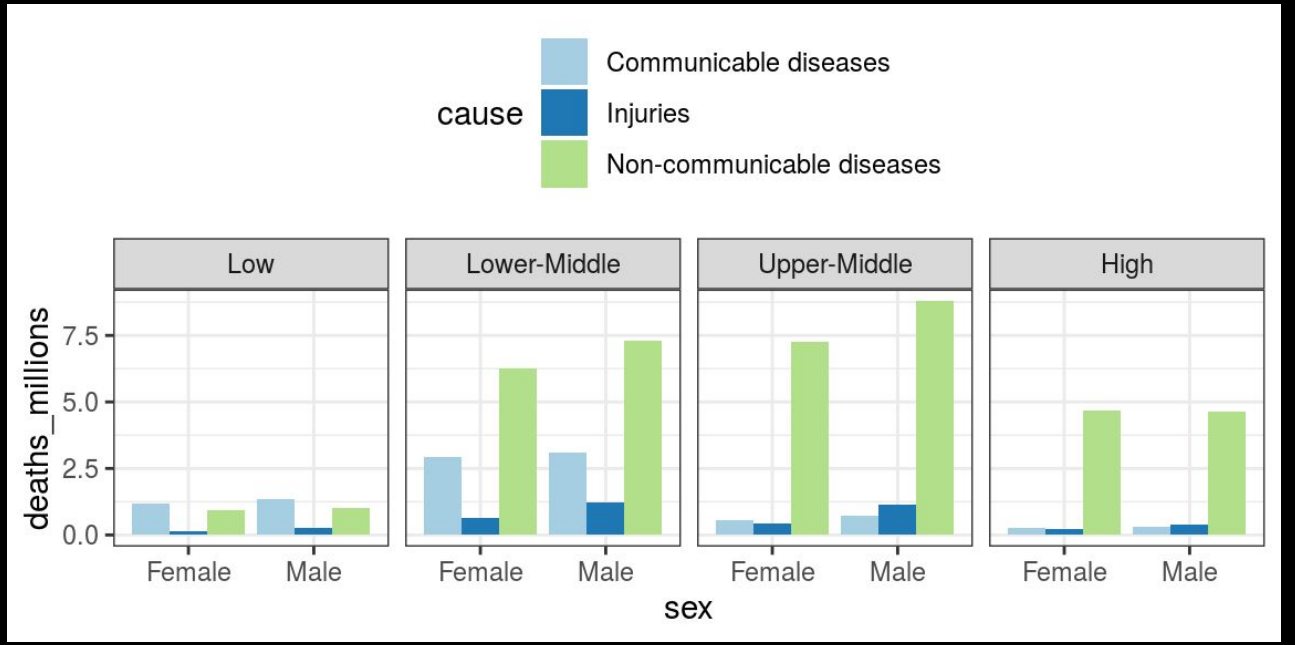
Data Example 2: Environmental Science (Point Processes)

3. Temporal omissions: missing data during equipment downtime; no measurement of sub-monthly rainfall intensity.



Many more…

If you can't identify these for any particular data set, it is an area you need to learn more about or you should find a different data set!

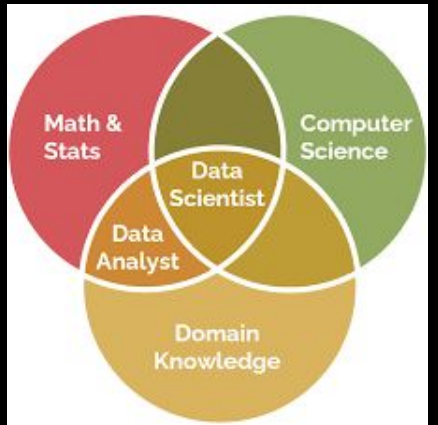


Group Discussion: What is Data Science?

What is Data Science?

Data science is the interdisciplinary field that uses scientific methods, algorithms, and systems to extract knowledge and insights from structured and unstructured data.

Data science transforms raw data into actionable insights—fueling decision-making in business, science, health, and beyond.



Final Projects

You can pick one:

1. Time Series Analysis and/or Prediction – e.g., apply statistical signal processing methods and forecasting techniques to temporal data.

You can pick one:

1. Time Series Analysis and/or Prediction

2. Categorical and Count Data Analysis – e.g., analyze experimental or observational data, comparing conditions and drawing inferences using basic hypothesis testing regimes or regression-based analysis (GLMs) in both or either frequentist and/or Bayesian frameworks.

