https://hao-ai-lab.github.io/dsc204a-w24/

DSC 204A: Scalable Data Systems
Winter 2024

Hao Zhang

@haozhangml

haozhang@ucsd.edu

Bio

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

Now: Asst. Prof @ HDSI, Affiliated with CSE, UCSD

- Ph.D. from CMU CS, 2020

- Project: Parameter servers, Data parallel ML, etc.

- Took 4-year leave to work for a startup (raised 100M+), 2016-2021

- Project: Petuum

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

- Project: Alpa, vLLM, Vicuna, lmsys.org

My Lab: Hao AI Lab

Research Area: Machine Learning + Systems

Recent topics:

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

Some ongoing projects:

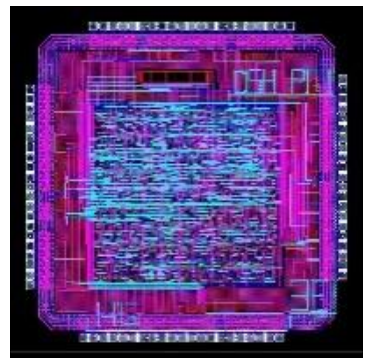
Alpa alpa.ai

lmsys.org

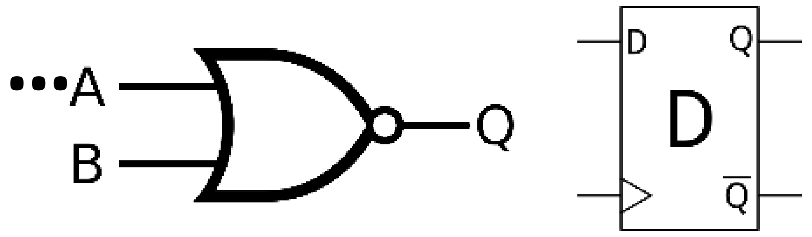
vLLM vllm.ai

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

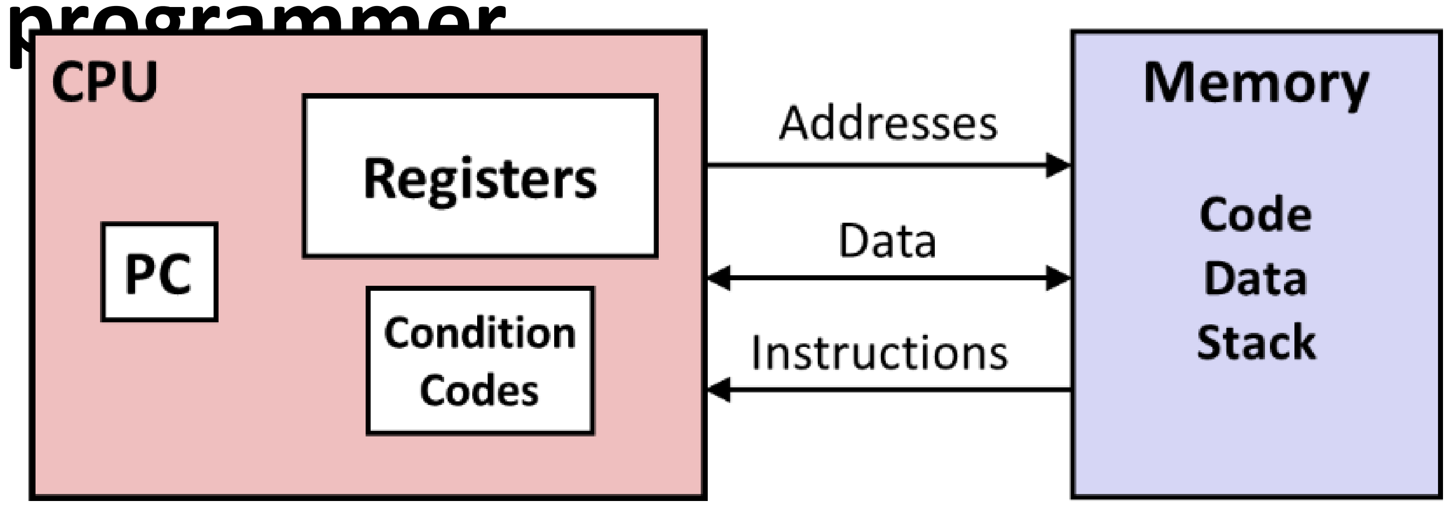
Computer Designer



Gates, clocks, circuit layout,



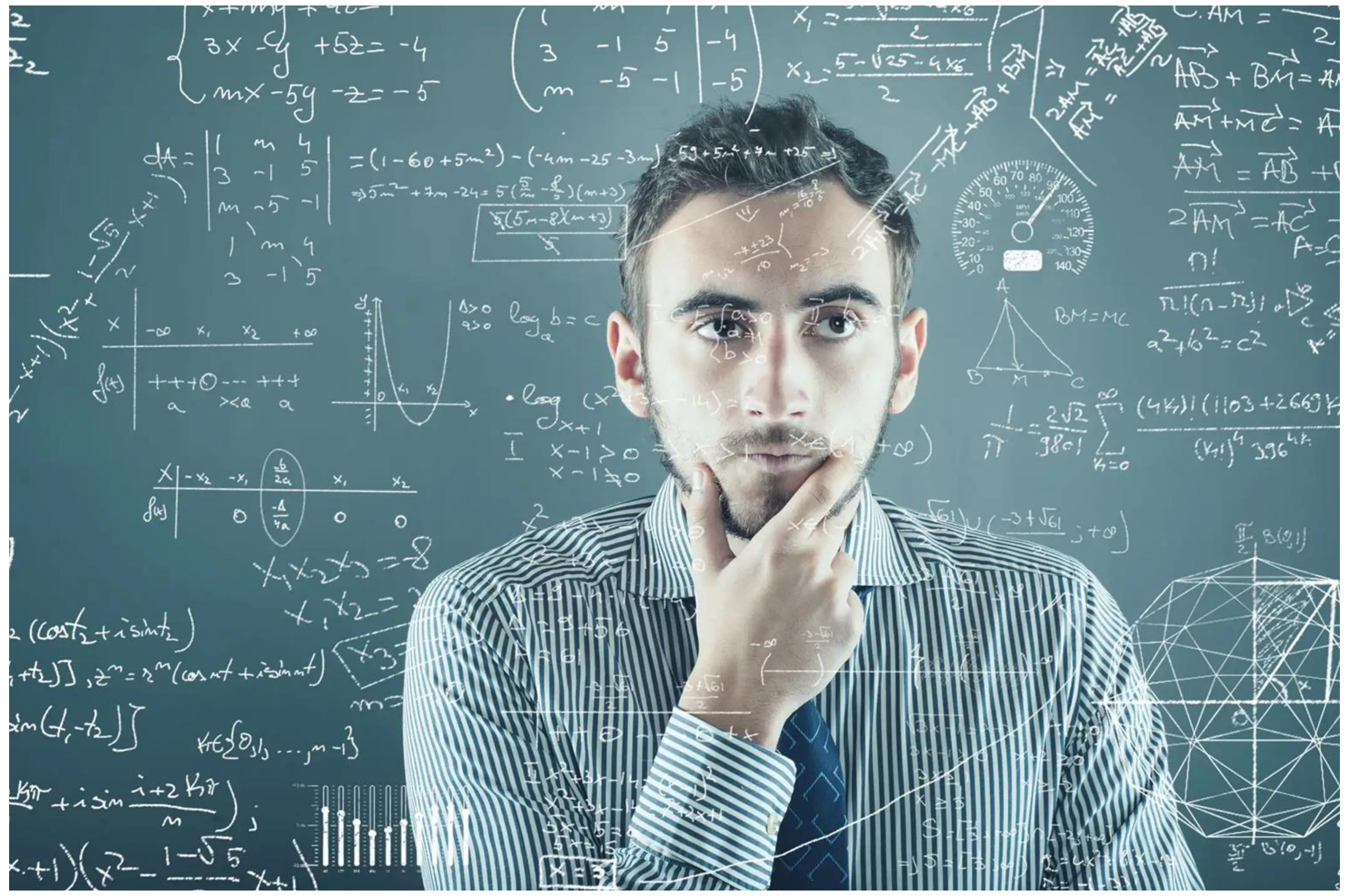
Assembly programmer



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

EVER
Company's 1000-table database on data lake with 100k attributes
DataFrame API

