

# ECS 119: Data Processing Pipelines

Fall 2025

# Welcome! (Am I in the right place?)

- This is a **required course** for data science majors
- Open as a CS non-major elective (100-119)
- Cap: 90, 12 currently waitlisted
- 4 units

# Basic Details

- **Instructor:** Caleb Stanford
- **TA:** Muhammad Hassnain
- **CRN:** 29022
- **Units:** 4
- **Lectures:** Monday, Wednesday, Friday 3:10-4pm in Walker Hall 1330
- **Discussion section:** Wednesdays at 11am in Young Hall 194
- **Office hours:** See Piazza
- **Final exam:** Thursday, December 11, 8:00am.

Syllabus + schedule + all of the above information is on [the class repository](#)

# About the Instructor

Started at Davis July 2023

Research: programming languages for systems applications  
(distributed systems, data processing systems, etc.)



DavisPL Research Group



[Website](#)

# Class TA

[Muhammad Hassnain](#)

Discussion section:  
Wednesdays at 11am in  
Young Hall 194



# Plan for today

1. Course introduction
2. Syllabus and logistics
3. Q+A

# Plan for today

1. Course introduction
2. Syllabus and logistics
3. Q+A

What is this class about?

# ECS 119: Data Processing Pipelines

[Home](#) • [Schedules and Classes](#) • [ECS 119: Data Processing Pipelines](#)

**Subject**

ECS 119

**Title**

Data Processing Pipelines

**Status**

Active

**Units**

4.0





# Discussion Question

You have compiled a spreadsheet of **website traffic data** for various popular websites (Google, Instagram, chatGPT, Reddit, Wikipedia, etc.). You have a **dataset of user sessions**, each together with time spent, login sessions, and click-through rates. You want to put together an app which identifies trends in website popularity, duration of user visits, and popular website categories over time.

1. Which of the following **tools** you would like to have access to to use to build your app and why? Python, R, Excel, MySQL, C++, Other
2. Can you estimate how many **hours** will it take to build your app? How many **computers**?
3. Suppose the data consisted of **10K (10,000) sessions, 10M (10,000,000) sessions, or 10B (10,000,000,000) sessions**. How would your answers to the above change in each of these scenarios?

## Sharing Your Answers

[tinyurl.com/3xw52mft](https://tinyurl.com/3xw52mft)

(requires login)

*Discussion questions are for completion only. There are no "right" answers.*



A common theme

**Software components, design constraints, tools and development techniques** differ depending on the size of the dataset...

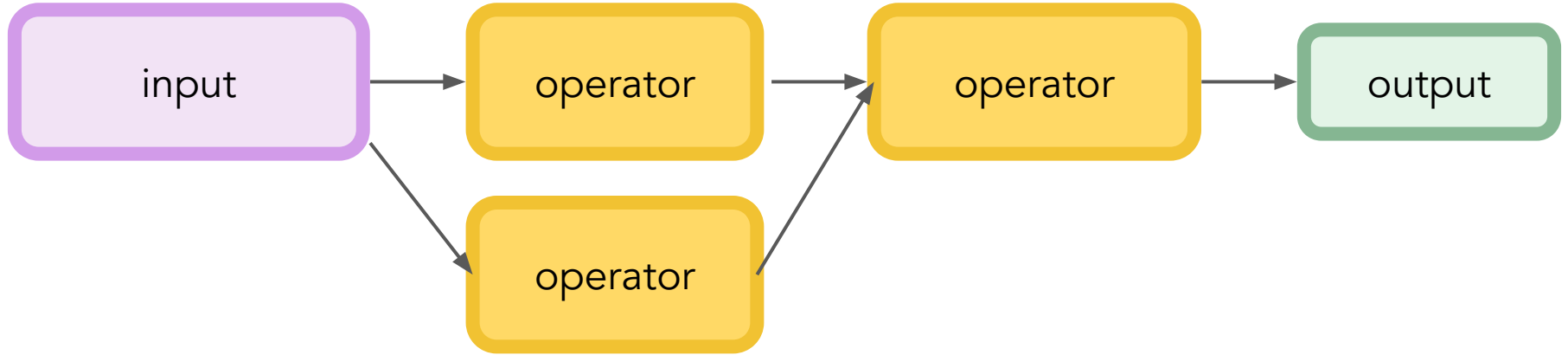
- ... 10K datapoints
- ... 10M datapoints
- ... 10B datapoints