You can pick one:

1. Time Series Analysis and/or Prediction

2. Categorical and Count Data Analysis

3. Scientific Software Package Development – e.g., implement a statistical method or algorithm in Python as a reusable package, with well-documented code, unit tests, and user tutorials.

You can pick one:

1. Time Series Analysis and/or Prediction

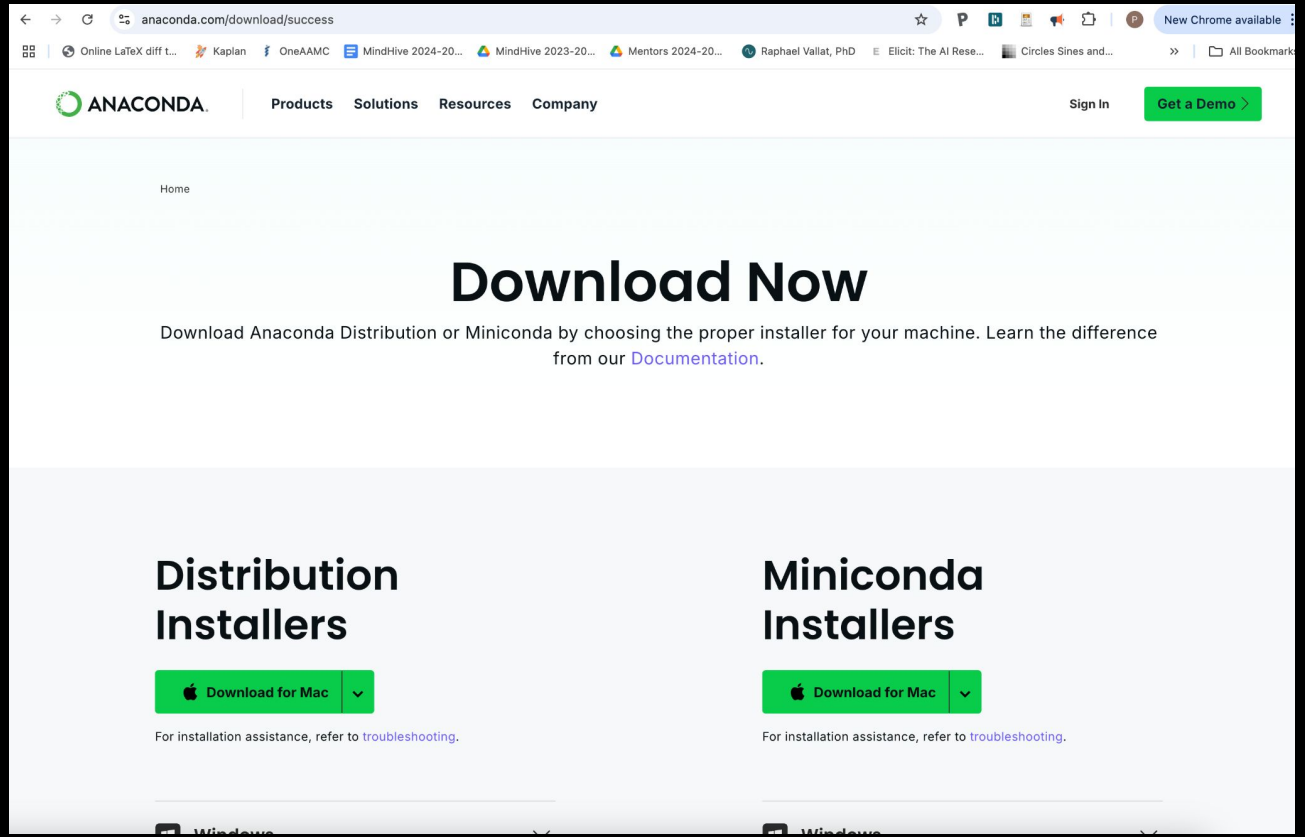
2. Categorical and Count Data Analysis

3. Scientific Software Package Development

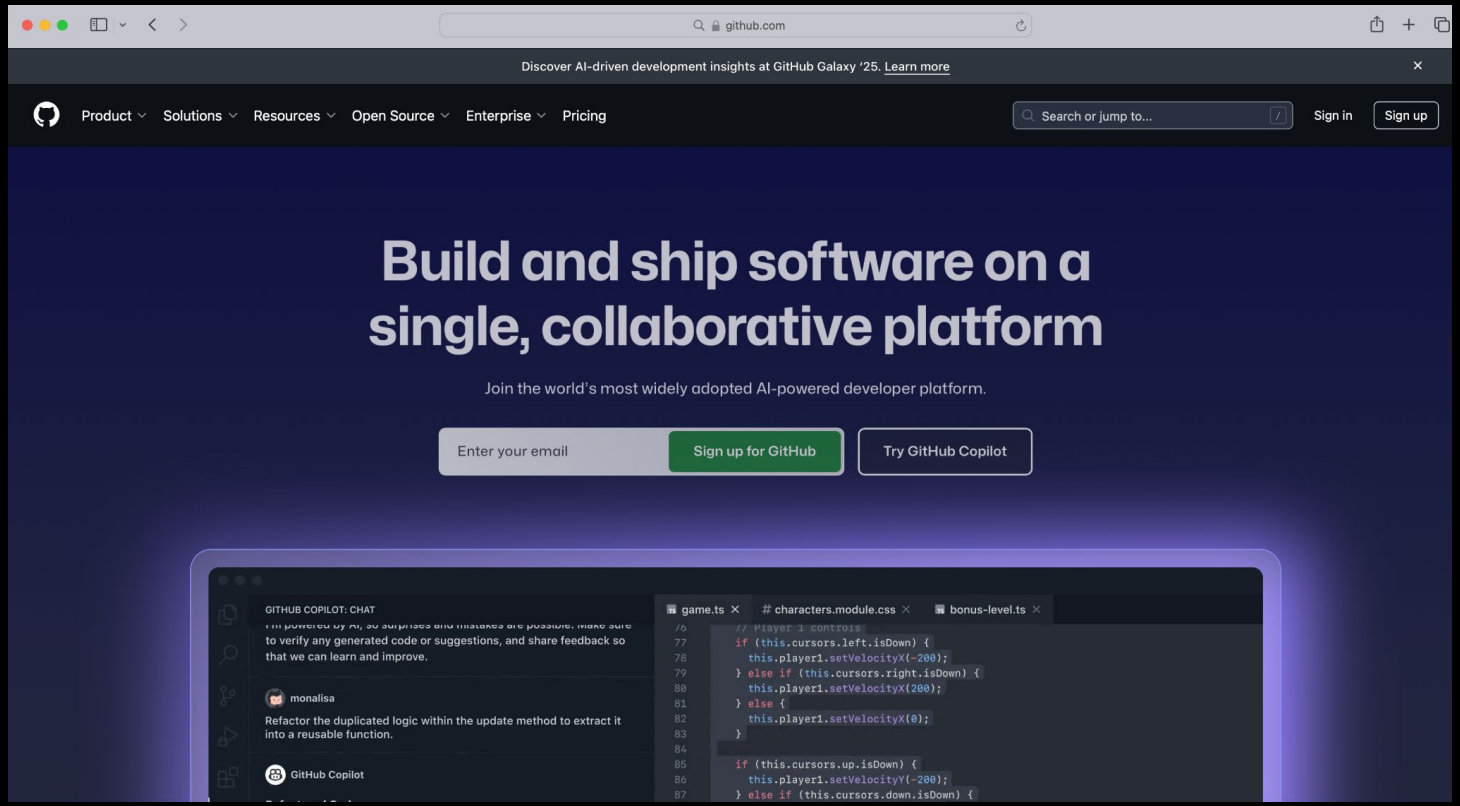
4. Teaching a Statistical Technique – e.g., create an interactive lesson in a Jupyter Notebook to explain a concept including narrative, examples, and exercises.