What is this course about: access and process big data

How to 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

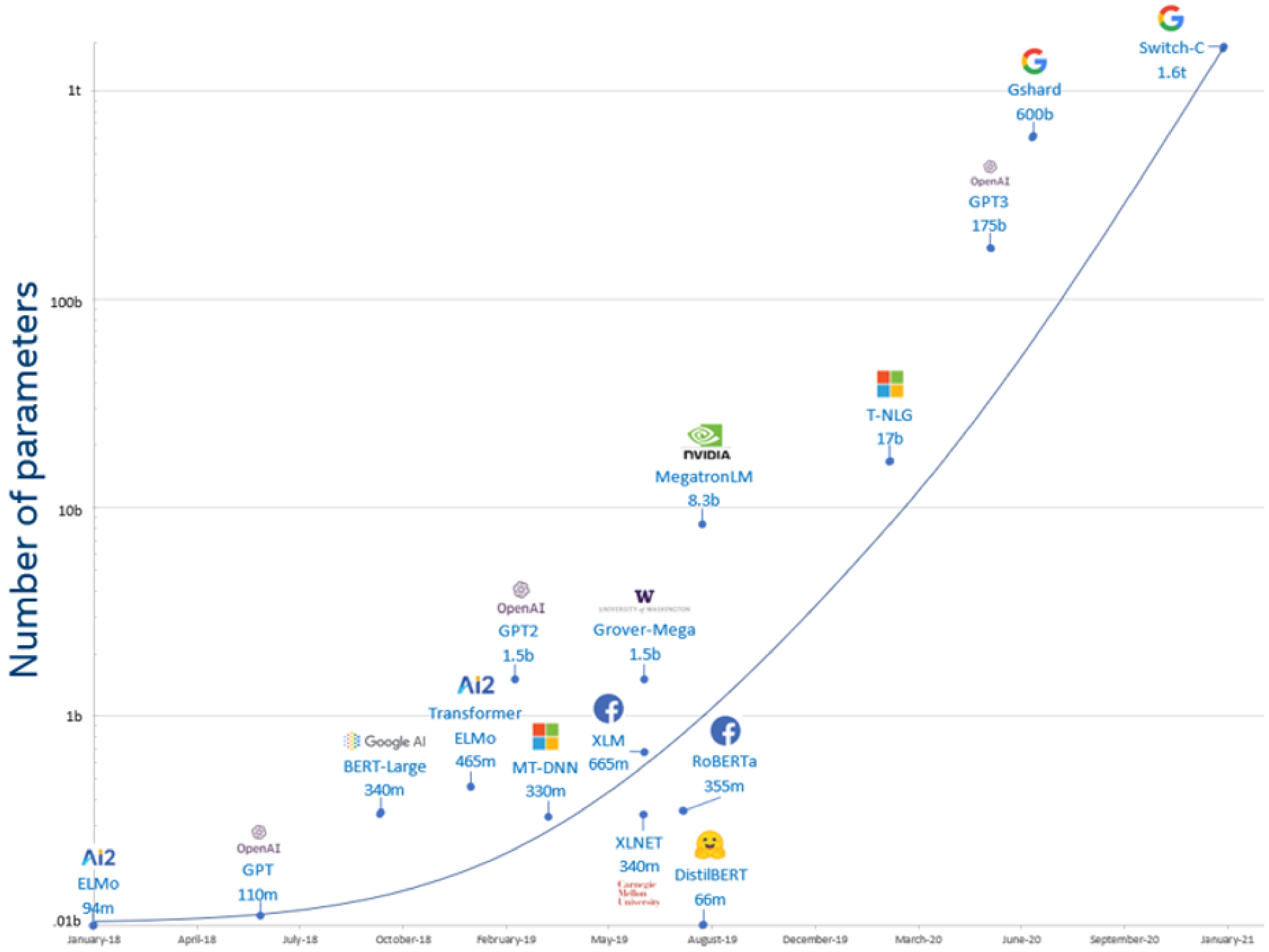


Figure 1: Exponential growth of number of parameters in DL models

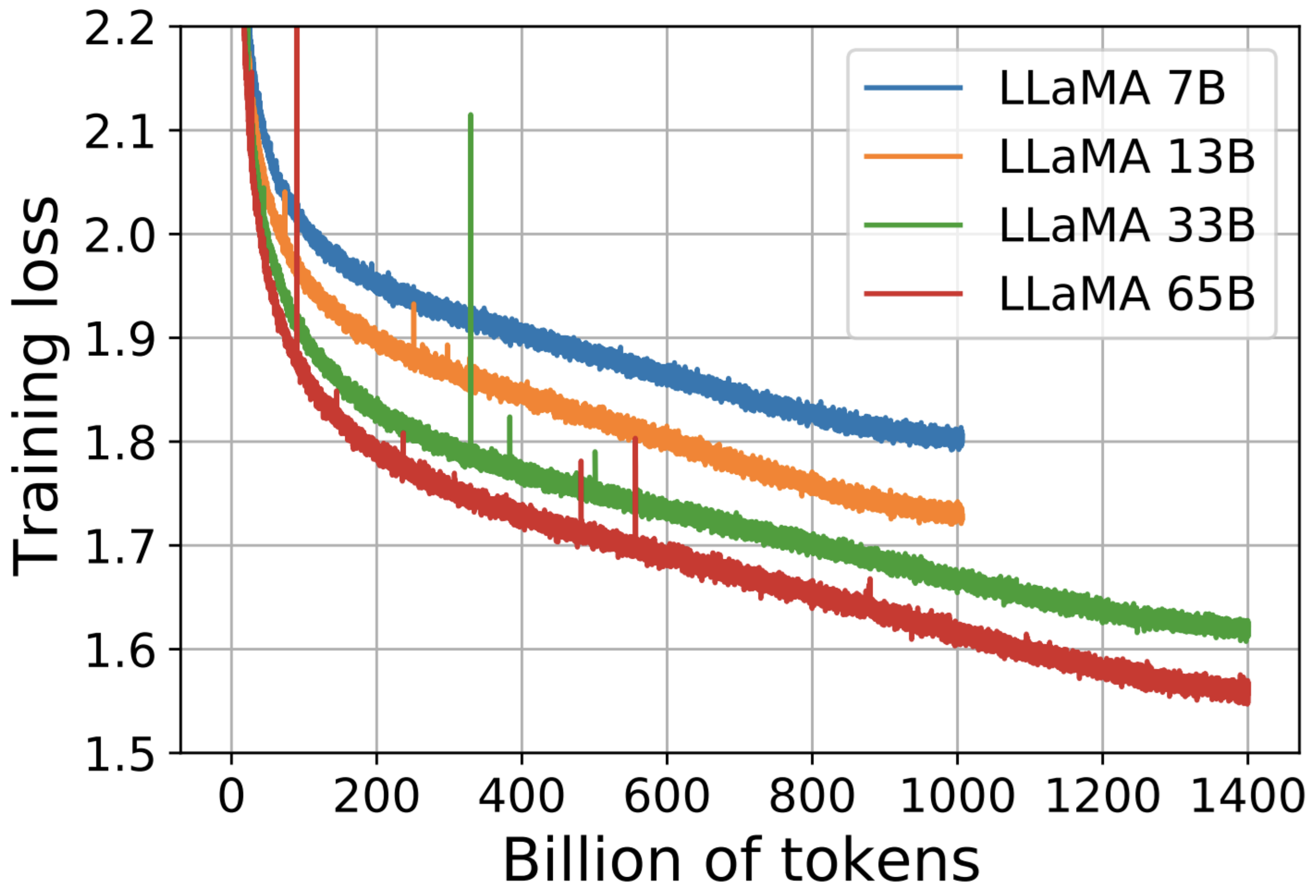


Figure 1: **Training loss over train tokens for the 7B, 13B, 33B, and 65 models.** LLaMA-33B and LLaMA-65B were trained on 1.4T tokens. The smaller models were trained on 1.0T tokens. All models are trained with a batch size of 4M tokens.