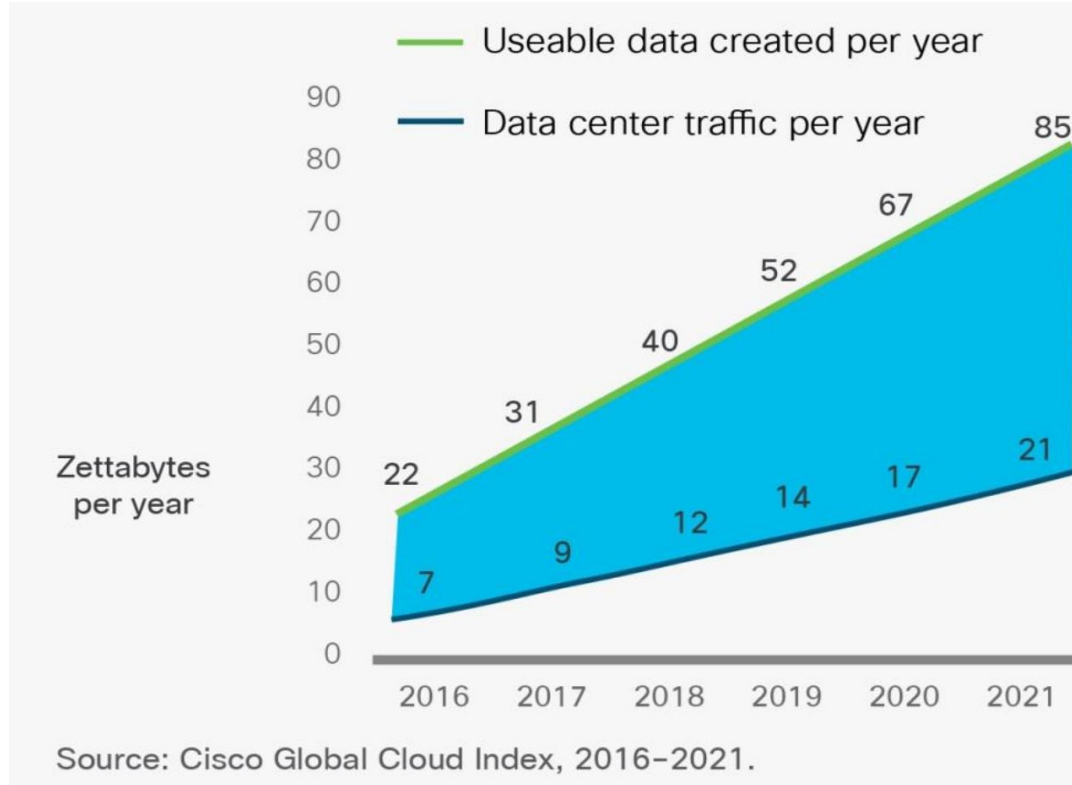
# Building applications for large datasets?



# Parallel computing



# Distributed and real-time computing





More data  
produced  
than can be  
feasibly  
stored

# Distributed computing – failures

## How did hackers compromise my EC2 instance?

Asked 8 years, 6 months ago   Modified 2 years, 8 months ago   Viewed 8k times

- 
- 11
- My EC2 instance was hacked recently. It doesn't really matter as I'm just starting my website and there was no sensitive information on my server yet, but I do plan for there to be in the future. I am going to terminate the compromised server and set up another one in order to secure my website.
- 



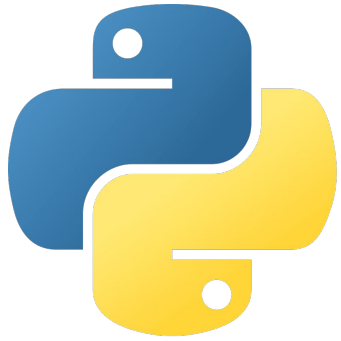
What this course is about



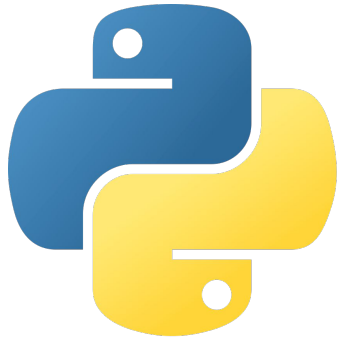
# Principles and practice of working with **large datasets**...



