https://hao-ai-lab.github.io/dsc204a-f25/

# DSC 204A: Scalable Data Systems
# Fall 2025


Staff
Instructor: Hao Zhang
TAs: Mingjia Huo, Yuxuan Zhang

@haozhangml @haoailab
haozhang@ucsd.edu

# Instructor

Hao Zhang (https://cseweb.ucsd.edu/~haozhang/)

- Ph.D. from CMU CS, 2020

- Projects: Parameter server, auto-parallelization

- Took 4-year leave to work for a "not-so-successful" startup (raised 100M+), 2016-2021

- Projects: Petuum, MLOps

- Then postdoc at UC Berkeley working on LLM+systems, 2021 – 2023

- Projects: vLLM, Vicuna, lmsys.org, Chatbot Arena

- Then co-founded a small startup and acquired by SNOW and started at UCSD

# My Lab: https://hao-ai-lab.github.io/

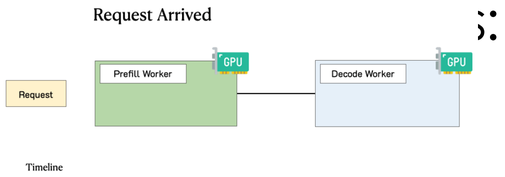
Research Area: Machine Learning + Systems

Recent topics (some will be covered in the final part of this course):

- Fast LLM Inference and Serving
- Large-scale distributed ML systems, Model parallelism, etc.
- Open source LLMs, data curation, evaluation

I also work for snowflake for 20% of my time (which is relevant to this course)

Sor :




DistServe

vLLM

vllm.ai

Starred 58.8k

FAST VIDEO

Starred 2.3k

# Today

**What is This Course and Why Study It**

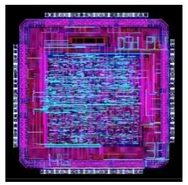
Course overview

Logistics

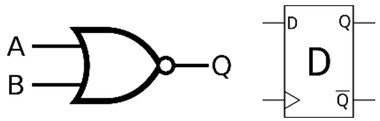
Warm up (If time permits)

# What is this course about: **data-centric system** course

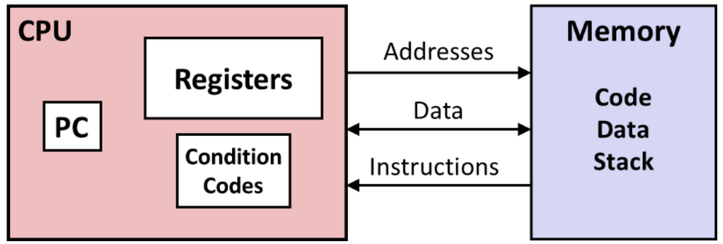
## Computer Designer



**Gates, clocks, circuit layout, …**



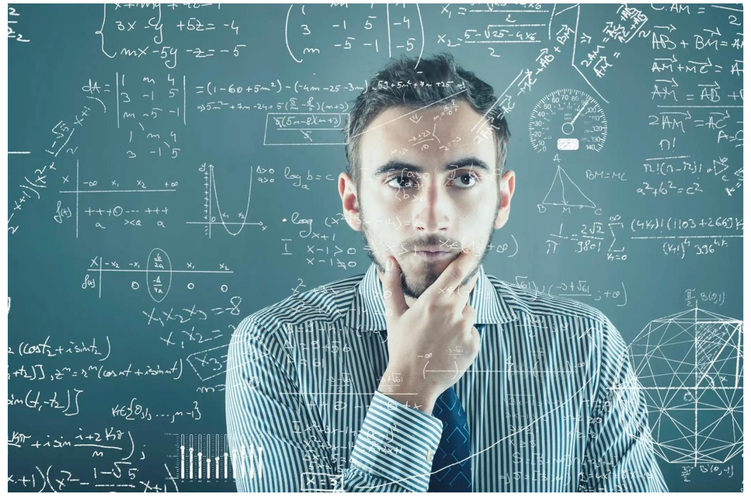

## Assembly programmer




## C programmer

```
#include <stdio.h>
int main(){
int i, n = 10, t1 = 0, t2 = 1, nxt;
for (i = 1; i <= n; ++i){
printf("%d, ", t1);
nxt = t1 + t2;
t1 = t2;
t2 = nxt; }
return 0; }
```

## Data science



# What is this course about: data

DATA

How to store and access the data?

- Computer Organizations
- OS
- Databases
- Data encoding

# What is this course about: drawing values from data

BIG

DATA

How to store and access **big** data?

- Cloud
- Distributed storage
- Parallelisms, partitioning
- Networking

# One classic example: Dataframe API

# What is this course about: access and process big data


How to access and process big data?

- Distributed computing
- Batch and stream processors, dataflow systems, programming models
- Big data tools: Hadoop, Spark, Ray