What is this course about: drawing values from data

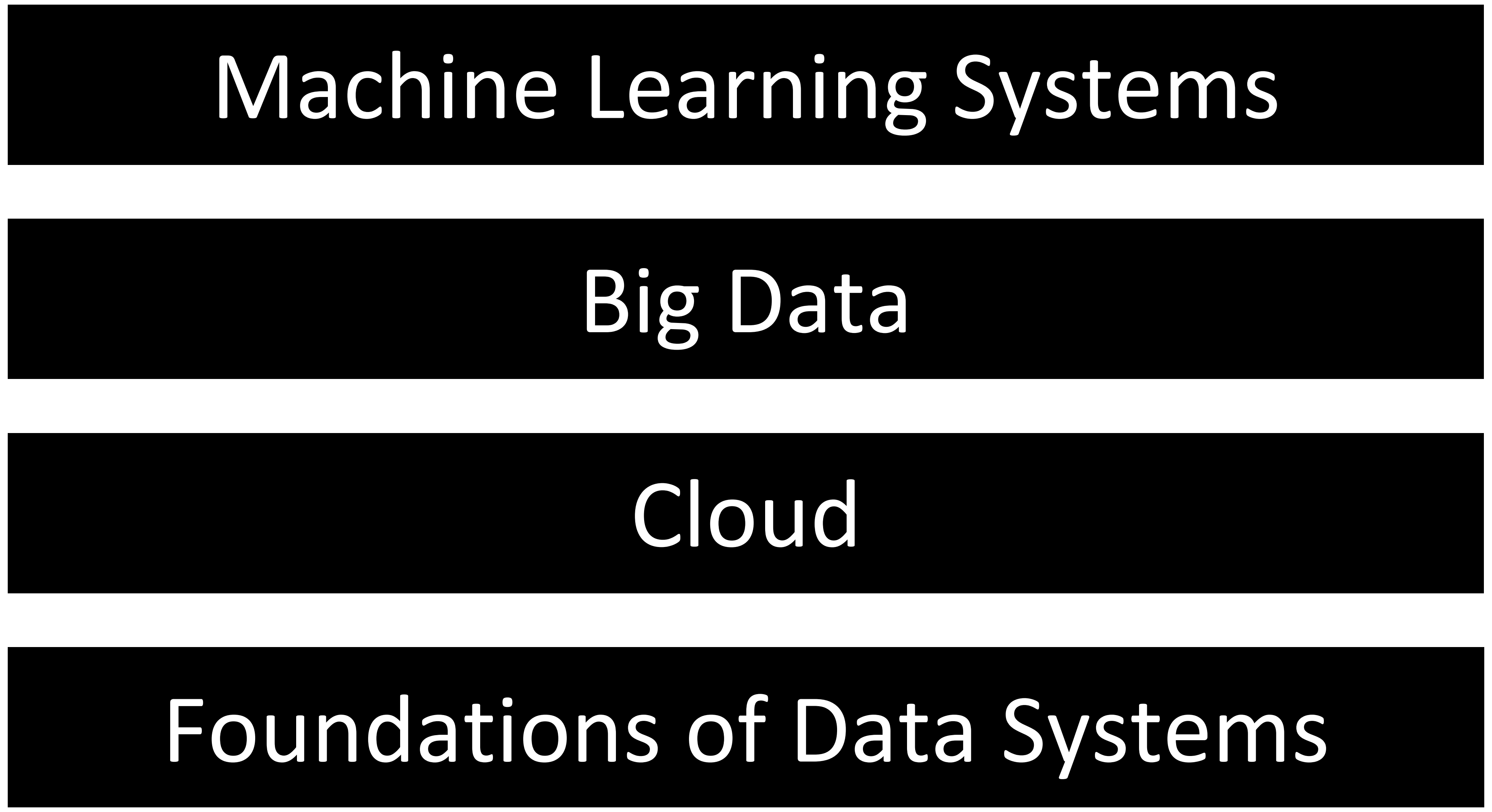
BIG
DATA+AI

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



Machine Learning Systems

Big Data

Cloud

Foundations of Data Systems

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~~

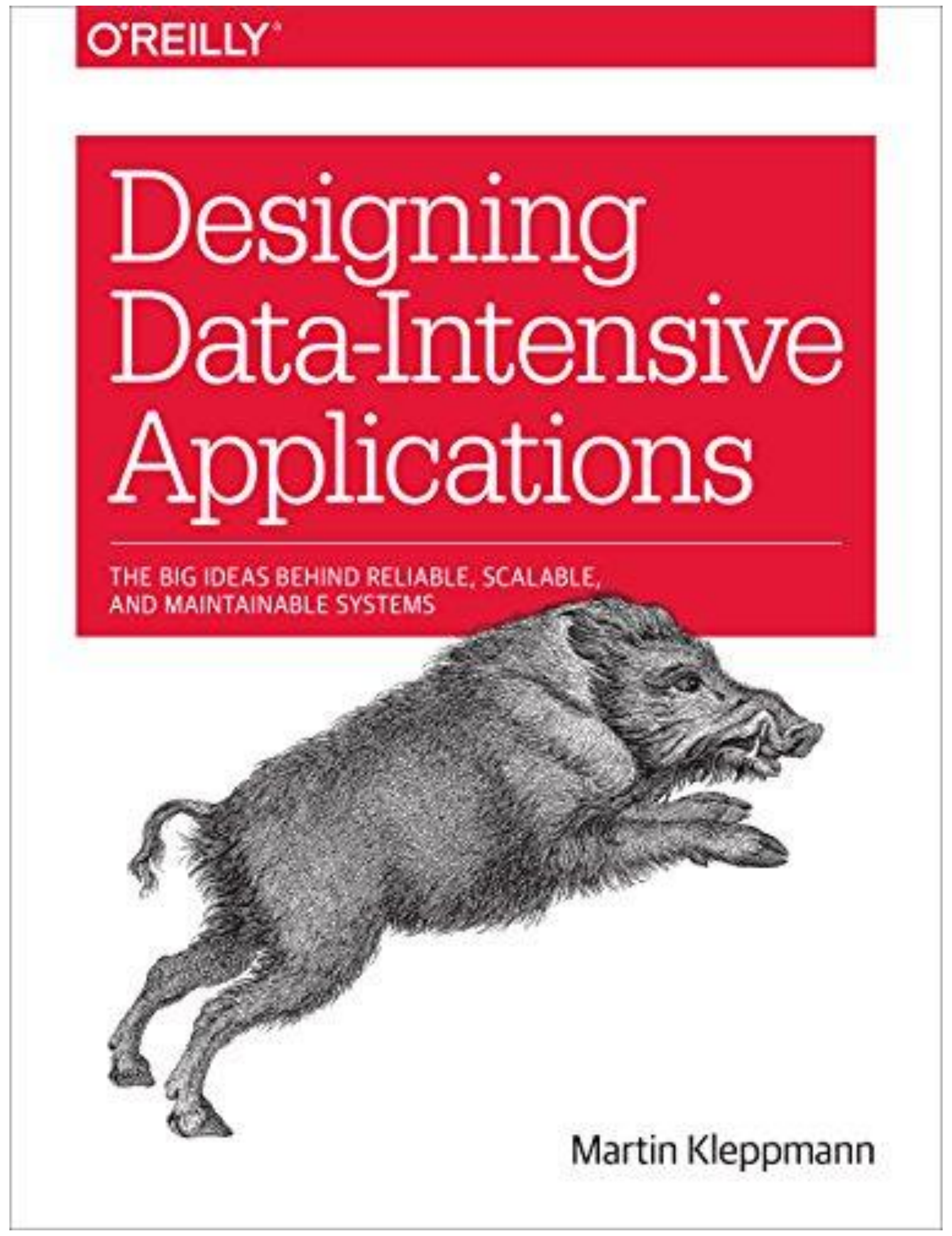
What is this course about?