Classroom Tools

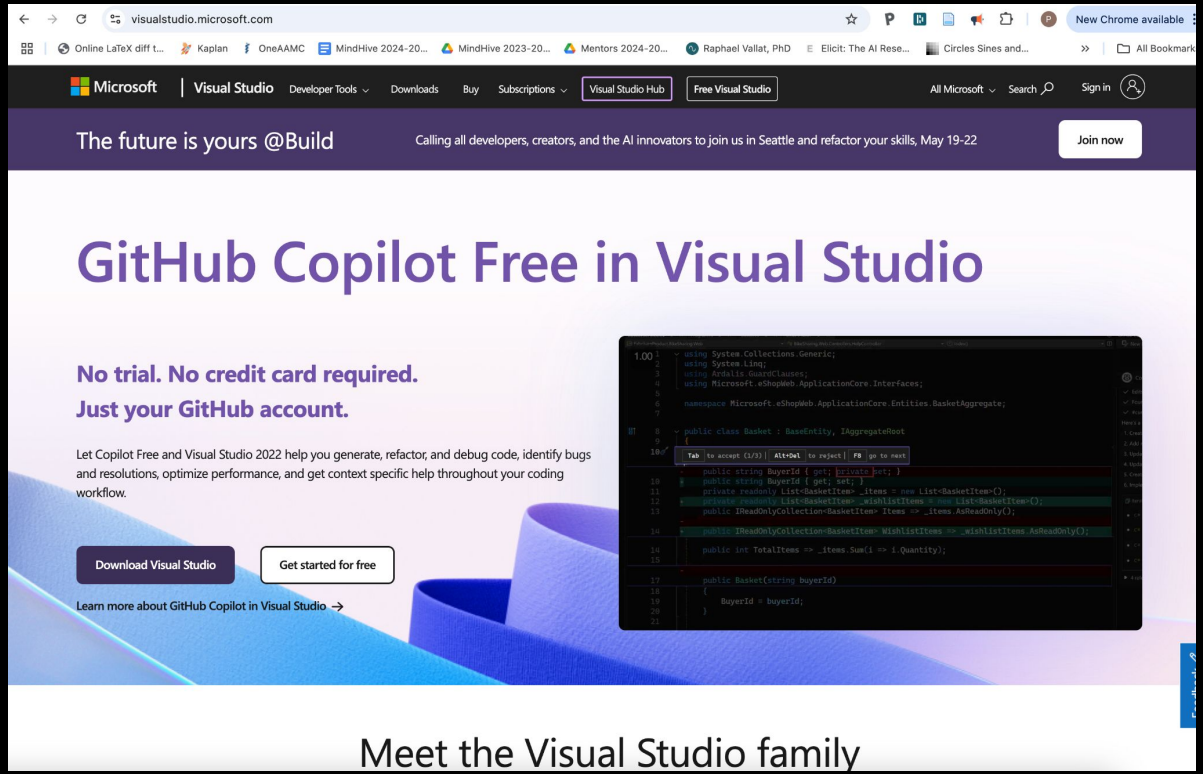
Python and Anaconda



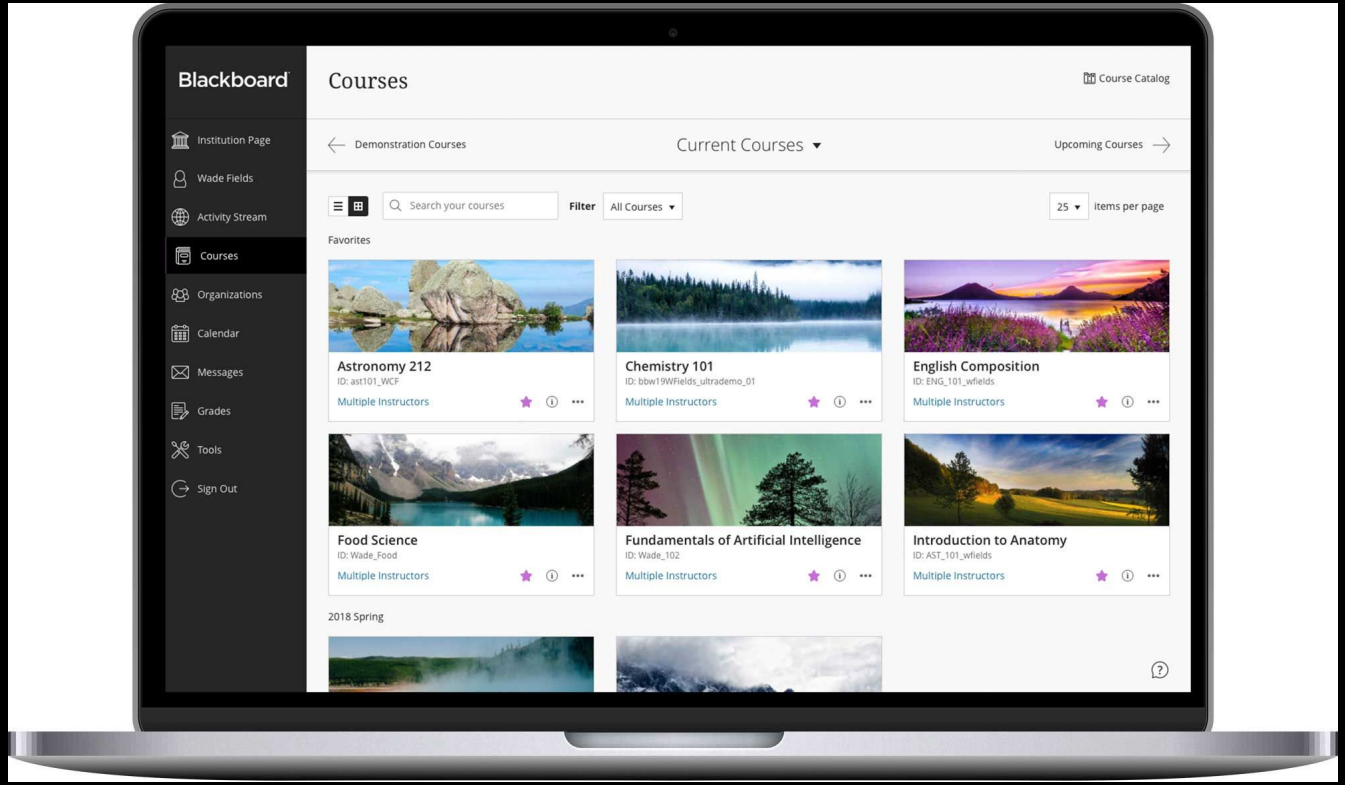
GitHub (and GitHub classroom)