# One Modern example: LLMs

AI: new ways of drawing values from big data

LLMs: powerful AI that can scale with **data size**

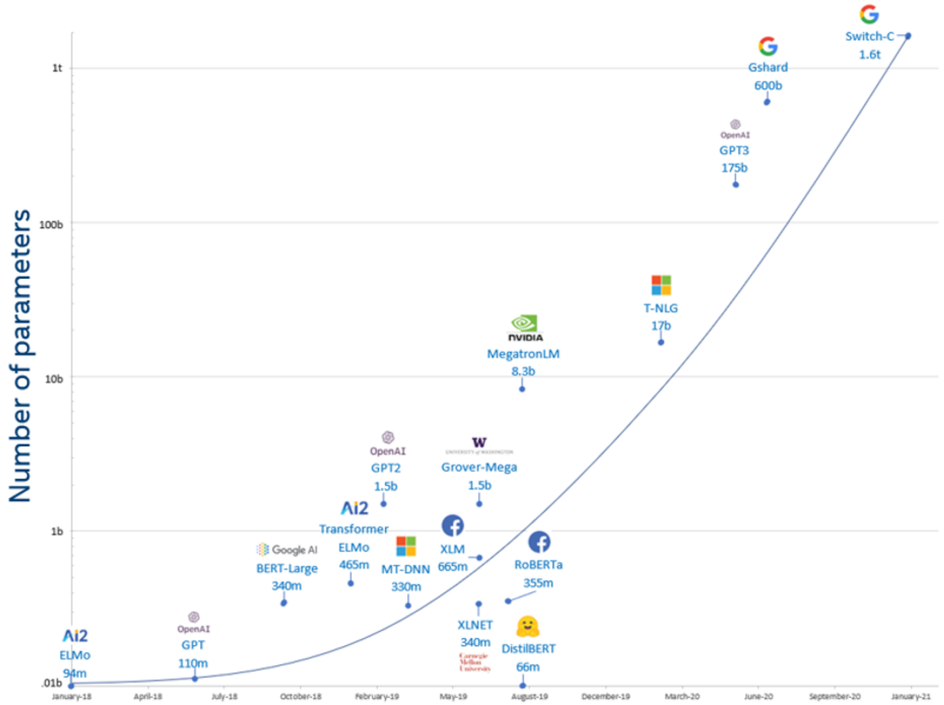



*Figure 1: Exponential growth of number of parameters in DL models*

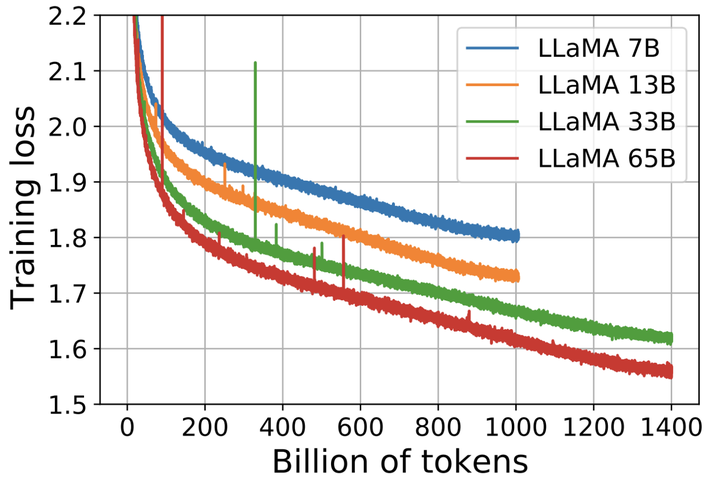



Figure 1: **Training loss over train tokens for the 7B, 13B, 33B, and 65 models.** LLaMA-33B and LLaMA-65B were trained on 1.4T tokens. The smaller models were trained on 1.0T tokens. All models are trained with a batch size of 4M tokens.

# What is this course about: drawing values from data

BIG DATA

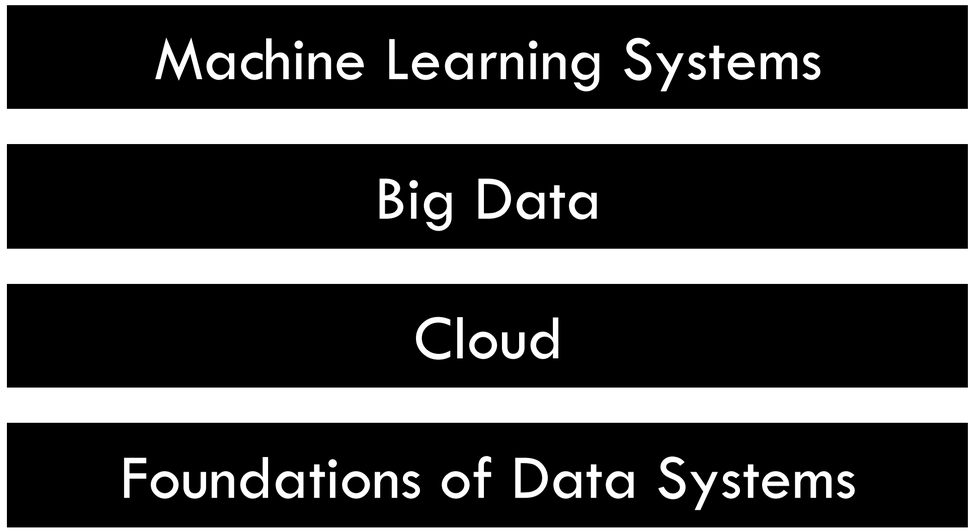
+AI

AI: New ways of drawing values from Big data

- ML frameworks, dataflow graphs
- Distributed ML systems, ML parallelisms
- Large language model systems

# Hence the course is organized into four parts

- Foundations of data systems: OS, storage, compute
- Cloud: Cloud storage, network, parallelism, etc.
- Big Data: data processing and programming
- ML systems: ML frameworks, parallelism, LLM training and serving



# What is this course about?

- Foundations of data systems
  - Data models, big data storage and retrieval, and how to encode information when you store data, etc.
  - ~~Transactions, synchronization, consistency, consensus~~

# What is this course about?

- Cloud and Distributed Systems
  - Cluster, cloud, network, replication, partition, consistency, etc.
  - ~~RPC, Caching, Fault tolerance, Paxos, Concurrency~~