- **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~~

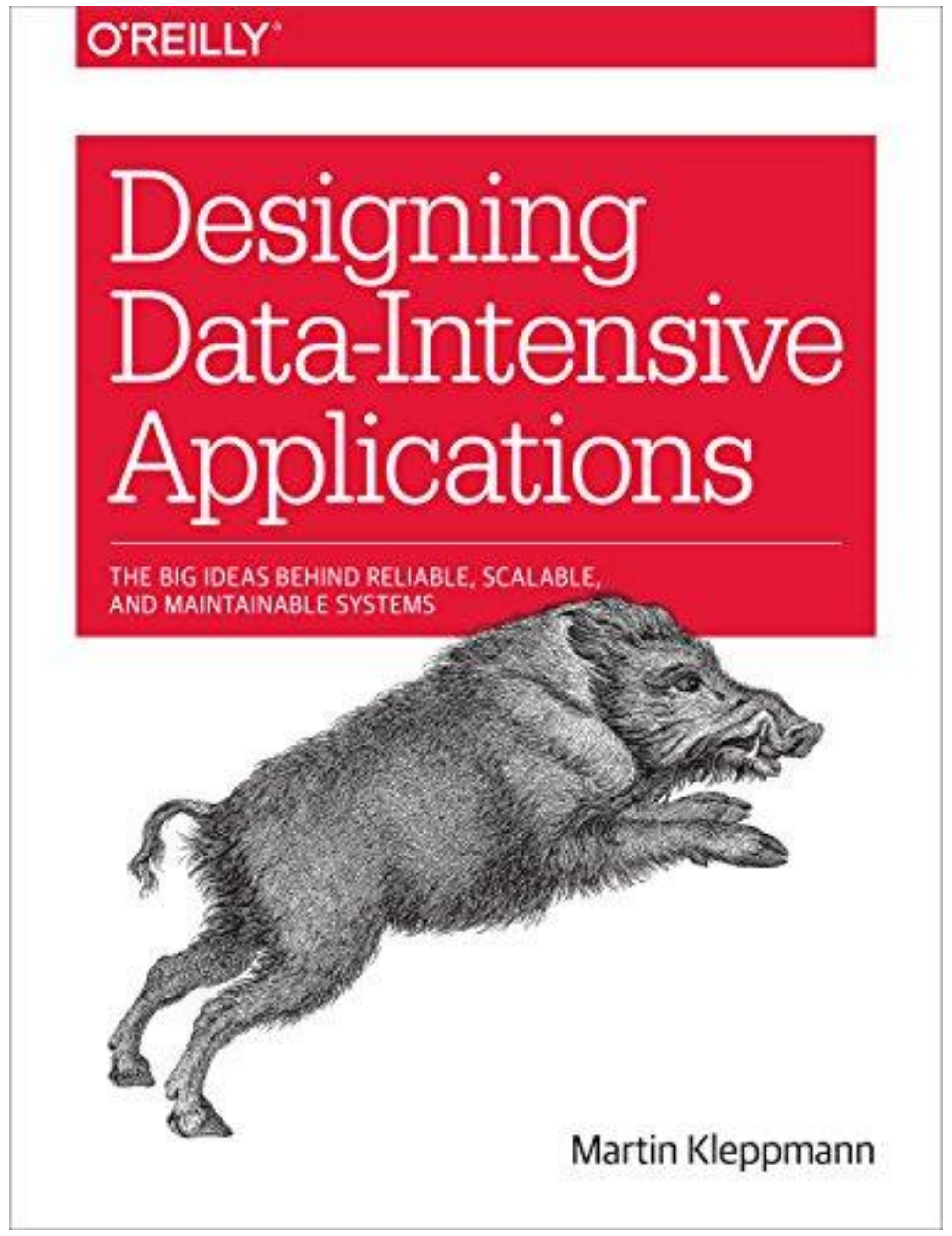
Suggested Textbooks

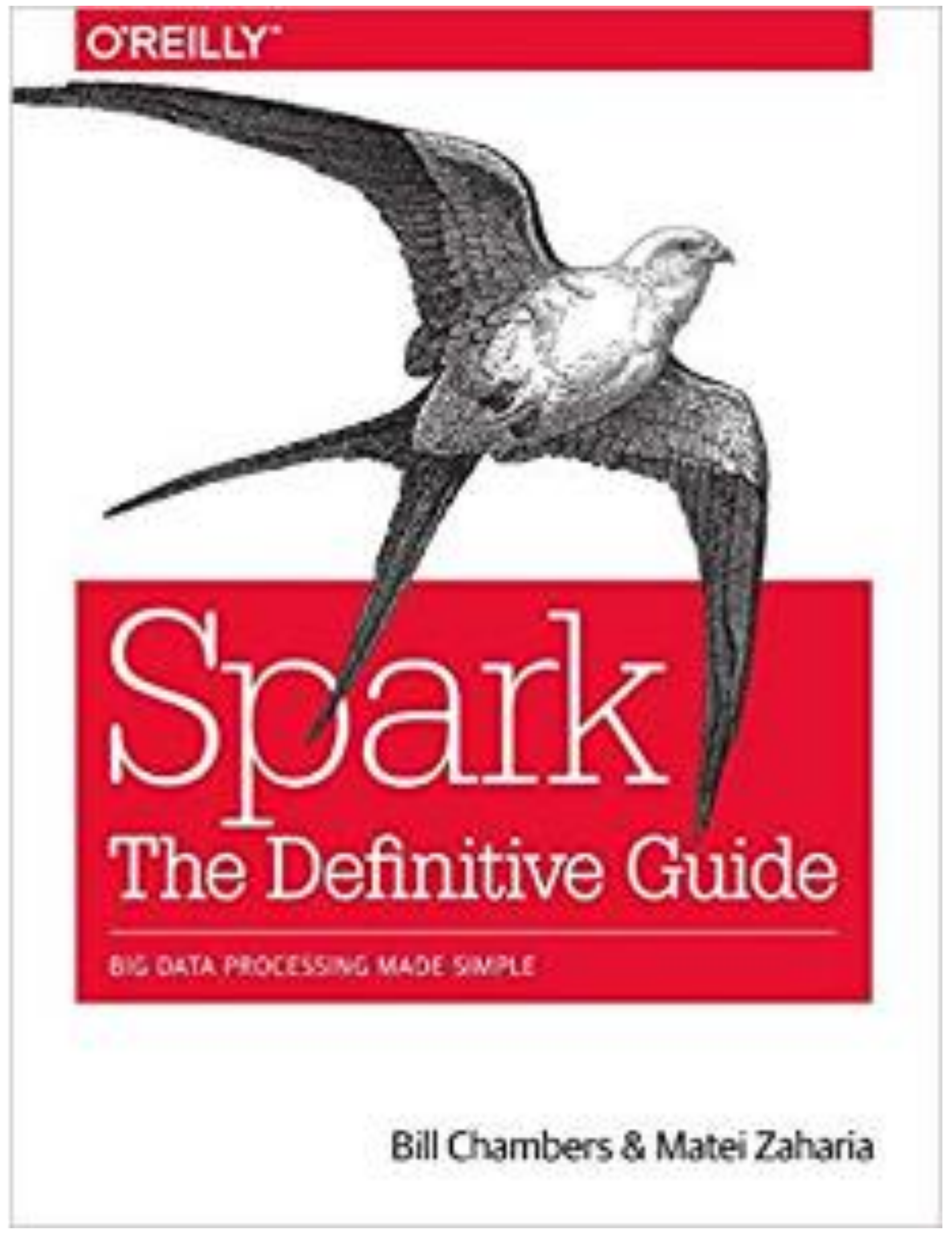

O'REILLY
Designing Data-Intensive Applications
THE BIG IDEAS BEHIND RELIABLE, SCALABLE, AND MAINTAINABLE SYSTEMS
Martin Kleppmann

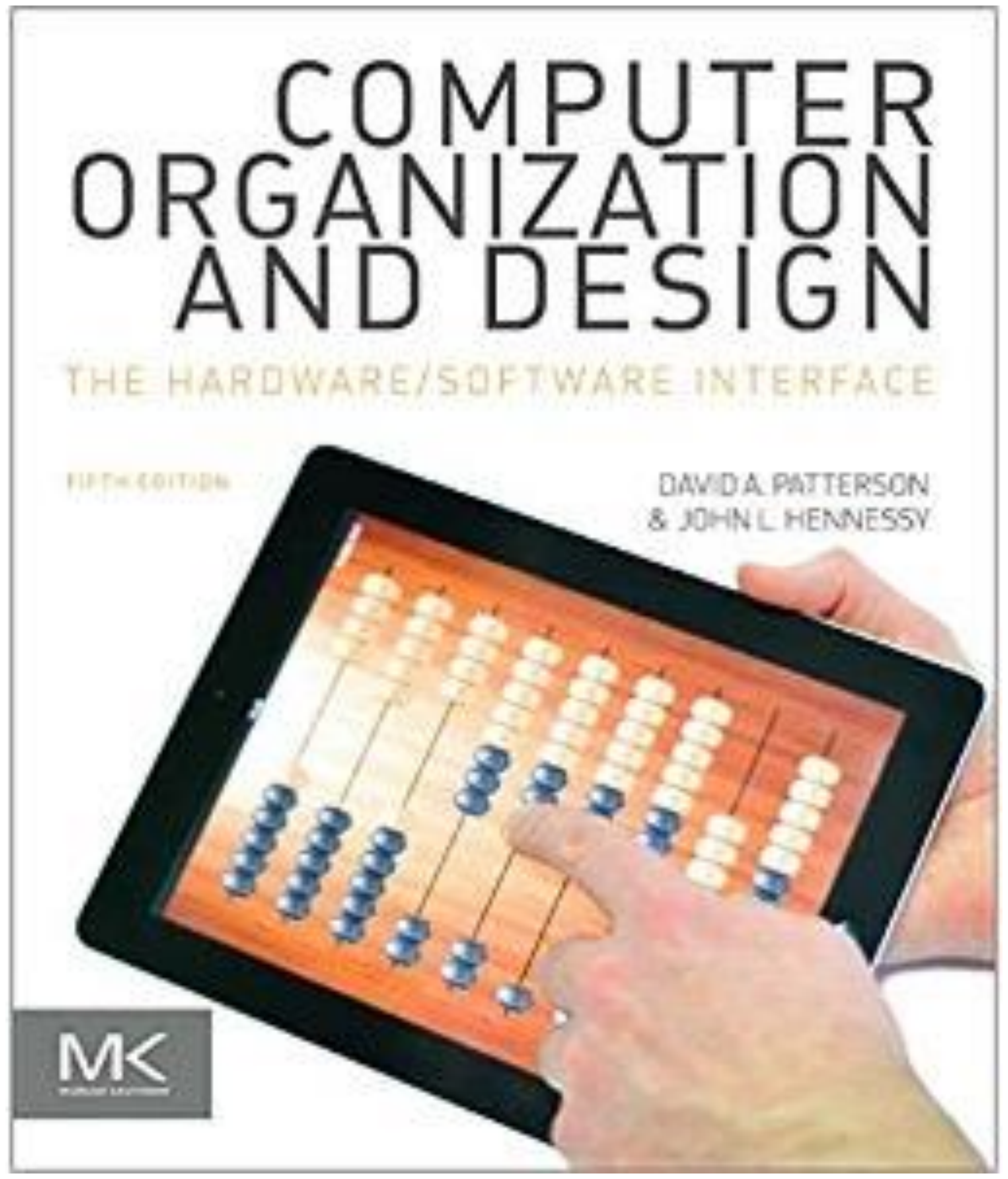
- Chapter 3. Storage and retrieval
- Chapter 4. Encoding and evolution
- Chapter 10. Batch processing
- Chapter 11. Stream processing
- Chapter 12. The future of data systems
- ~~The other chapters~~

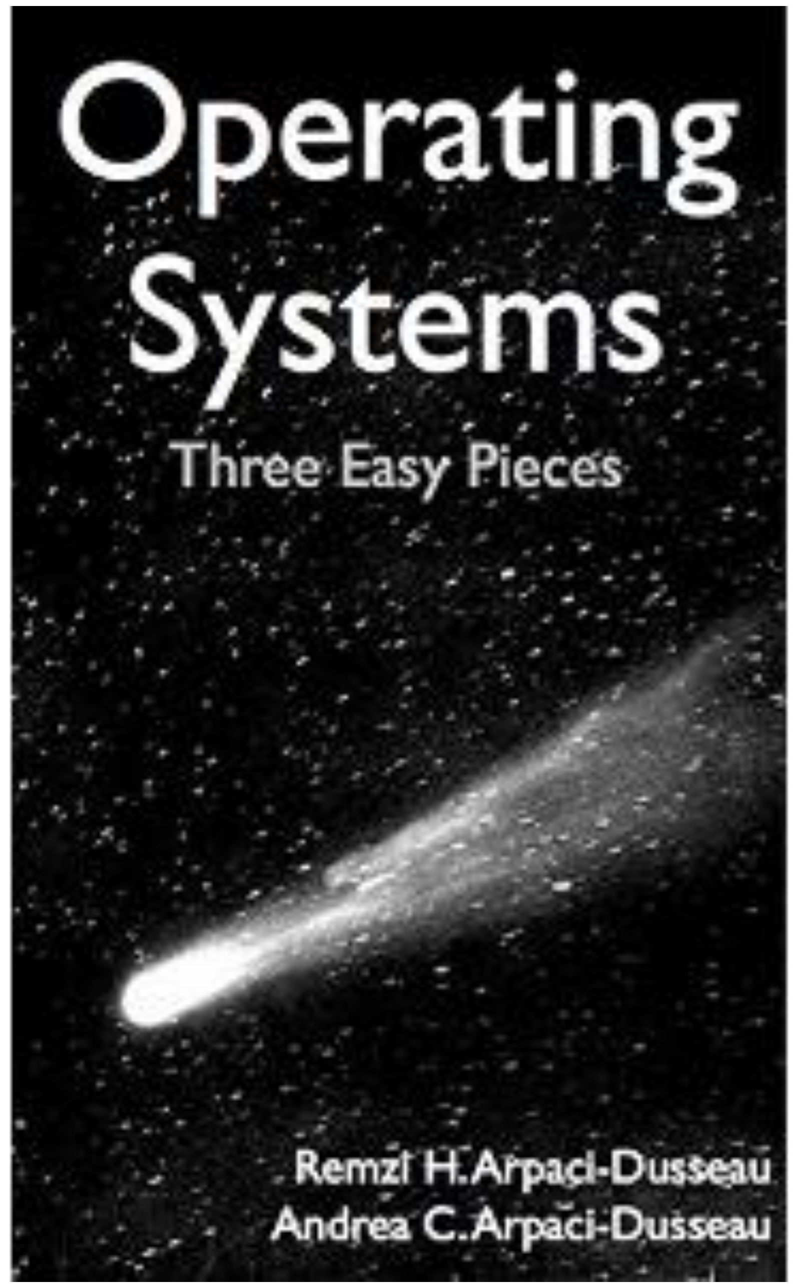
Suggested Textbooks

Computer systems are about carefully layering levels of abstraction.