... using **real tools** used in industry



... using **real tools** used in industry



✨ Hands-on Approach ✨



# ! Disclaimer



Some things will be **experimental!**

Q: What are the prerequisites?

ECS 116 or 165A

Ability to program [FizzBuzz](#)

Q: Is this an AI/ML course?

Short answer: No

Long answer: No, but the tools and software used in this class are of fundamental importance for building, training, and deploying large ML models in industry

# Q: Is this course right for me as an elective?

**Short answer:** Yes in some cases

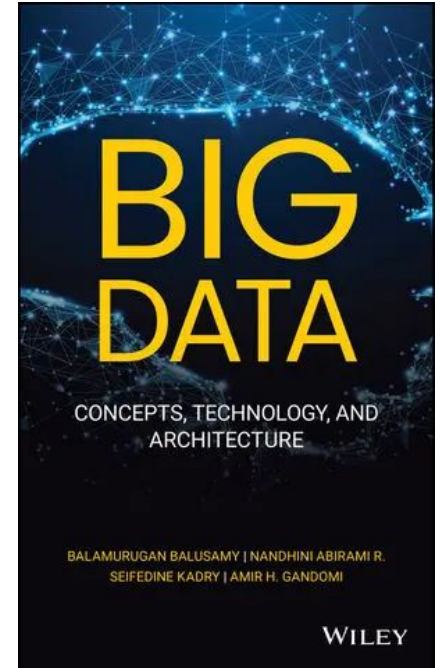
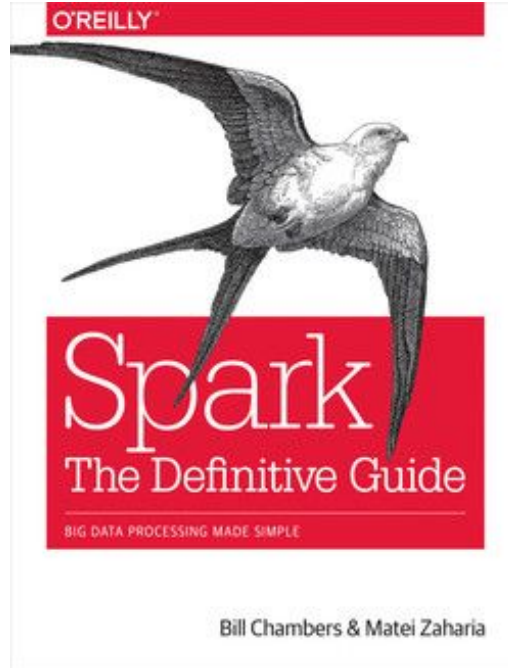
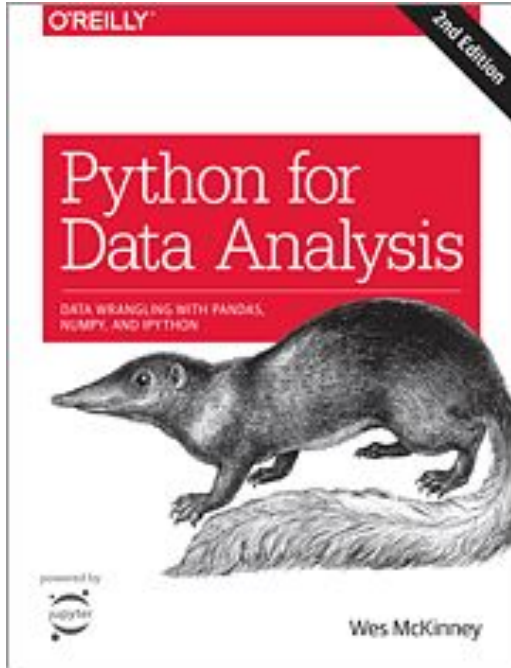
**Disclaimer:** Non-major CS class – does not count as an upper-division elective\*

- \* I am hoping to get this changed, however, this probably won't happen for this quarter

**Long answer:** The material will be relevant to you if:

- You want to know about the fundamental principles & practice behind large-scale data processing tasks
- You want to gain hands-on experience with the tools used in industry

## Textbooks (Optional)





# Learning objectives

See [Syllabus](#)

