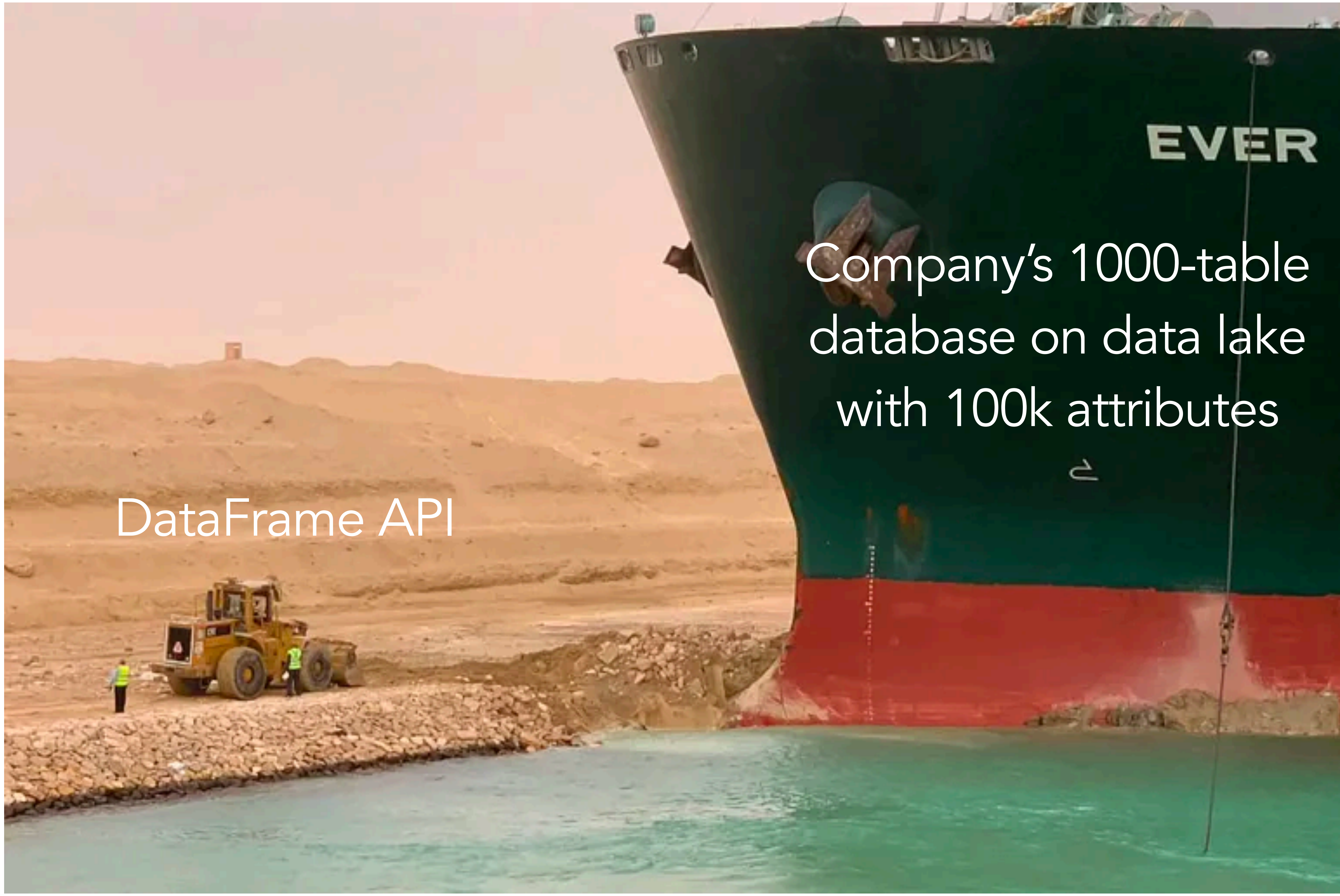