O'REILLY
Designing Data-Intensive Applications
THE BIG IDEAS BEHIND RELIABLE, SCALABLE, AND MAINTAINABLE SYSTEMS
Martin Kleppmann


O'REILLY
Spark
The Definitive Guide
BIG DATA PROCESSING MADE SIMPLE
Bill Chambers & Matei Zaharia


COMPUTER ORGANIZATION AND DESIGN
THE HARDWARE/SOFTWARE INTERFACE
FIFTH EDITION
DAVID A. PATTERSON
& JOHN L. HENNESSY
MK



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
  - But you'll need to study how to use them by yourself
- Not a machine learning course
  - We focus more on system and data

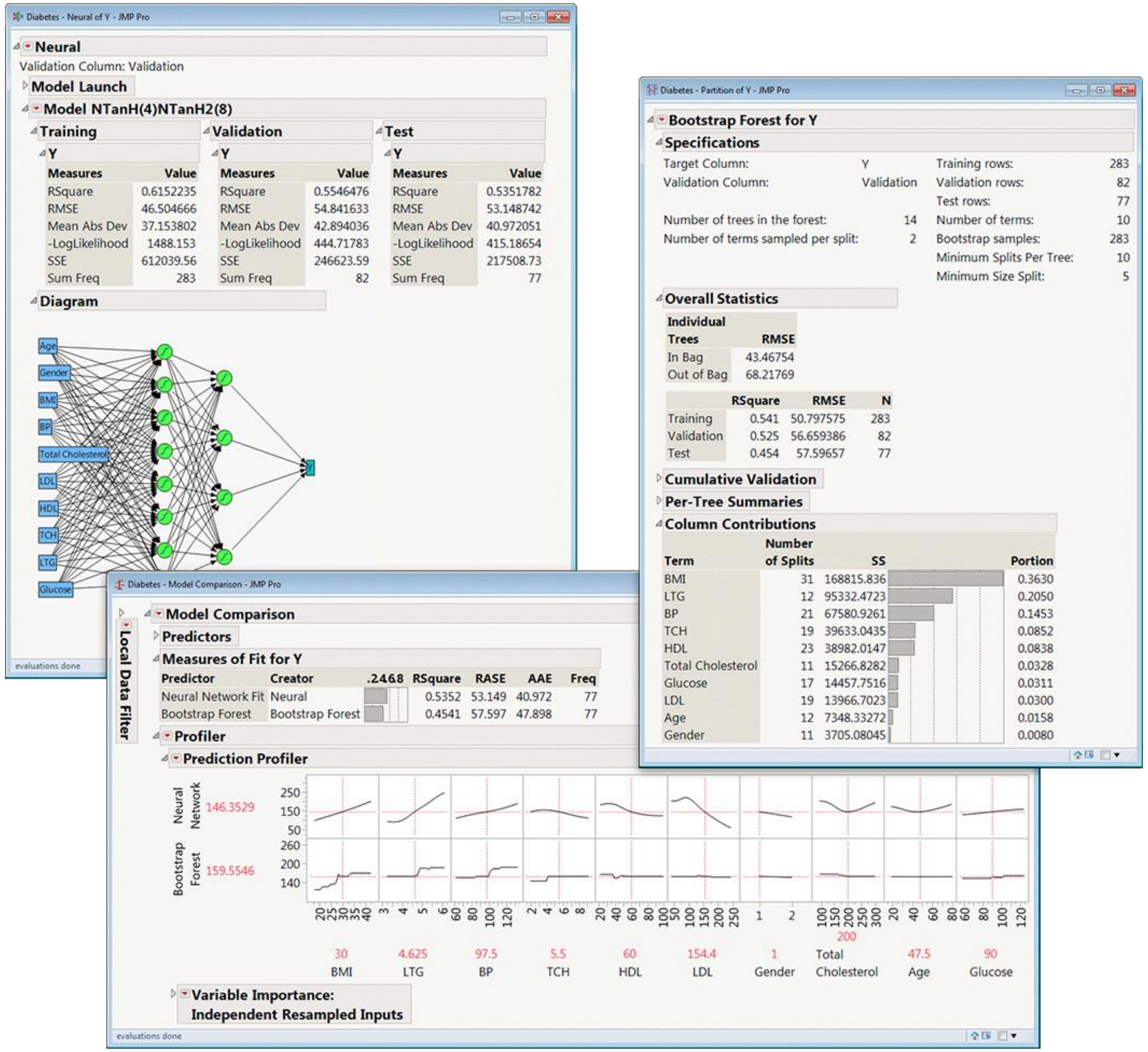
Big Deltas of this year offering

- The pace will be faster: less basics, more advanced stuffs
    - Take DSC 202 or DSC102 instead if you expect more basics (pre-requisite)
- ~1/4 will be about new systems developed between 16 – 22
    - Data + ML systems: TensorFlow, PyTorch, Ray
    - Machine learning parallelism
    - LLM systems
- Homework redesigned to be based on Ray
- No midterm exam, more paper readings, scribe notes

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

“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



Neural

Validation Column: Validation

Model Launch

Model NTanH(4)NTanH2(8)

Training — Y

Measures	Value
RSquare	0.6152235
RMSE	46.504666
Mean Abs Dev	37.153802
-LogLikelihood	1488.153
SSE	612039.56
Sum Freq	283

Validation — Y

Measures	Value
RSquare	0.5546476
RMSE	54.841633
Mean Abs Dev	42.894036
-LogLikelihood	444.71783
SSE	246623.59
Sum Freq	82

Test — Y

Measures	Value
RSquare	0.5351782
RMSE	53.148742
Mean Abs Dev	40.972051
-LogLikelihood	415.18654
SSE	217508.73
Sum Freq	77

Diagram

Bootstrap Forest for Y

Specifications

Target Column:	Y	Training rows:	283
Validation Column:	Validation	Validation rows:	82
		Test rows:	77
Number of trees in the forest:	14	Number of terms:	10
Number of terms sampled per split:	2	Bootstrap samples:	283
		Minimum Splits Per Tree:	10
		Minimum Size Split:	5

Overall Statistics

Individual Trees	RMSE
In Bag	43.46754
Out of Bag	68.21769

	RSquare	RMSE	N
Training	0.541	50.797575	283
Validation	0.525	56.659386	82
Test	0.454	57.59657	77

Cumulative Validation

Per-Tree Summaries

Column Contributions

Term	Number of Splits	SS	Portion
BMI	31	168815.836	0.3630
LTG	12	95332.4723	0.2050
BP	21	67580.9261	0.1453
TCH	19	39633.0435	0.0852
HDL	23	38982.0147	0.0838
Total Cholesterol	11	15266.8282	0.0328
Glucose	17	14457.7516	0.0311
LDL	19	13966.7023	0.0300
Age	12	7348.33272	0.0158
Gender	11	3705.08045	0.0080

Model Comparison

Predictors

Measures of Fit for Y

Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
Neural Network Fit	Neural		0.5352	53.149	40.972	77
Bootstrap Forest	Bootstrap Forest		0.4541	57.597	47.898	77