# Plan for today

1. Course introduction
2. Syllabus and logistics
3. Q+A

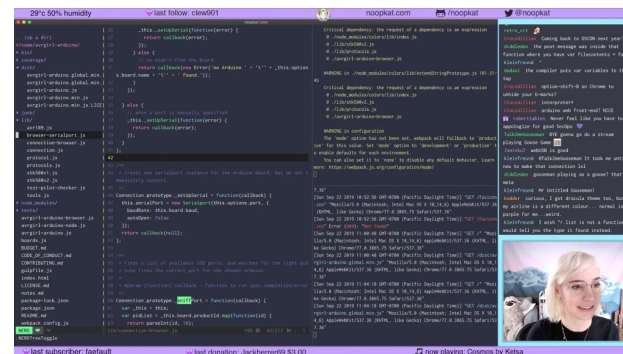
# Lectures

- I generally lecture with live coding (not slides)

- Programming can only be learned by **doing!**

Attendance: Lectures and discussion section are encouraged, but not required

- In-class polls can be made up at any time



# To follow along...

<https://github.com/DavisPL-Teaching/119>

```
> git clone git@github.com:DavisPL-Teaching/119.git
```

```
> git pull
```

# Piazza

Please join the [Piazza](#)

# Grade breakdown

- **Participation (10%)**: via in-class quizzes
- **Homeworks (35%)**: I plan to do 3 assignments, plus homework 0
- **Midterm (20%)**: covering the first half and main concepts of the course
- **Final Exam (35%)**: covering all topics covered in the course.

**Minimum** grade cutoffs: 93%=guaranteed A, 90%=guaranteed A-, etc.

**Exams may be curved to a lower maximum score**

# Attendance and participation (10%)

Fill out the in-class polls (participation points only)

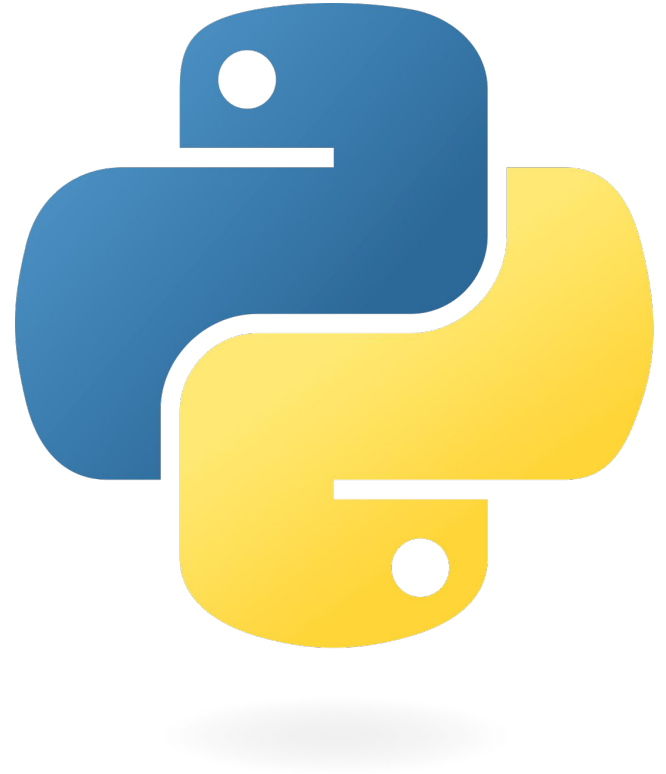
**If you are sick:** Starting from the first day of class, you may join the class remotely via Zoom (the quality may not be as good)