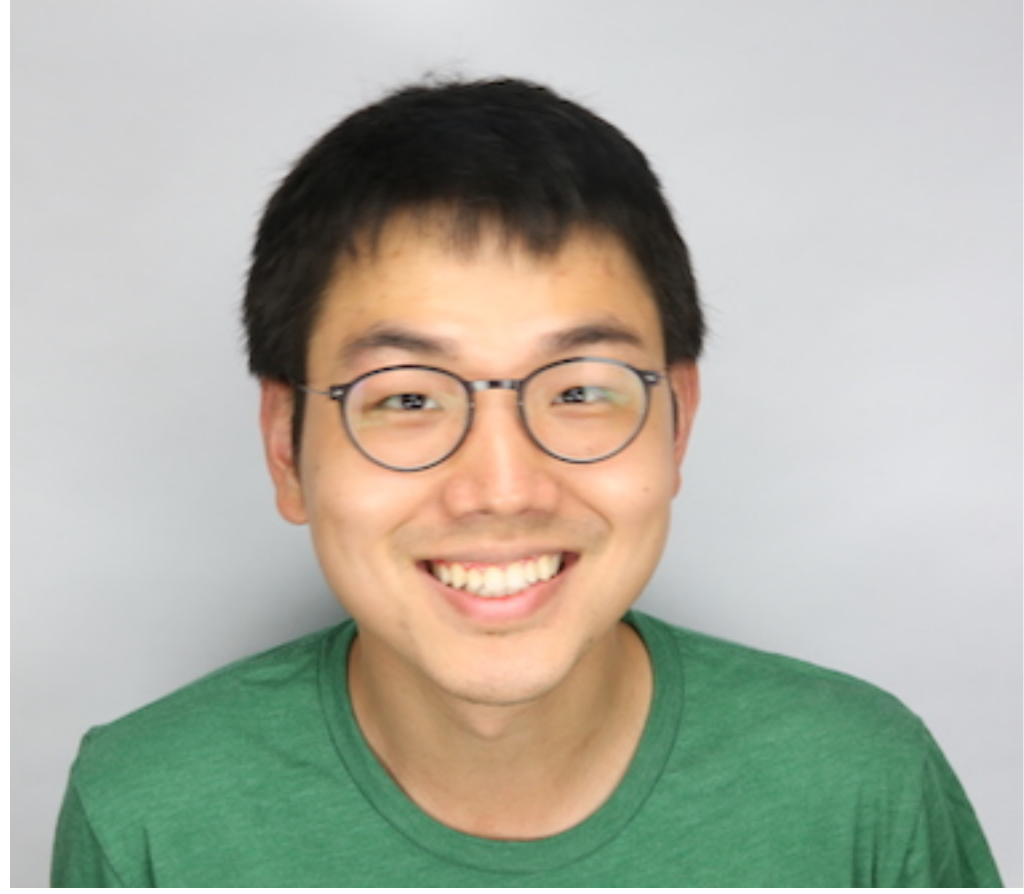
DSC 204a Scalable Data Systems