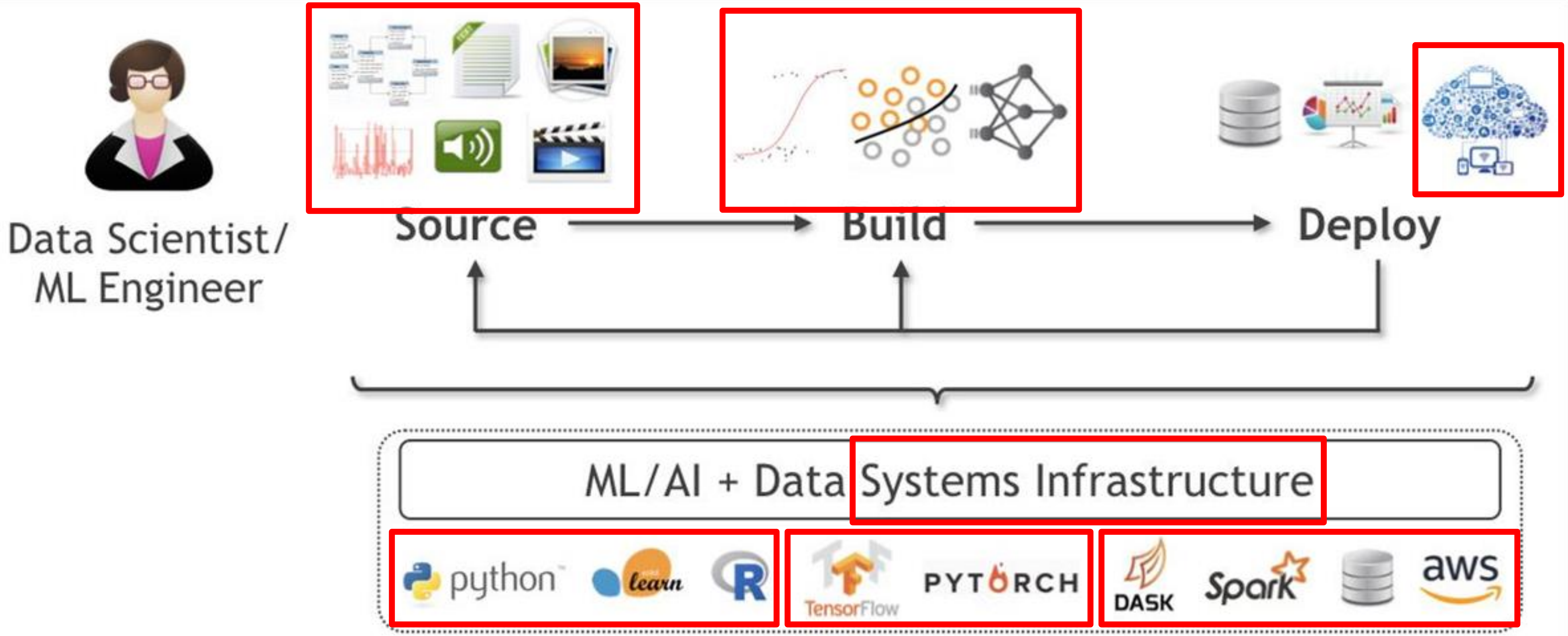
Profiler

Prediction Profiler

Variable Importance: Independent Resampled Inputs

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

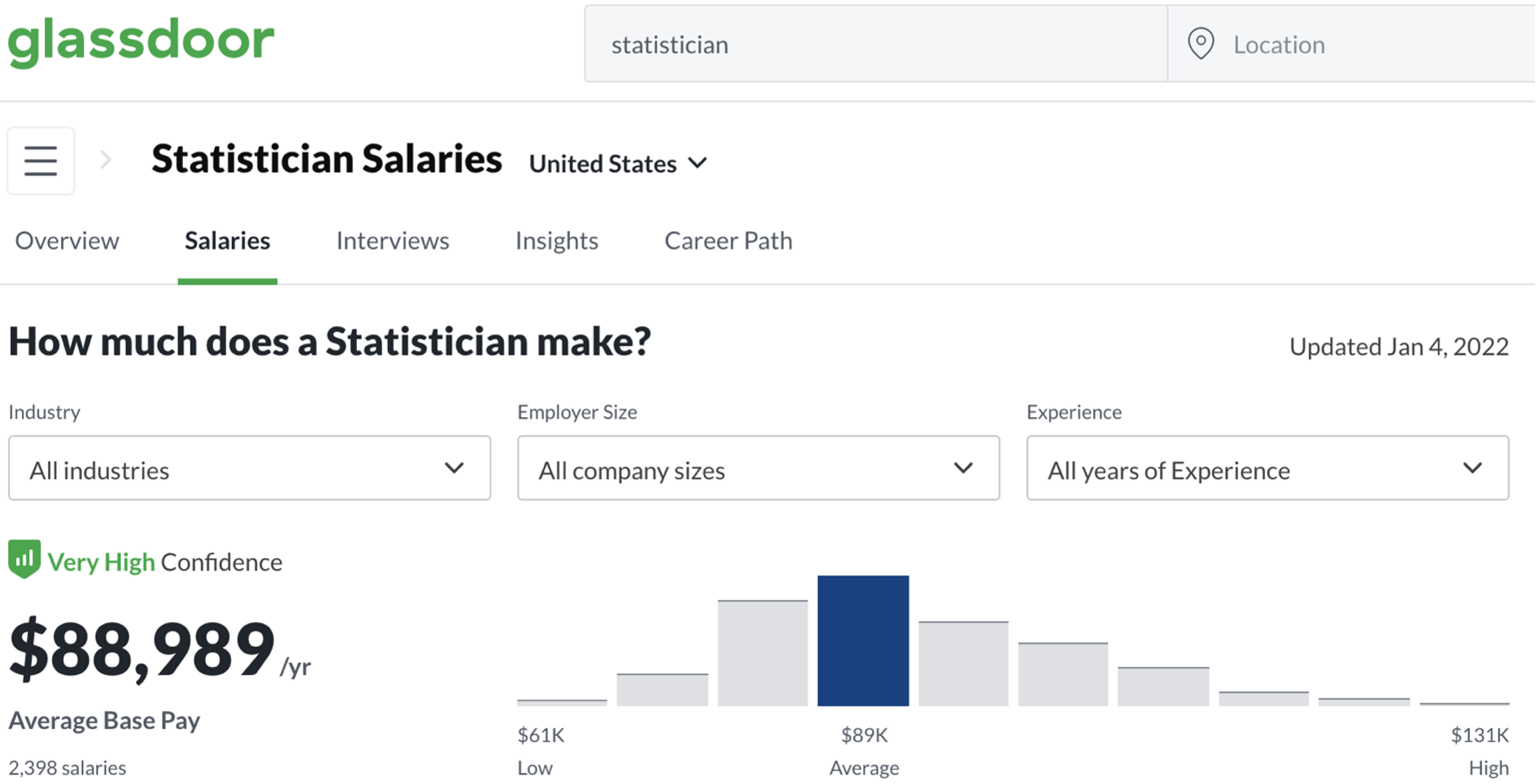
In the era of 2020s:



Data acquisition
Data preparation

Feature Engineering
Training & Inference
Model Selection

Serving
Monitoring



Statistician Salaries

United States

How much does a Statistician make?