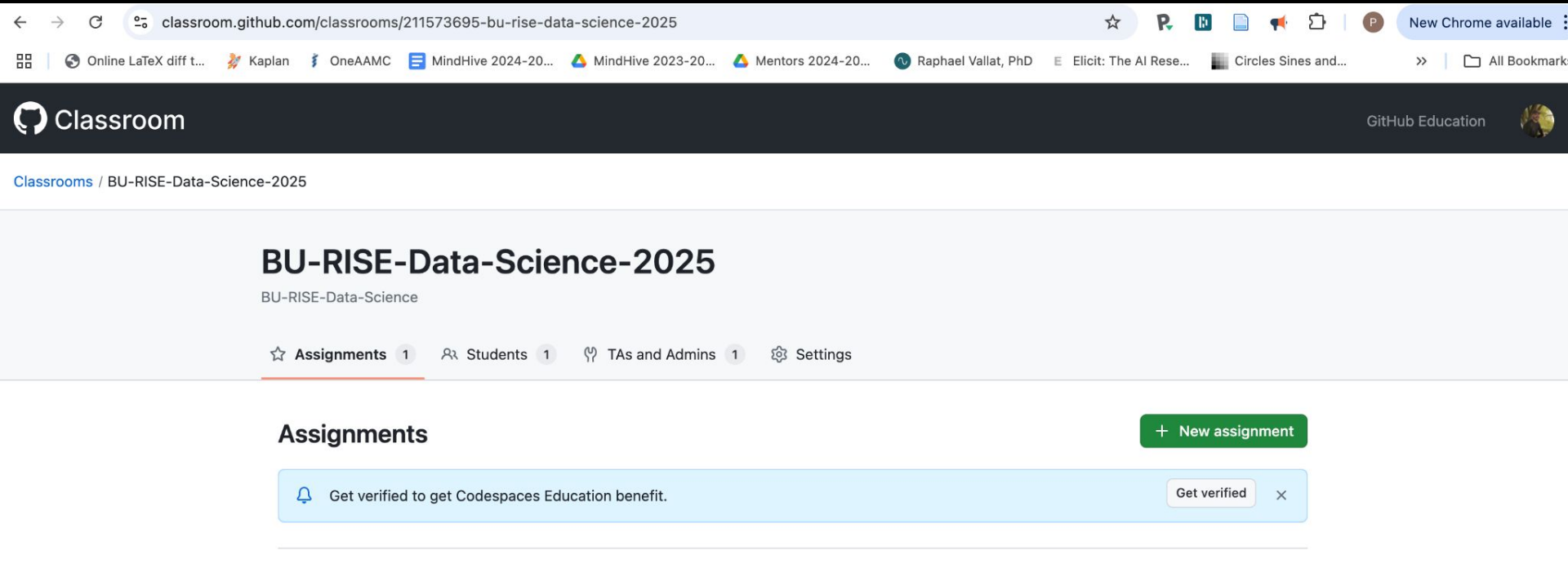
VSCode (Visual Studio)