- Haojian Jin


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

Meme idea credit: https://datasystemsfun.tumblr.com/

Bio



Haojian Jin (http://haojianj.in/)

Asst. Prof @ UCSD-HDSI

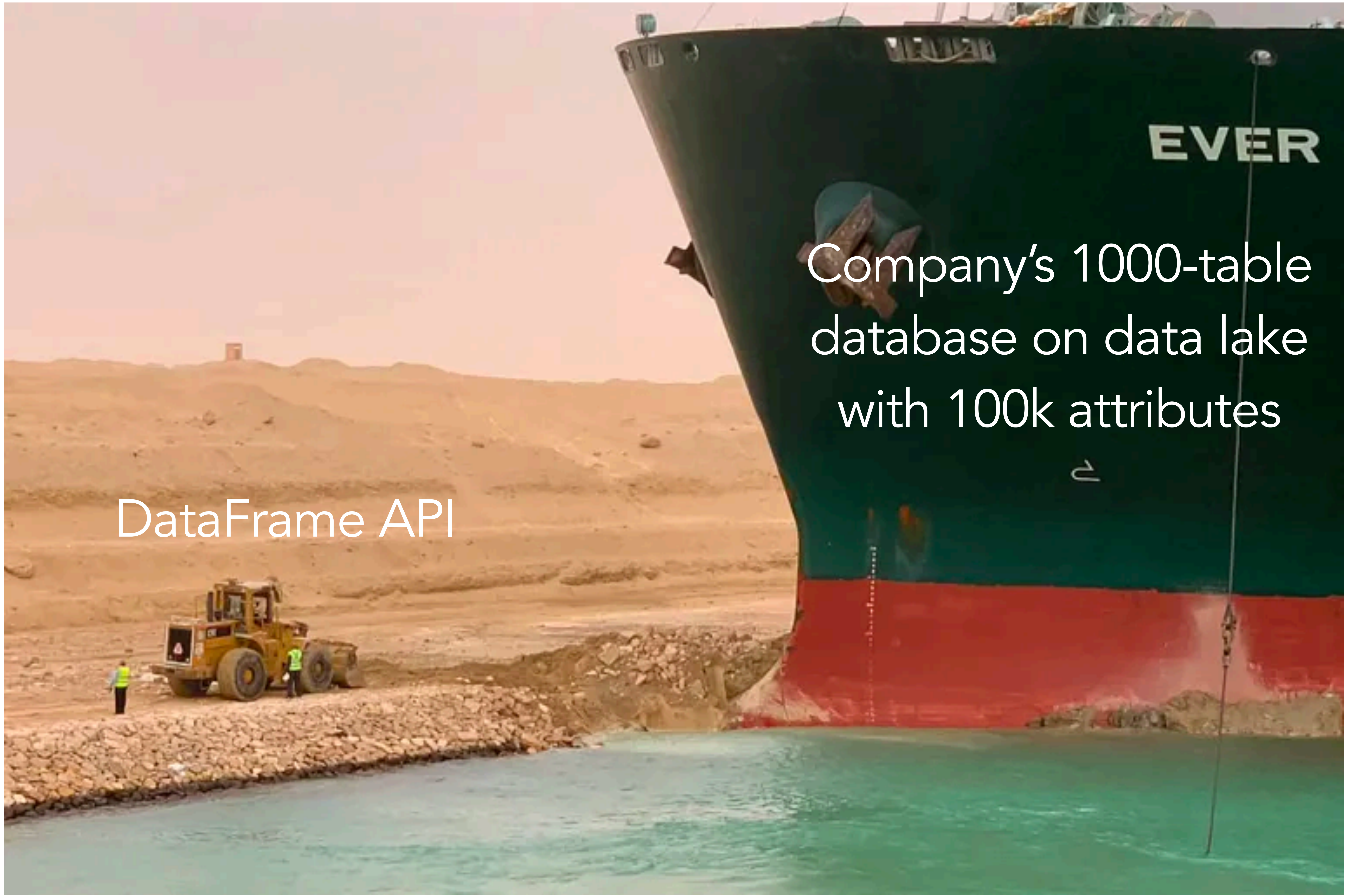
Data Smith Lab:

We study the security and privacy of data systems by researching the people who design, implement, and use these systems.

Ph.D. from CMU Human-Computer Interaction Institute

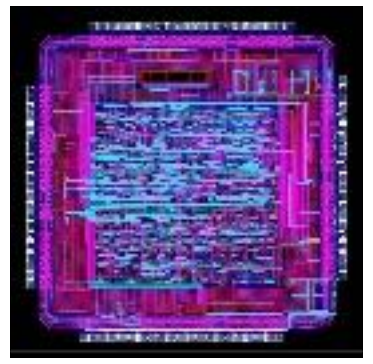
Before Ph.D.: worked at Yahoo Research, ran a startup

What is this course about?