Updated Jan 4, 2022

Industry: All industries

Employer Size: All company sizes

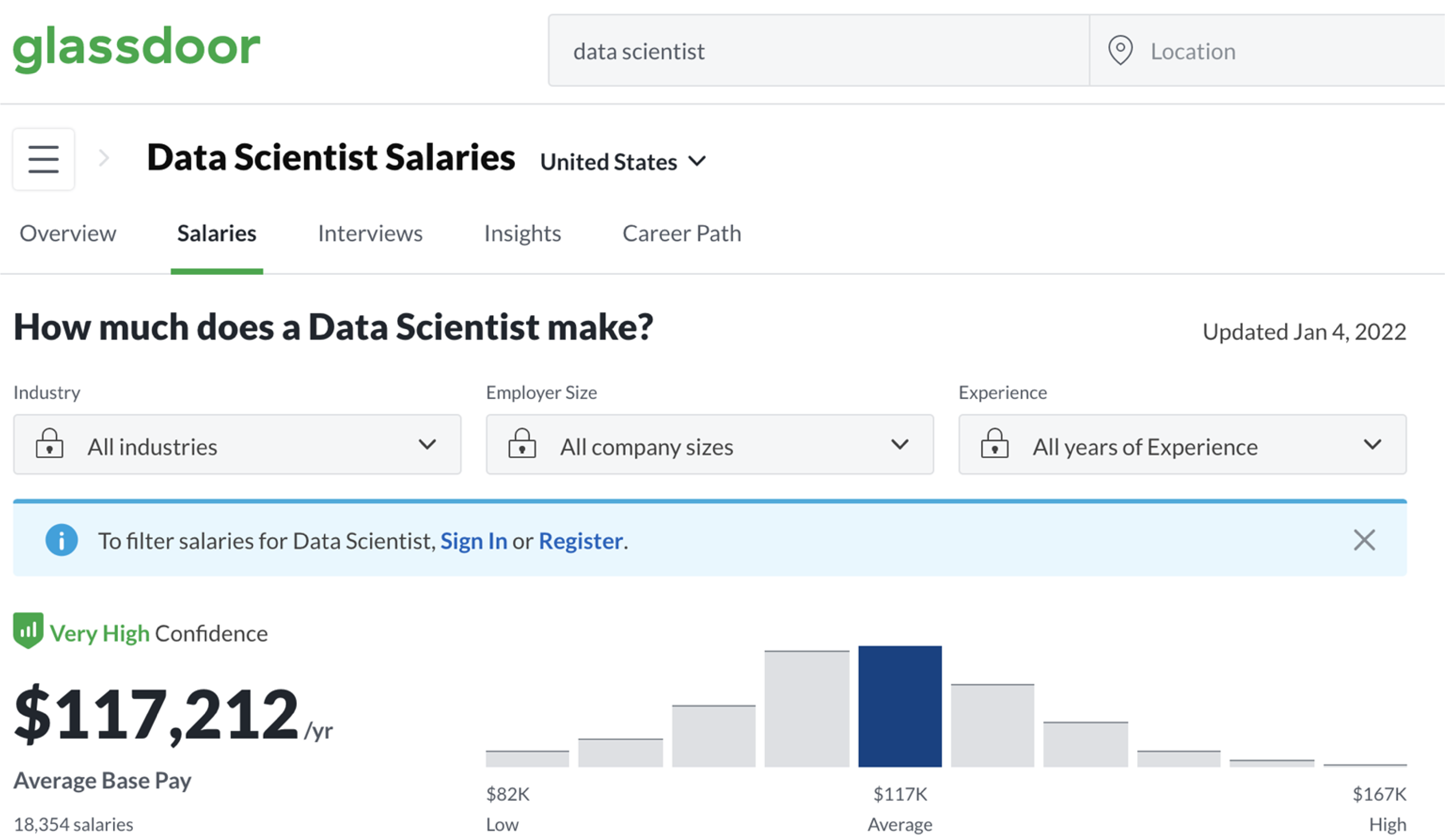
Experience: All years of Experience

Very High Confidence

$88,989 /yr

Average Base Pay

2,398 salaries



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

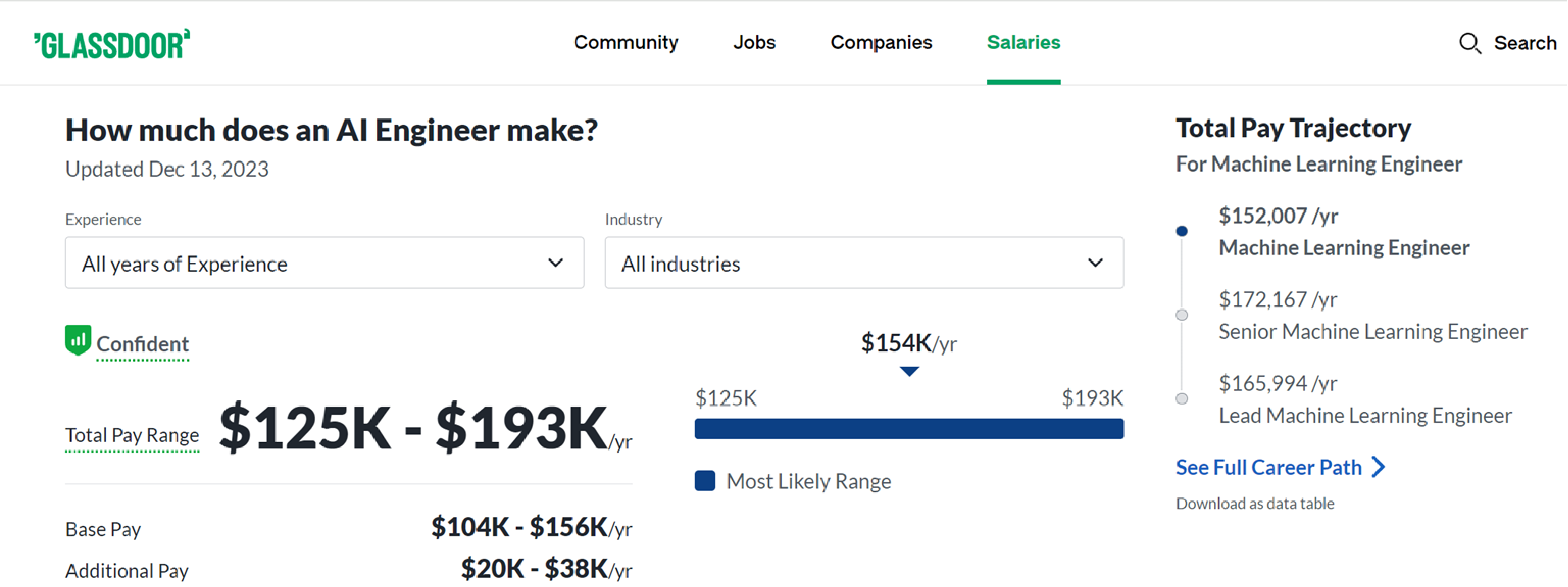
$117,212 /yr

Average Base Pay

18,354 salaries

— $88,989

= $28,223!



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

$154K/yr

$125K $193K

■ Most Likely Range

Base Pay	**$104K - $156K**/yr
Additional Pay	**$20K - $38K**/yr

Total Pay Trajectory

For Machine Learning Engineer

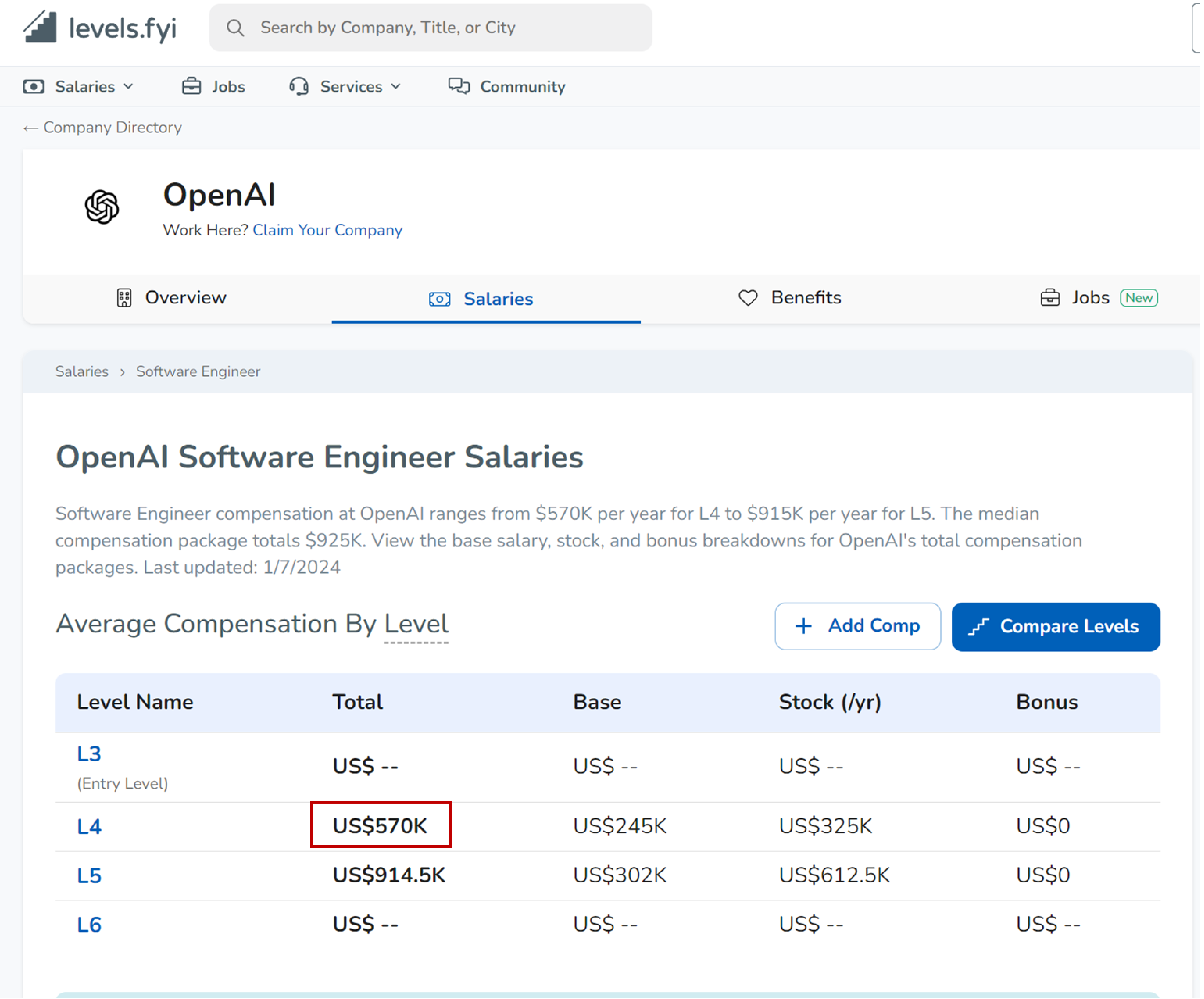
- $152,007 /yr
  Machine Learning Engineer
- $172,167 /yr
  Senior Machine Learning Engineer
- $165,994 /yr
  Lead Machine Learning Engineer

See Full Career Path >

Download as data table

— $88,989

= $65011!



OpenAI

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

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 model companies: they operate on either big model or big models.

Fastest-growing data companies

snowflake

databricks

Fastest-growing model companies

OpenAI

ANTHROP\C

Questions?