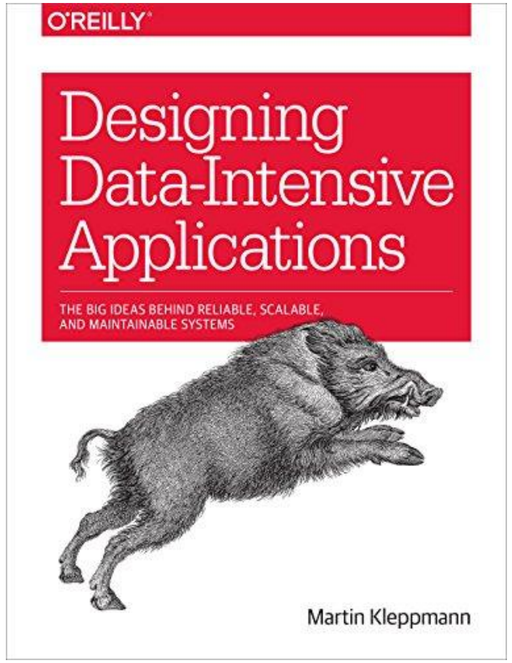
# What is this course about?

- Big Data Processing and Programming model
  - Batch processing, stream processing, MapReduce, Hadoop, Spark, Ray, etc.

# What is this course about?

- ML Systems
  - ML frameworks, dataflow graph representation of ML, ML parallelism, LLMs, LLM training and serving
  - ~~ML architecture details, learning algorithms/theory, optimizations, NLP~~

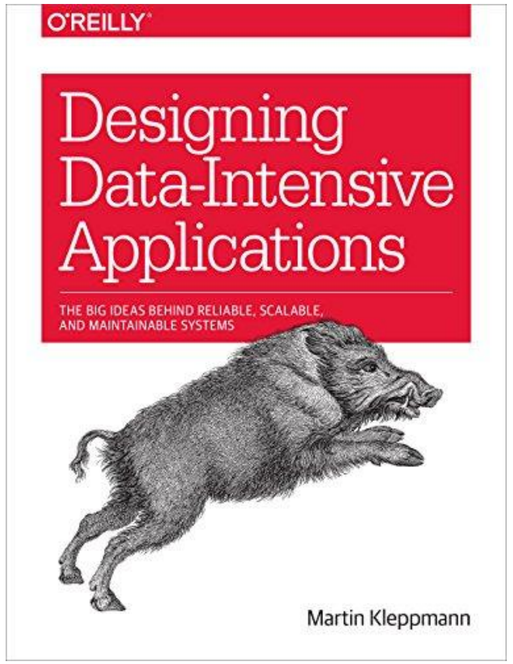
# Suggested Textbooks

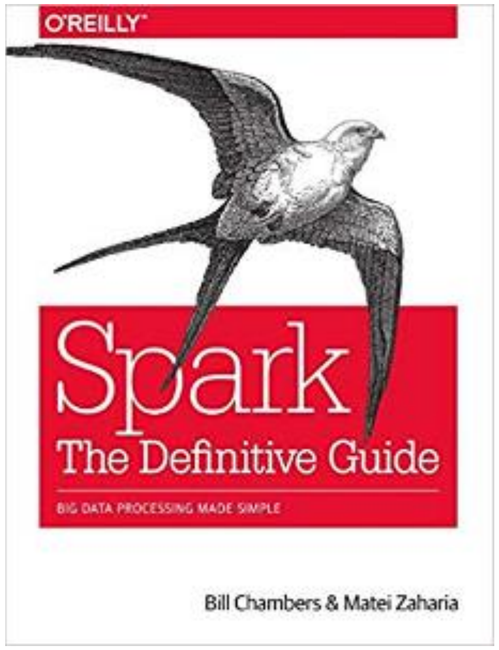



- Chapter 3. Storage and retrieval
- Chapter 4. Encoding and evolution
- Chapter 10. Batch processing
- Chapter 11. Stream processing
- Chapter 12. The future of data systems

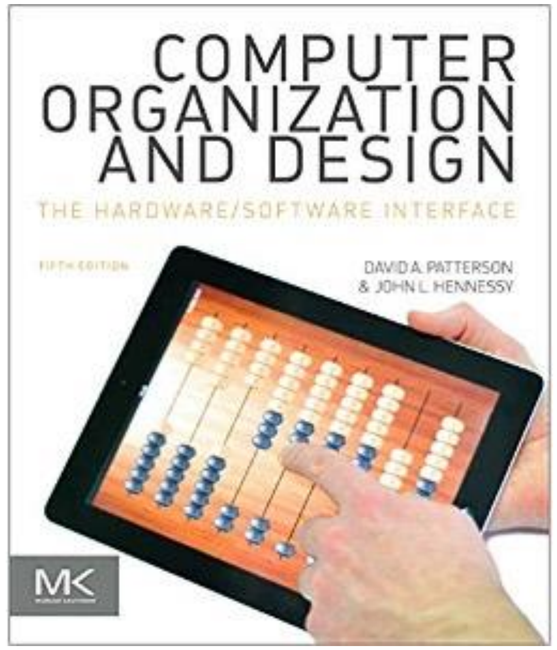
# Suggested Textbooks

Computer systems are about carefully layering levels of abstraction.