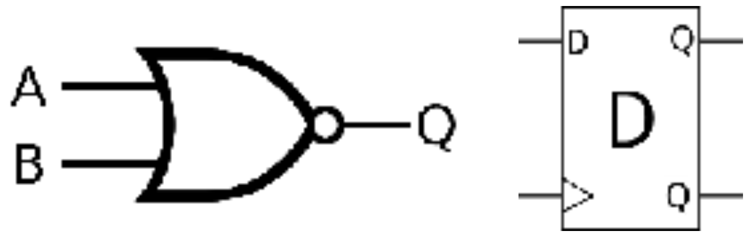

EVER
Company's 1000-table database on data lake with 100k attributes
DataFrame API

Levels of Abstraction

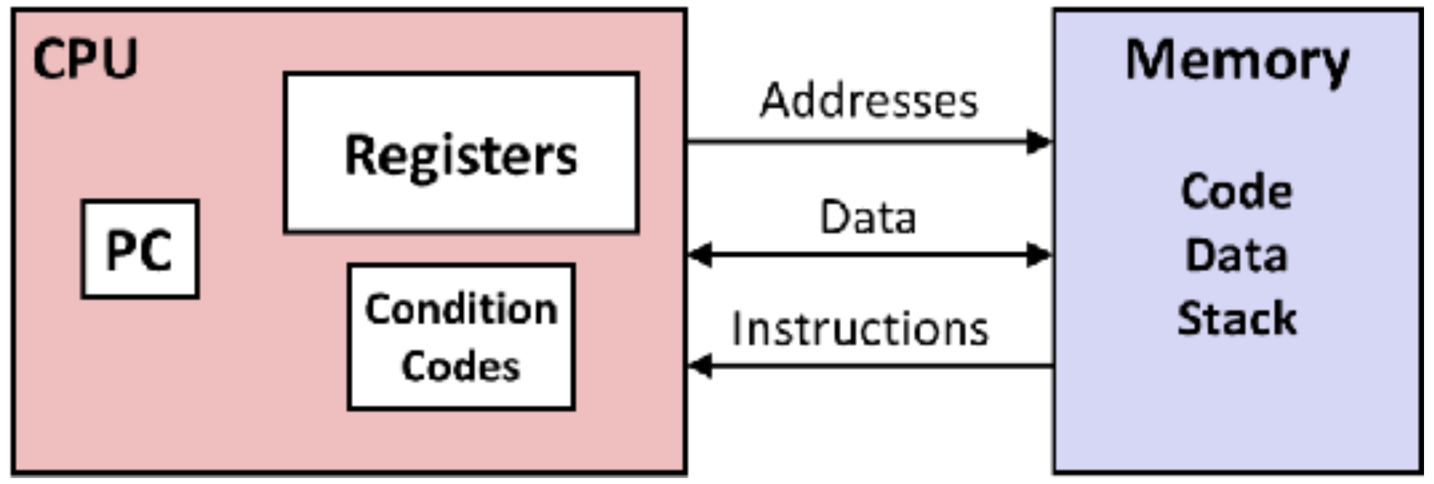
Computer Designer



Gates, clocks, circuit layout, …



Assembly programmer



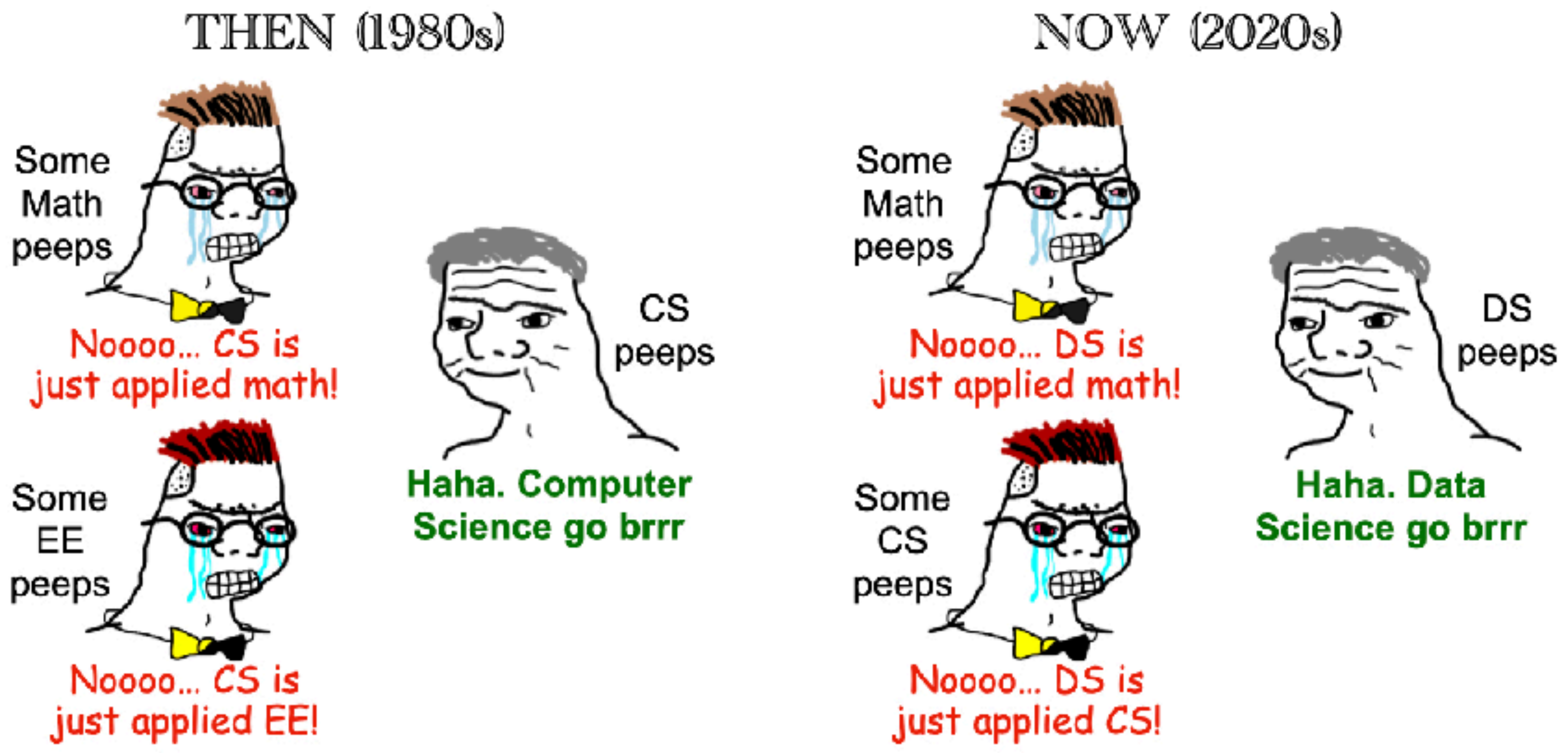