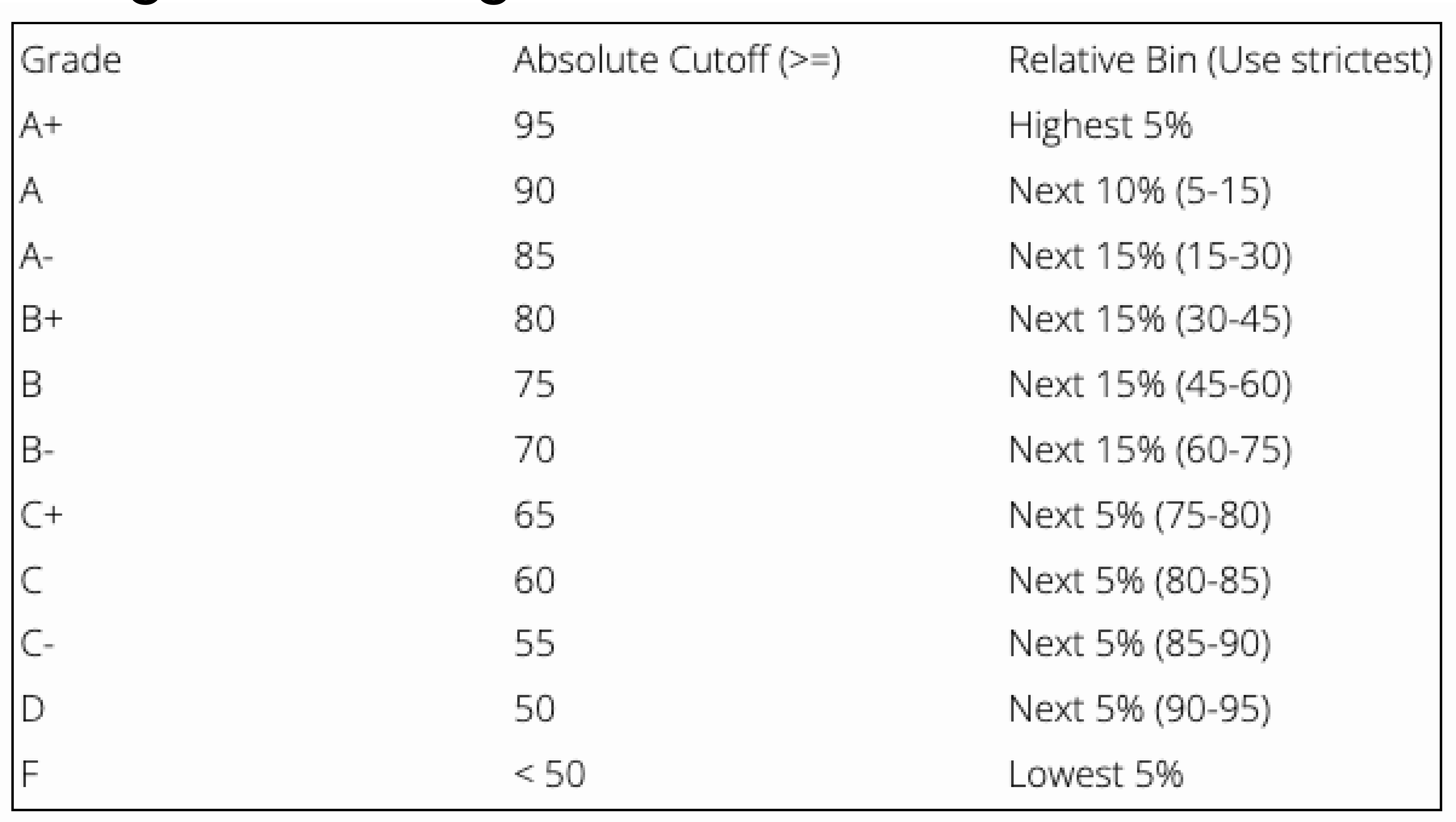
Prerequisites

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

Components and Grading

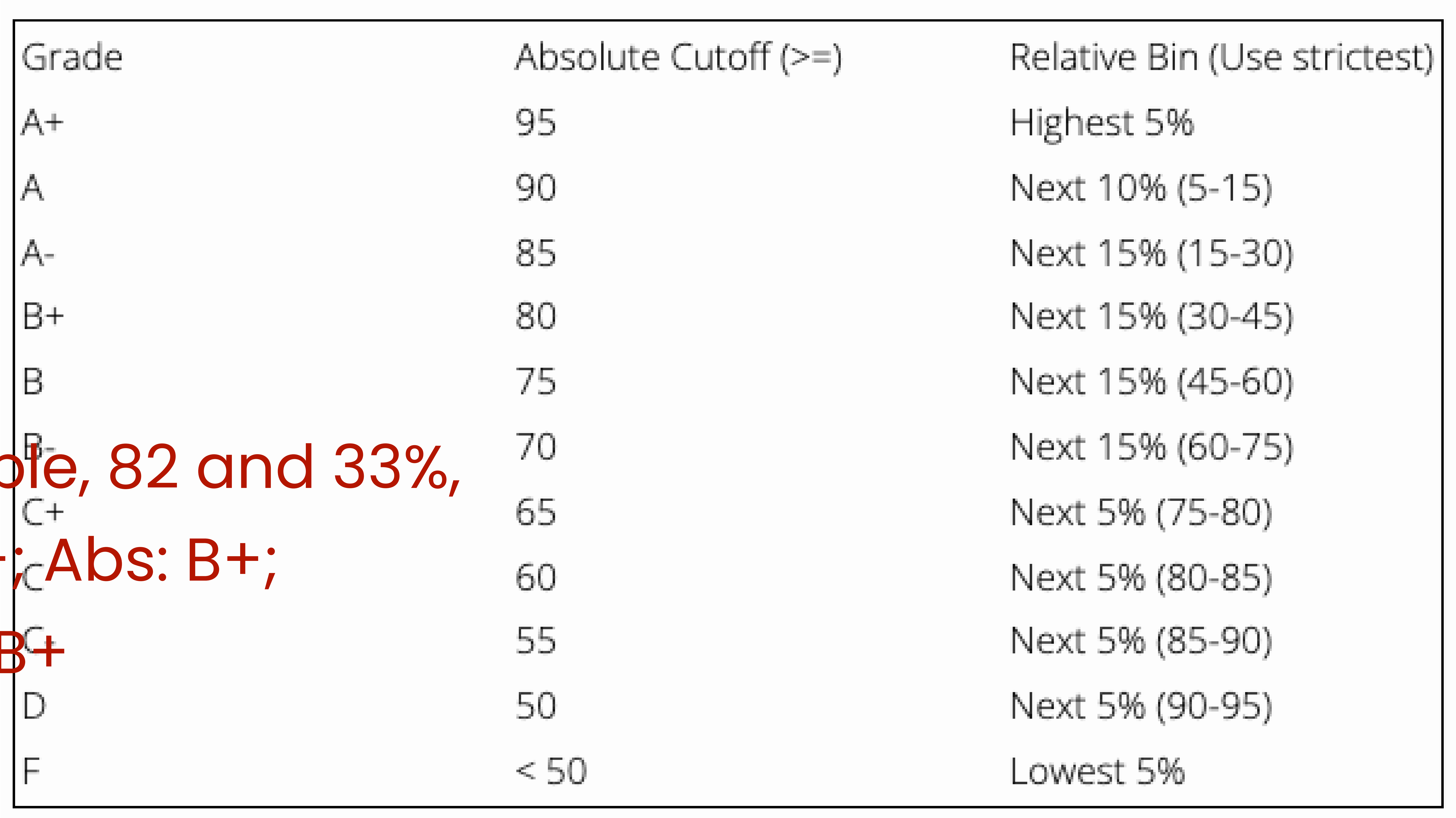
- 3 Programming Assignments: **44%** (12% + 16% + 16%)
  - No late days! Plan your work well ahead.
- No Midterm (cheers!)
- Final Exam (03/22/2024 8am-11am): **36%**
- Scribe Duties: 8%
- **Reading summary: 12%**
- Extra Credit: **5%**

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-w24/

Programming Assignments

- Three newly designed PAs
- Will be based on Ray: https://www.ray.io/
- Topics: exploring distributed data exploration, processing, and distributed ML
- The school be allocating $50 AWS credits to each student
- You only have $50 AWS credit! Close the instance when you finish.

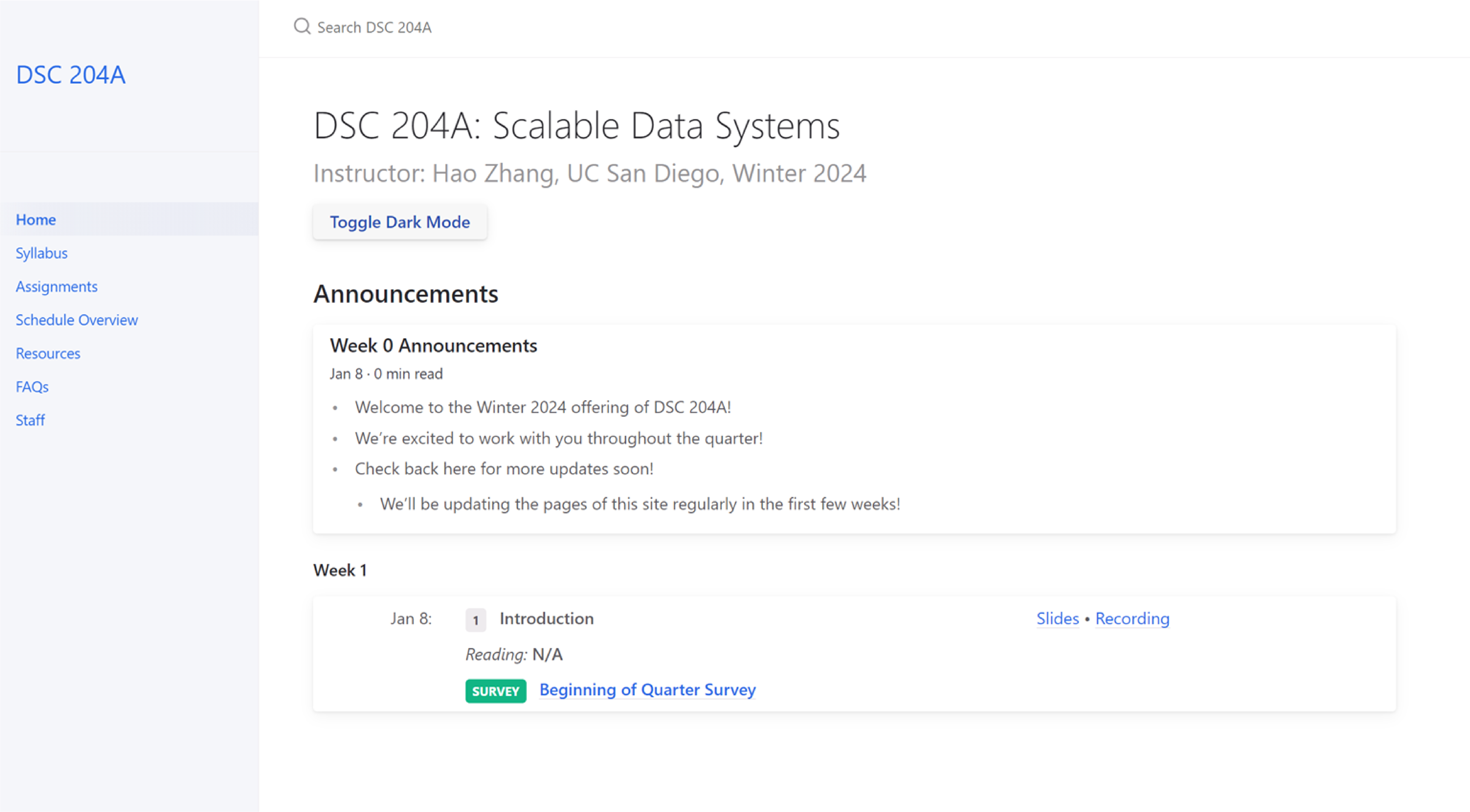
Expectations on the PAs

- 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!

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



DSC 204A: Scalable Data Systems

Instructor: Hao Zhang, UC San Diego, Winter 2024

Announcements

Week 0 Announcements

Jan 8 · 0 min read

- Welcome to the Winter 2024 offering of DSC 204A!
- We're excited to work with you throughout the quarter!
- Check back here for more updates soon!
  - We'll be updating the pages of this site regularly in the first few weeks!

Week 1

Jan 8: 1 Introduction Slides • Recording

Reading: N/A

SURVEY Beginning of Quarter Survey

🌎 https://hao-ai-lab.github.io/dsc204a-w24/

Exploring Contents at website

🌎 https://hao-ai-lab.github.io/dsc204a-w24/

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 204:” as subject prefix

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!

Questions?