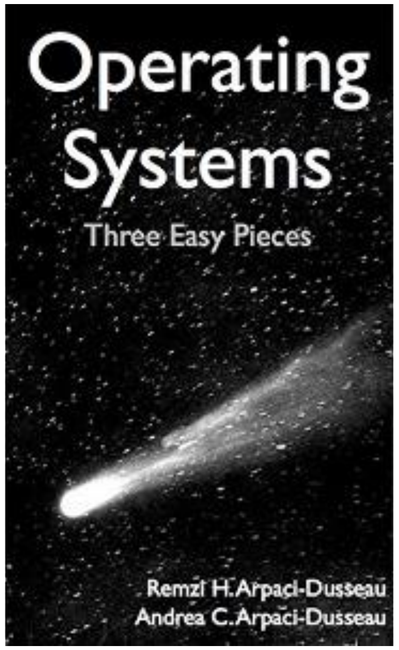



Scalable data flows




Low-level system software



# Learning outcomes of this course

- Explain the basic principles of data systems, distributed systems, and data programming model.
- Identify the abstract data access patterns of, and opportunities for parallelism and efficiency gains in data processing at scale.
- Gain hands-on experience in creating end-to-end pipelines for data preparation, feature engineering, and distributed model training.
- Reason critically about practical tradeoffs between accuracy, runtimes, scalability, usability, and total cost.
- **Enter** the current trends of Big data + Big Models

# What this course is NOT about

- Not a course on database, relational model, or SQL
  - Take DSC 202 instead (pre-requisite)
- Not a course on how to build scalable data systems
  - Take Distributed Systems, Operating Systems, Cloud Computing, …
- Not a training module for how to use Spark or PyTorch
  - We focus more on principles.
- Not a machine learning course
  - We focus more on system and data
- Not a machine learning system course
  - Take my CSE/DSC 291: deep learning systems in 26 Spring.
  - But could be a warm-up

# **Delta** of this year's offering by Hao

- The pace will be faster: less basics, more advanced stuffs
  - Take DSC 202 or DSC102 instead if you expect more basics (pre-requisite)
- More new stuffs, less classic stuff: ~1/4 will be about new systems developed between 2016 – 2024
  - Data + ML systems: PyTorch, Ray
  - Machine learning parallelism
  - LLM systems
- Homework will be based on Ray and vLLM
- No mid-term, no in-class quiz
- More offline paper readings, scribe notes

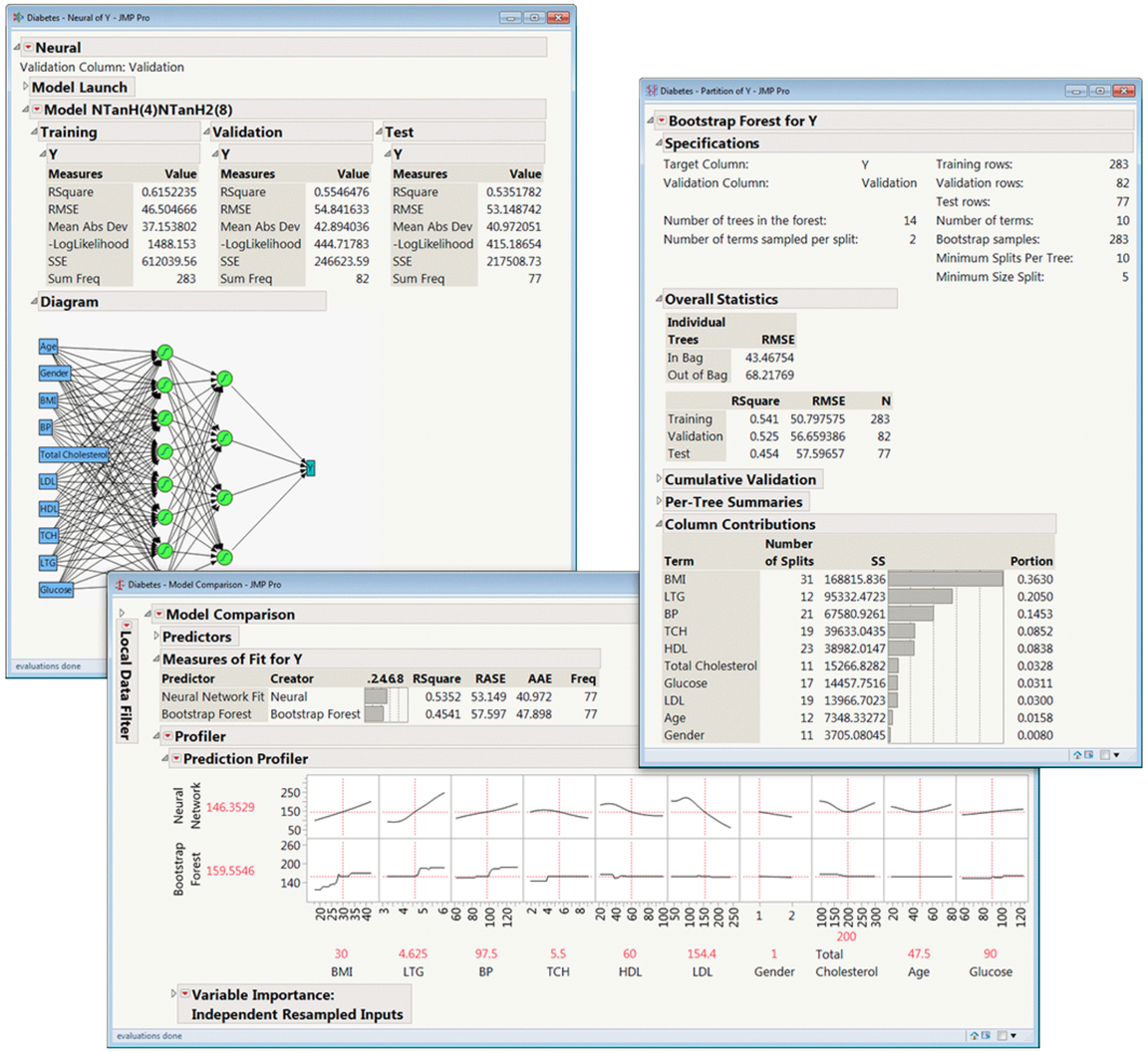
Why bother learning such low-level system-related stuff in Data Science?