C programmer

```
#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 science professionals ought to be familiarized with data systems from a **user**'s standpoint, as opposed to the conventional approach of a **system implementer**.


THEN (1980s)
Some Math peeps
Noooo... CS is just applied math!
Some EE peeps
Noooo... CS is just applied EE!
CS peeps
Haha. Computer Science go brrr
NOW (2020s)
Some Math peeps
Noooo... DS is just applied math!
Some CS peeps
Noooo... DS is just applied CS!
DS peeps
Haha. Data Science go brrr

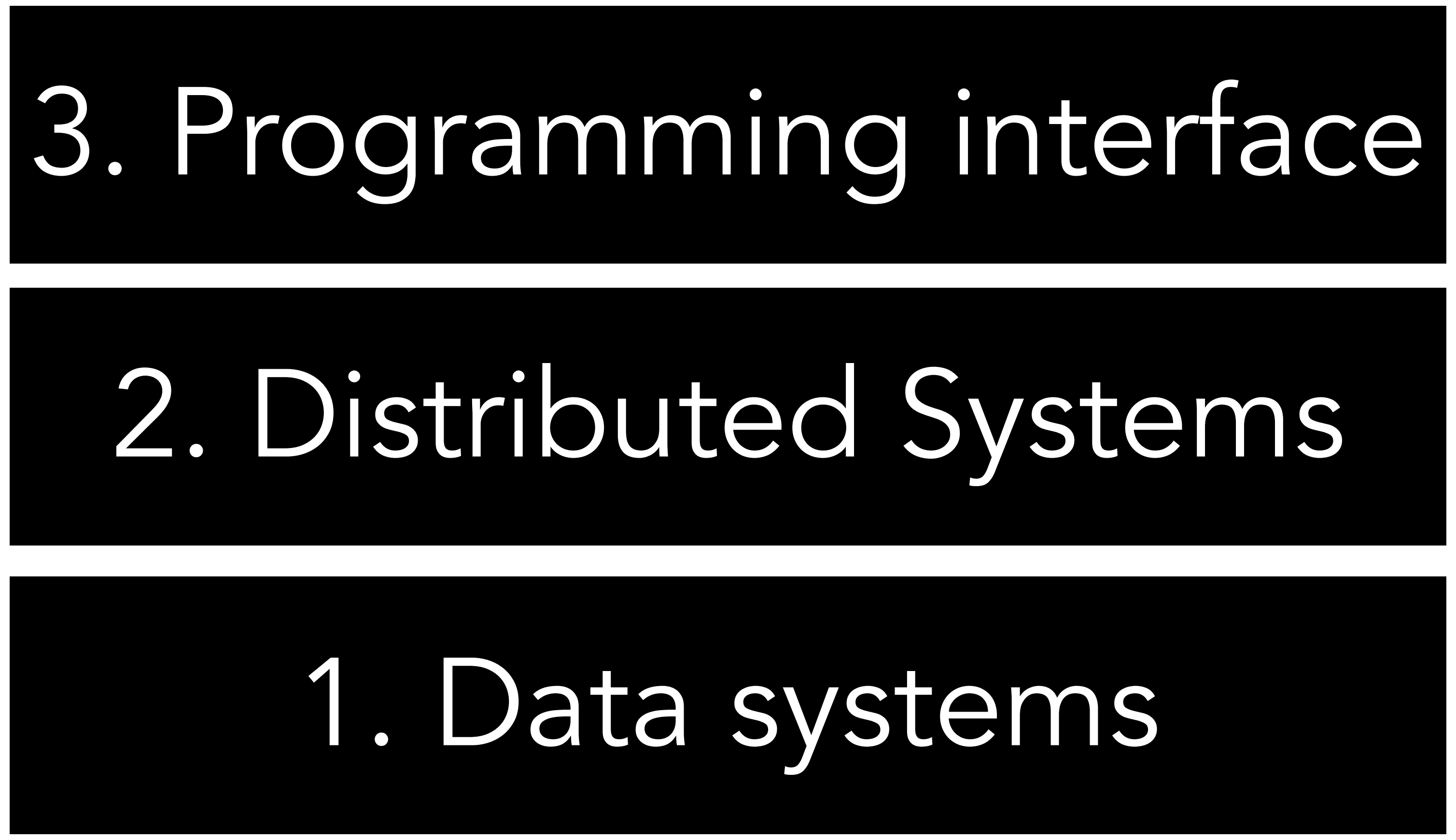