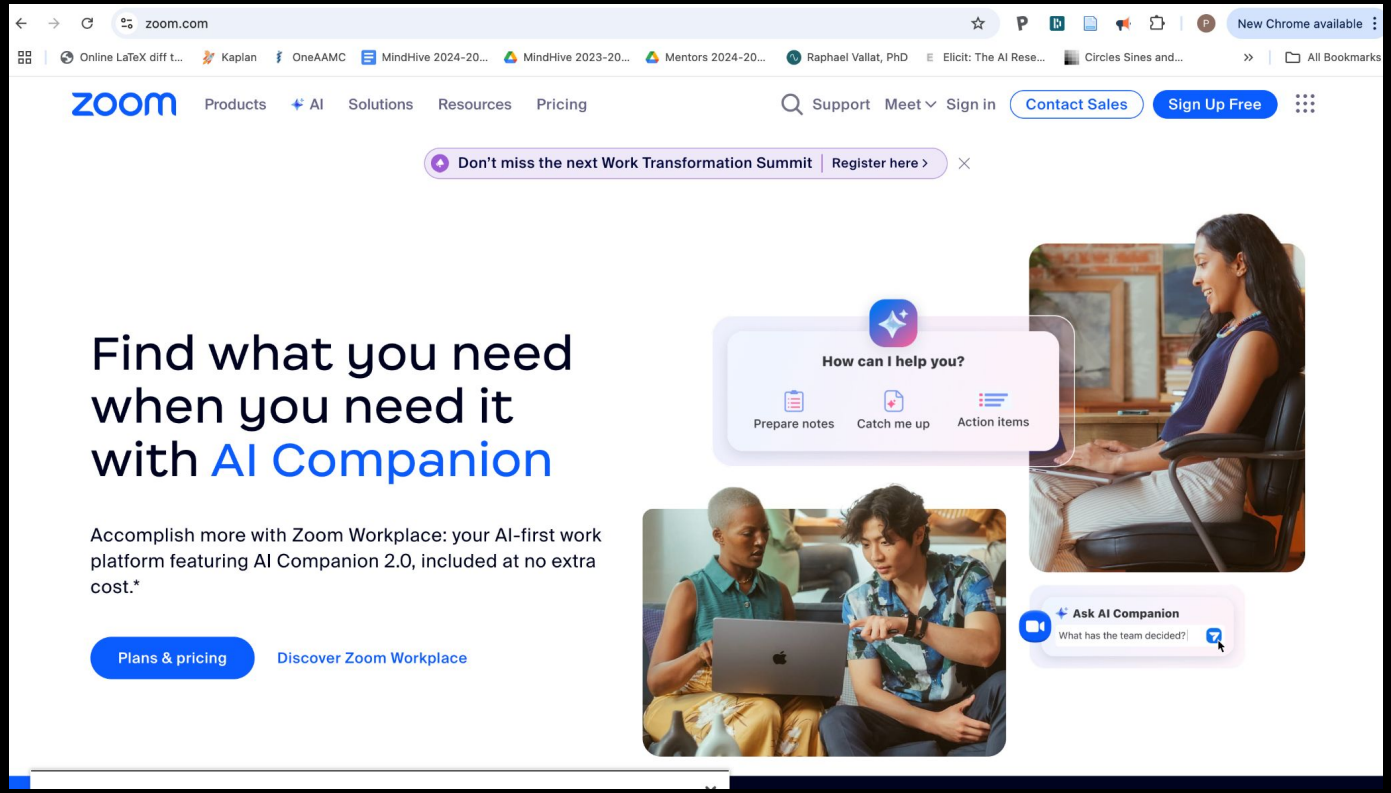
BlackBoard


Blackboard
Courses
Course Catalog
Institution Page
Wade Fields
Activity Stream
Courses
Organizations
Calendar
Messages
Grades
Tools
Sign Out
Demonstration Courses
Current Courses
Upcoming Courses
Search your courses
Filter
All Courses
25
items per page
Favorites
Astronomy 212
ID: ast101_WCF
Multiple Instructors
Chemistry 101
ID: bbw19WFields_ultrademo_01
Multiple Instructors
English Composition
ID: ENG_101_wfields
Multiple Instructors
Food Science
ID: Wade_Food
Multiple Instructors
Fundamentals of Artificial Intelligence
ID: Wade_102
Multiple Instructors
Introduction to Anatomy
ID: AST_101_wfields
Multiple Instructors
2018 Spring

GitHub Classroom (auto-grade your assignments)



Zoom (via your BU account, install before tomorrow)



Start Introduction to Python Coding Activity 1 (Blackboard and Google Colab)