# I will Provide 2 Arguments

1. Operating Large, distributed systems is an essential skill today
2. The tech world is scaling and accelerating…
3. You might be able to make more money if you know how to deal with distributed systems ☺

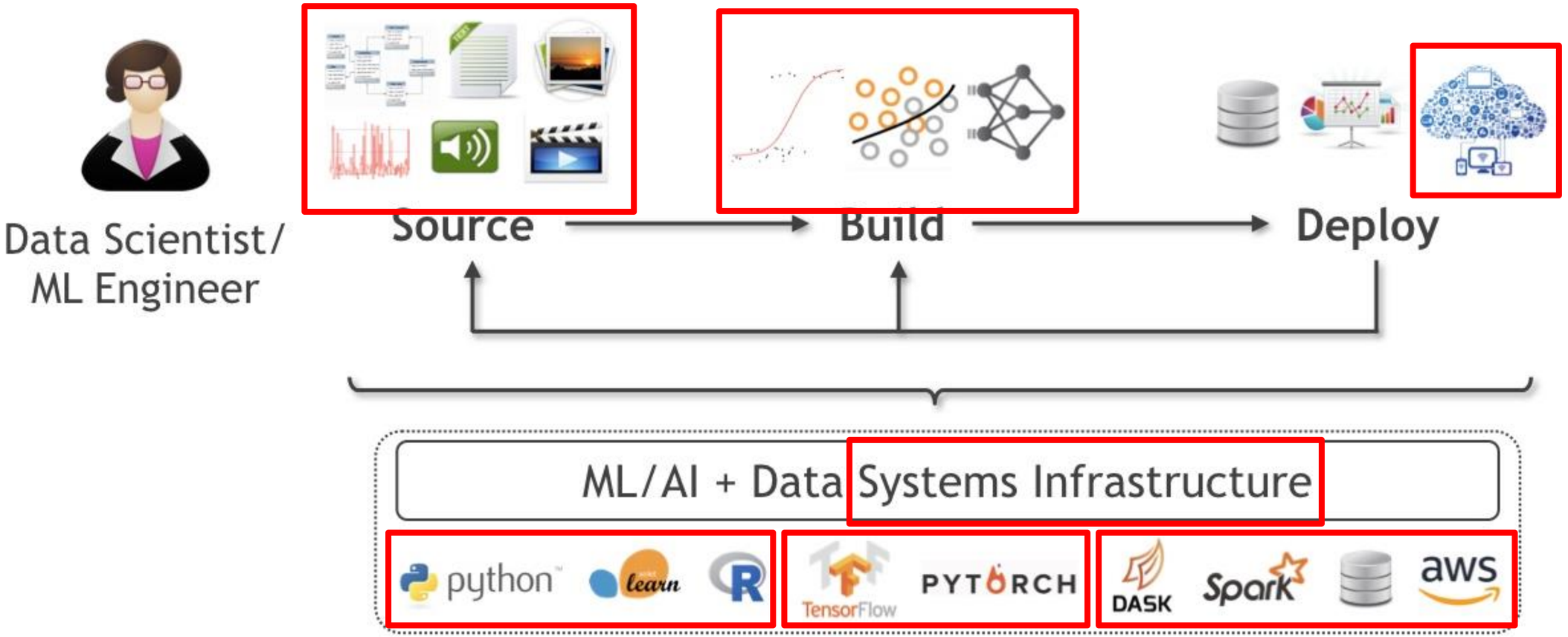
# “Statisticians”/“Analysts” 20 years ago

- Methods: Sufficed to learn just math/stats, maybe some SQL
- Types: Mostly tabular (relational), maybe some time series
- Scale: Mostly small (KBs to few GBs)
- Tools: Simple GUIs for both analysis and deployment; maybe an R-like console




https://www.jmp.com/en_au/offers/jmp-pro-for-academic-research.html
https://www.technologymagazine.com/data-and-data-analytics/sas-tops-worldwide-advanced-and-predictive-analytics-market-share