**If you miss class:** Lectures are recorded. You can make up the in-class polls at any time



# Homeworks (35%)

We will be working in Python for most of the course



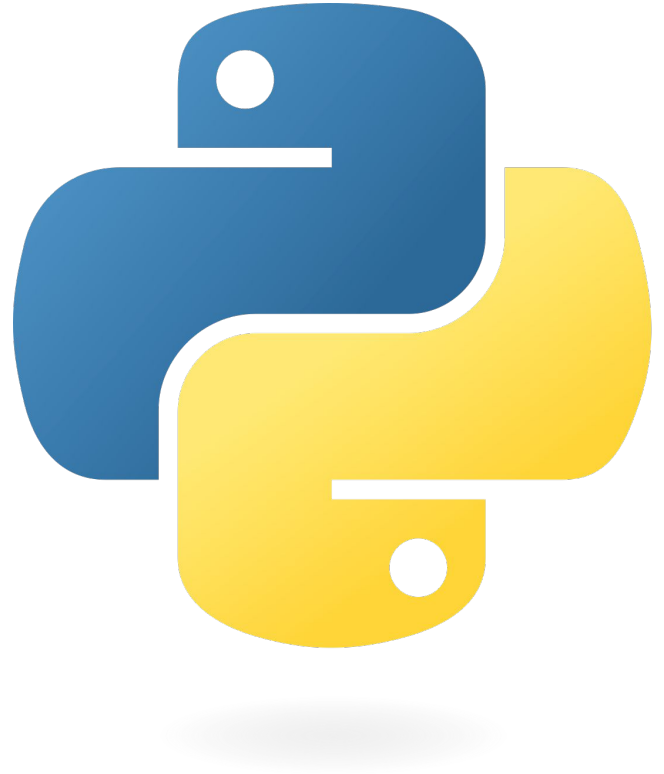


# Homeworks (35%)

We will be working in Python for most of the course

Rough plan for HWs:

- HW1 (Pandas): Data wrangling, data cleaning, and data validation
- HW2 (PySpark): Parallel processing
- HW3: Advanced topics & distributed processing



# Homework 0: installation help



(submission link posted soon, probably Friday)



# Homework grading

Most important: please run your code

- Software engineering is about running software!
- Experienced engineers will run and test their code frequently
- We cannot give points to code that does not run 😞

# Exams

- There will be one midterm (20%) and one final exam (35%)
- I allow one-sided cheat sheets, practice questions given (The goal is for you to pass!)

# Platforms

Class discussion, Q+A, and announcements:

- [ECS 119 on Piazza](#)
- Don't email me, post to Piazza!
- Make your post public and (if you prefer) anonymous

Homeworks and Exams: Gradescope

# AI Policy

**AI is a powerful tool! Please use it to help you (and not the other way around)**

- I do allow AI use on homework assignments. (I even encourage it! If you are using it responsibly)
- People in industry are using it – but an expert with an AI tool is much different than a non-expert with a tool
- [Advice from Jason Lowe-Power](#)
- Midterm and final exam will be in-class and closed-book

# Collaboration Policy

- Collaboration is encouraged!
- Everyone should submit their own solution
- Please list your collaborators at the top of your homework

# Schedule

## Tentative Schedule

**Please note:** there will be **no class on Wednesday, October 1** (1 week from today) as I will be away at a conference.



# Communication reminders

TA: **Hassnain**

Office hours: See Piazza

Please use Piazza for questions (not email)

# Other reminders

Job scams:

- [Job Scams | Career Center](#)
- [UC Davis Phish Bowl](#)

# Respect and discrimination

✨ **Please be nice!** ✨

Include everyone in group discussions

Reach out to me in case of any problems

# Waitlist

We currently have: 90 registered, 12 waitlisted

## **Please note:**

- Waitlist adds are done by the department and graduate advising (I am not allowed to issue PTAs)
- Priority will be given to students who met the prerequisite requirement

You are welcome to keep attending lectures until we know – I am sorry that I cannot guarantee a spot!

If you are in the first 5-10 spots and you attend consistently there is a good chance that you could get in.

# Plan for today

1. Course introduction
2. Syllabus and logistics
3. Q+A

Questions for me?

Reminder

Please join the [Piazza](#)





# Rough topic list...

- Basics of data processing
- Software engineering tools
- Input data sources
- Parallelism and concurrency
- Distributed computing
- Real-time and streaming data processing
- Cloud computing tools

# Learning objectives

- Use Python and other scripting tools to manage and manipulate data on a single machine.
- Understand the components, techniques, tools, and performance metrics of setting up data processing jobs in Python.
- Understand the concepts of parallelism, types of parallelism, and parallelization mechanisms, via tools like MapReduce, Hadoop and Spark.
- Understand how software engineering tools and configuration are integrated into a data project, via tools like Git and the shell and other orchestration.
- Understand the concepts of distributed computing and distributed data processing, including distributed consistency requirements, and how it manifests in real-world applications.
- Understand advanced topics including programming over real-time and streaming data sources and using cloud platforms such as AWS, Azure, and Google Cloud.