# In the era of 2020s:

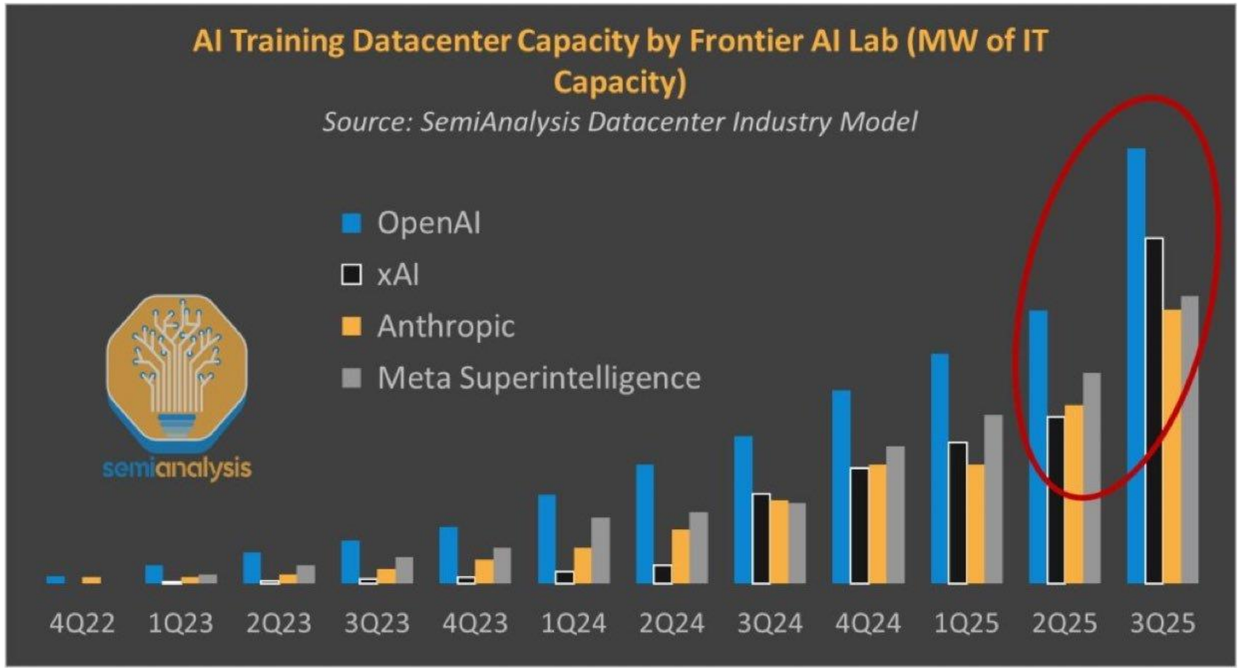


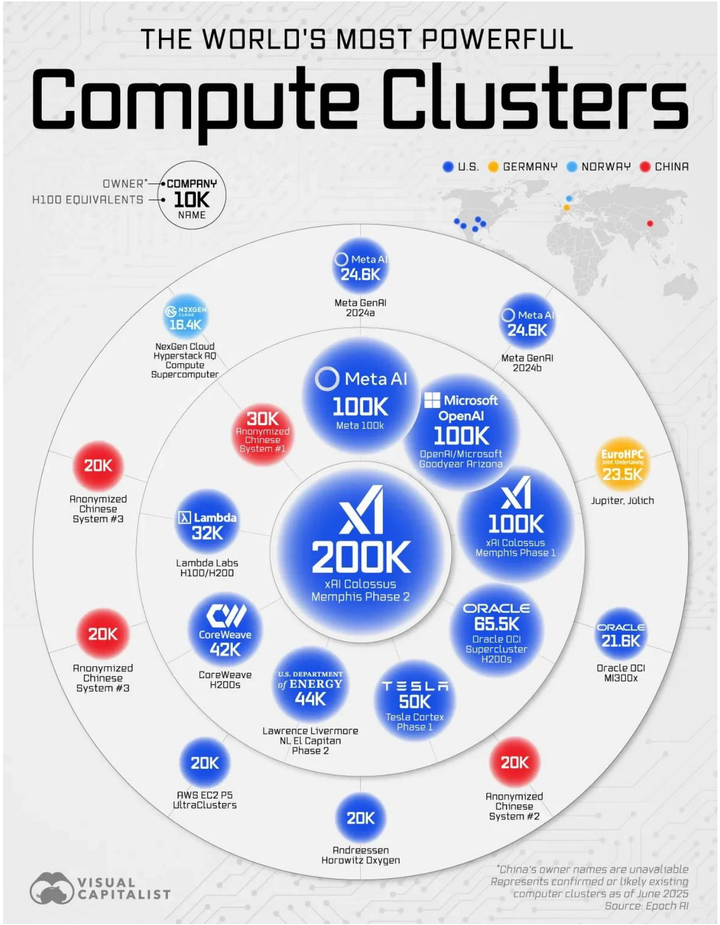

Data acquisition
Data preparation

Feature Engineering
Training & Inference
Model Selection

Serving
Monitoring

# The Entire Tech World Now is About Scaling




Q: what skills are most needed to scale on the software side?

glassdoor

statistician | Location

# Statistician Salaries United States

Overview | **Salaries** | Interviews | Insights | Career Path

## How much does a Statistician make?

Updated Jan 4, 2022

Industry: All industries

Employer Size: All company sizes

Experience: All years of Experience

Very High Confidence

**$88,989** /yr

**Average Base Pay**

2,398 salaries

$61K Low | $89K Average | $131K High

glassdoor

data scientist | Location

# Data Scientist Salaries United States

Overview | **Salaries** | Interviews | Insights | Career Path

## How much does a Data Scientist make?

Updated Jan 4, 2022

Industry: All industries

Employer Size: All company sizes

Experience: All years of Experience

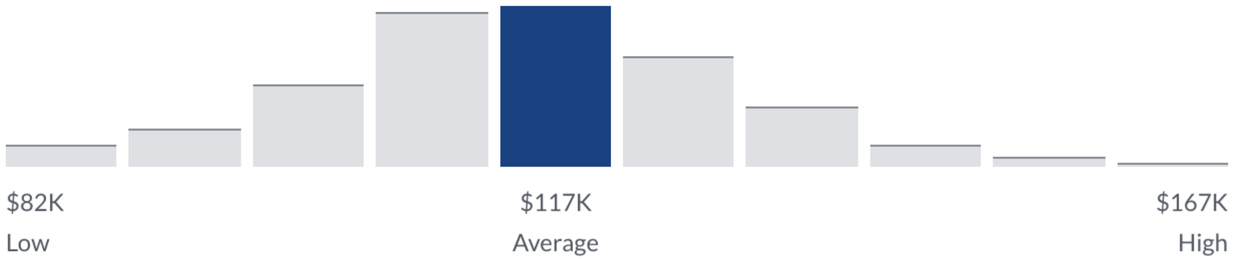
To filter salaries for Data Scientist, Sign In or Register.

Very High Confidence

**$117,212** /yr

Average Base Pay

18,354 salaries




— $88,989

= $28,223!

GLASSDOOR

Community Jobs Companies Salaries Search

# How much does an AI Engineer make?

Updated Dec 13, 2023

Experience: All years of Experience

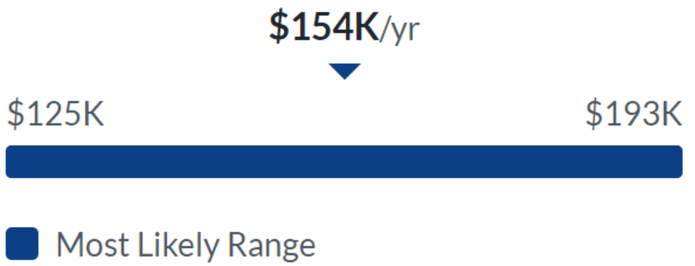
Industry: All industries

Confident

Total Pay Range **$125K - $193K**/yr

Base Pay **$104K - $156K**/yr

Additional Pay **$20K - $38K**/yr




## Total Pay Trajectory

For Machine Learning Engineer

- **$152,007 /yr**
  **Machine Learning Engineer**
- $172,167 /yr
  Senior Machine Learning Engineer
- $165,994 /yr
  Lead Machine Learning Engineer

See Full Career Path >

Download as data table

— $88,989

= $65011!

# OpenAI

Work Here? Claim Your Company



## OpenAI Software Engineer Salaries

Software Engineer compensation at OpenAI ranges from $570K per year for L4 to $915K per year for L5. The median compensation package totals $925K. View the base salary, stock, and bonus breakdowns for OpenAI's total compensation packages. Last updated: 1/7/2024

### Average Compensation By Level

+ Add Comp | Compare Levels

| Level Name | Total | Base | Stock (/yr) | Bonus |
|---|---|---|---|---|
| L3 (Entry Level) | US$ -- | US$ -- | US$ -- | US$ -- |
| L4 | US$570K | US$245K | US$325K | US$0 |
| L5 | US$914.5K | US$302K | US$612.5K | US$0 |
| L6 | US$ -- | US$ -- | US$ -- | US$ -- |

# Another Perspective

The fastest growing companies in SV is either data or AI companies: they operate either big data or big models.

## Fastest-growing data companies

snowflake

databricks

## Fastest-growing model companies

OpenAI

ANTHROP\C

Questions?

# Prerequisites

- DSC 200, 202 (or equivalent).
- Proficiency in Python programming & Unix Terminals
- Network and Operation System basics
- Deep learning basics: pytorch, tensorflow,
- For all other cases, email me with proper justification; a waiver can be considered (I normally approve all students)

# Components and Grading

- 3 Programming Assignments: 44% (12% + 16% + 16%)
  - In total 5 late days! Plan your work well ahead.
- No Midterm (cheers!)
- Final Exam (06/14/2023 3pm-6pm): 36%
- Scribe Duties: 8%
- Reading summary: **12%**
- Extra Credit: 5%

# Grading Scheme (grade is the better of the two)

| Grade | Absolute Cutoff (>=) | Relative Bin (Use strictest) |
|---|---|---|
| A+ | 95 | Highest 5% |
| A | 90 | Next 10% (5-15) |
| A- | 85 | Next 15% (15-30) |
| B+ | 80 | Next 15% (30-45) |
| B | 75 | Next 15% (45-60) |
| B- | 70 | Next 15% (60-75) |
| C+ | 65 | Next 5% (75-80) |
| C | 60 | Next 5% (80-85) |
| C- | 55 | Next 5% (85-90) |
| D | 50 | Next 5% (90-95) |
| F | < 50 | Lowest 5% |

# Grading Scheme (grade is the better of the two)

| Grade | Absolute Cutoff (>=) | Relative Bin (Use strictest) |
|---|---|---|
| A+ | 95 | Highest 5% |
| A | 90 | Next 10% (5-15) |
| A- | 85 | Next 15% (15-30) |
| B+ | 80 | Next 15% (30-45) |
| B | 75 | Next 15% (45-60) |
| B- | 70 | Next 15% (60-75) |
| C+ | 65 | Next 5% (75-80) |
| C | 60 | Next 5% (80-85) |
| C- | 55 | Next 5% (85-90) |
| D | 50 | Next 5% (90-95) |
| F | < 50 | Lowest 5% |

Example, 82 and 33%,

Rel: B-; Abs: B+;

Final: B+

# The structure of the course

| | Topics | |
|---|---|---|
| Week 1-2 | Foundations of Data Systems | Single Machine: CompOrg, OS, Storage |
| Week 3-5 | Cloud | Cloud: Storage, network, parallelism, etc. |
| Week 6-8 | Big Data | Big Data Processing, dataflow, Programming models |
| Week 8-10 | Machine Learning Systems | MLSys: GPUs, ML libs, ML parallelism, LLM training/serving |

🌍 https://hao-ai-lab.github.io/dsc204a-f25/

# Programming Assignments

Three PAs

Will be based on Ray

- Good to study and try Ray from today if you have zero experience

Topics: exploring distributed data exploration, processing, and distributed ML

Most of the PAs should be doable using your laptop

- However, if you have trouble (due to hardware issue), please contact TAs

# Expectations on the PAs

- Expectations on the PAs:
  - Individual projects; see webpage on academic integrity
- TAs will explain and demo the tools; handle all Q&A
- You are expected to put in the effort to learn the details of the tools' APIs using their documentation on your own!

- In short: if you want to learn something solid, do the PAs
- PAs will be the most challenging part of this course

# Scribe Duties

Sign up your scribe duty here:

https://docs.google.com/spreadsheets/d/1NawbzzFapaUqaaldwgHx3CVxjRZyWxeq94F40N-pF-Y/edit?gid=0#gid=0

You should

- Scribe with as many details as possible
- Collaborate with other scribers
- Submit PRs to course website repo
- Reviewed and maybe iterated with the TA

# Exams

- No Mid-term
- In-person Final exam (36%)
- All MCQs (select one and all that apply)
- You can bring as many books/cheat sheets/paper you want
- No phone/laptop/Internet/ChatGPT
- Data: TBD

# Exams

Hao's lectures will feature some MCQs (that may appear in final exams) every week, so make sure to attend lectures or watch recordings.

TAs will give special recitations for preparing finals to help you navigate

# MCQ Example: Who originally developed PyTorch?



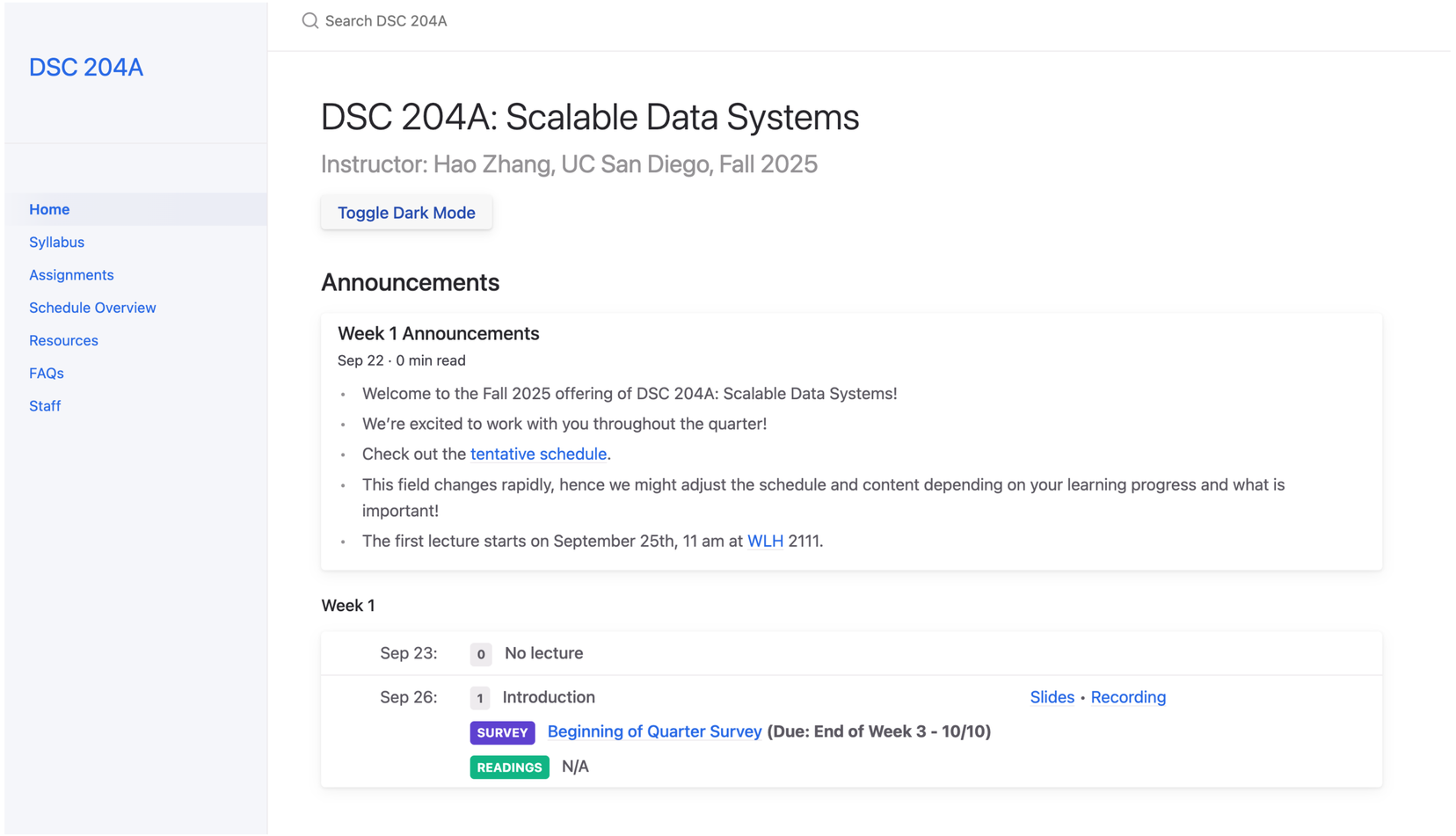

# Karma Points

- Participation: lectures / piazza
- Guest lecture: ask hard questions to challenge our guests ☺
- Completing course surveys and evaluation: it helps me, helps TAs and help yourself

# Respecting TAs' time

- Use piazza first, seeking helps from your peers
- Students answering questions on Piazza will be rewarded
- Office hours are for getting ideas on how to debug or better approach your homework.
- Write a description! Try to narrow down your problem area as much as possible.
- If you don't have a description, TA can reject your questions.
- Respect TA's working hours.
  - Respond in 24 hours.
  - Members may send msgs at night or on weekends, but only expect to receive a reply on weekday.

# Course website




🌍 https://hao-ai-lab.github.io/dsc204a-f25/

# General Dos and Do NOTs

- Do:
  - Follow all announcements on Piazza
  - Try to join the lectures/discussions live
  - Participate in discussions in class / on Piazza
  - Raise your hand before speaking
  - View/review podcast videos asynchronously by yourself
  - To contact me/TAs, use piazza first; if you really need to email, use "DSC 204A:" as subject prefix
  - Use LLMs to help your learning

# General Dos and Do NOTs

- Do NOT:
  - Harass, intimidate, or intentionally talk over others
  - Violate academic integrity on the PAs, exams, or other components; I (and the school) am very strict on this matter!

# TODOs after Today's lecture

1. Make sure you are enrolled with Piazza, Canvas, Gradesope
2. Check all contents of course website (Schedule, Syllabus, Exam time)
3. Signup your scribe duty
4. Finish Start-of-quarter survey
5. Start the reading of week 2 (which is due on Wed of week 4)

Questions?

# Warmup: History of Compute and Data

- ~= History of "which is the most valuable company in tech"