What is this course about?

- Relational databases
- NoSQL datastores
- Stream or batch processors
- Message brokers
- Spark, MapReduce, Hadoop, Kafka, HDFS
- Data lakes, column database
- ….

How to use and operate them more effectively?

What is this course about?

- Foundations of data systems
- Scaling distributed systems
- Data Processing and Programming model.



3. Programming interface

2. Distributed Systems

1. Data systems

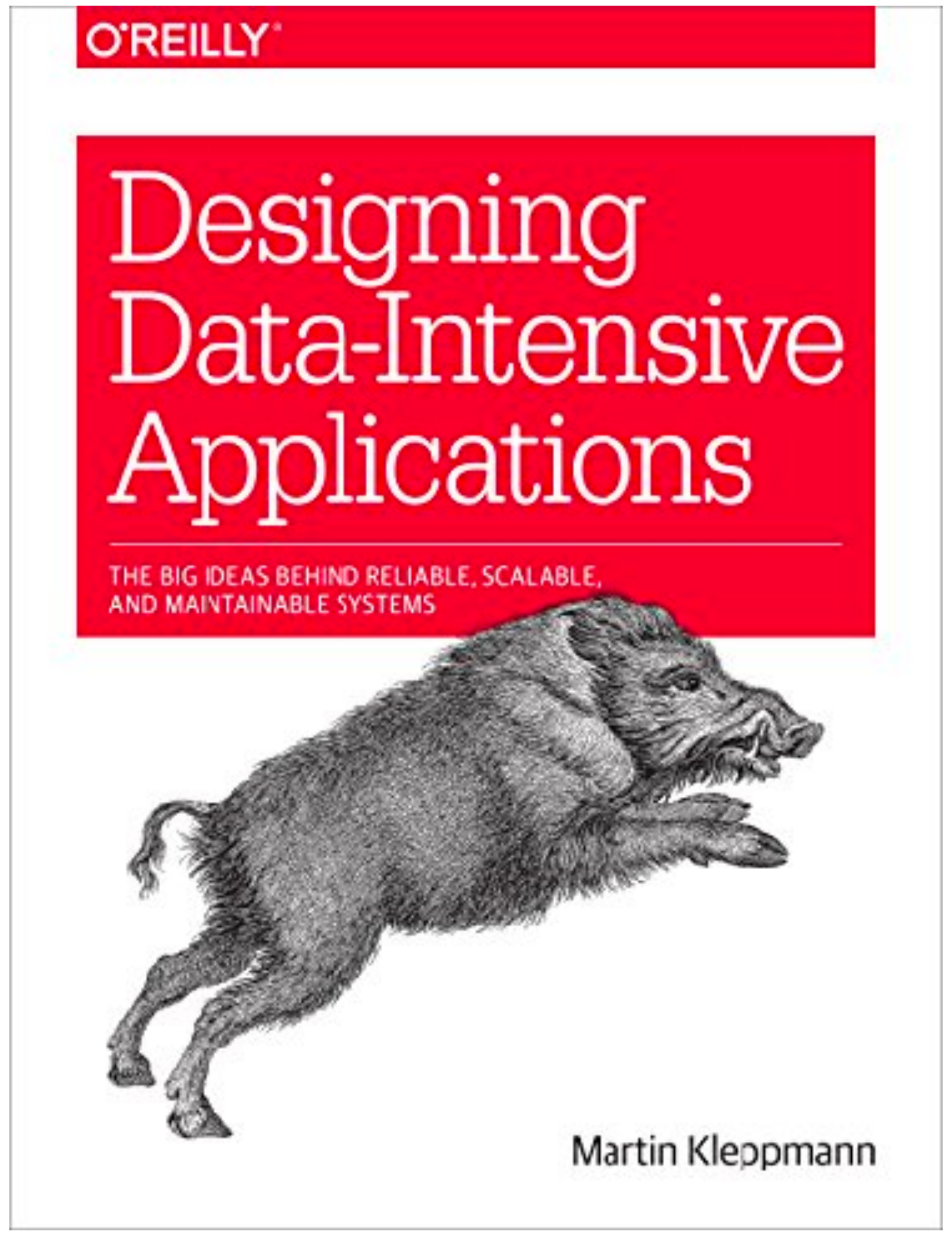
What is this course about?

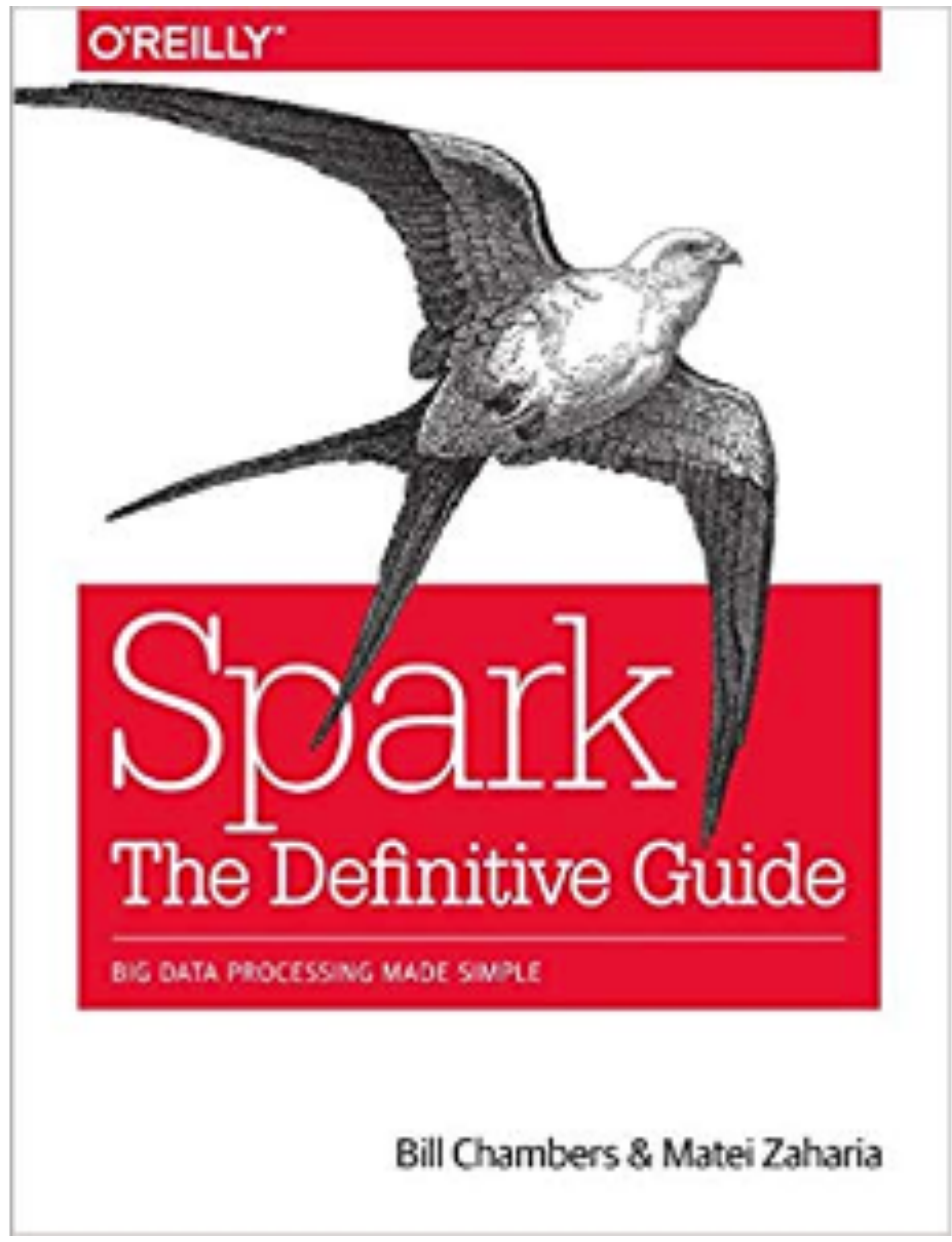
The data sci relevant components in the following course

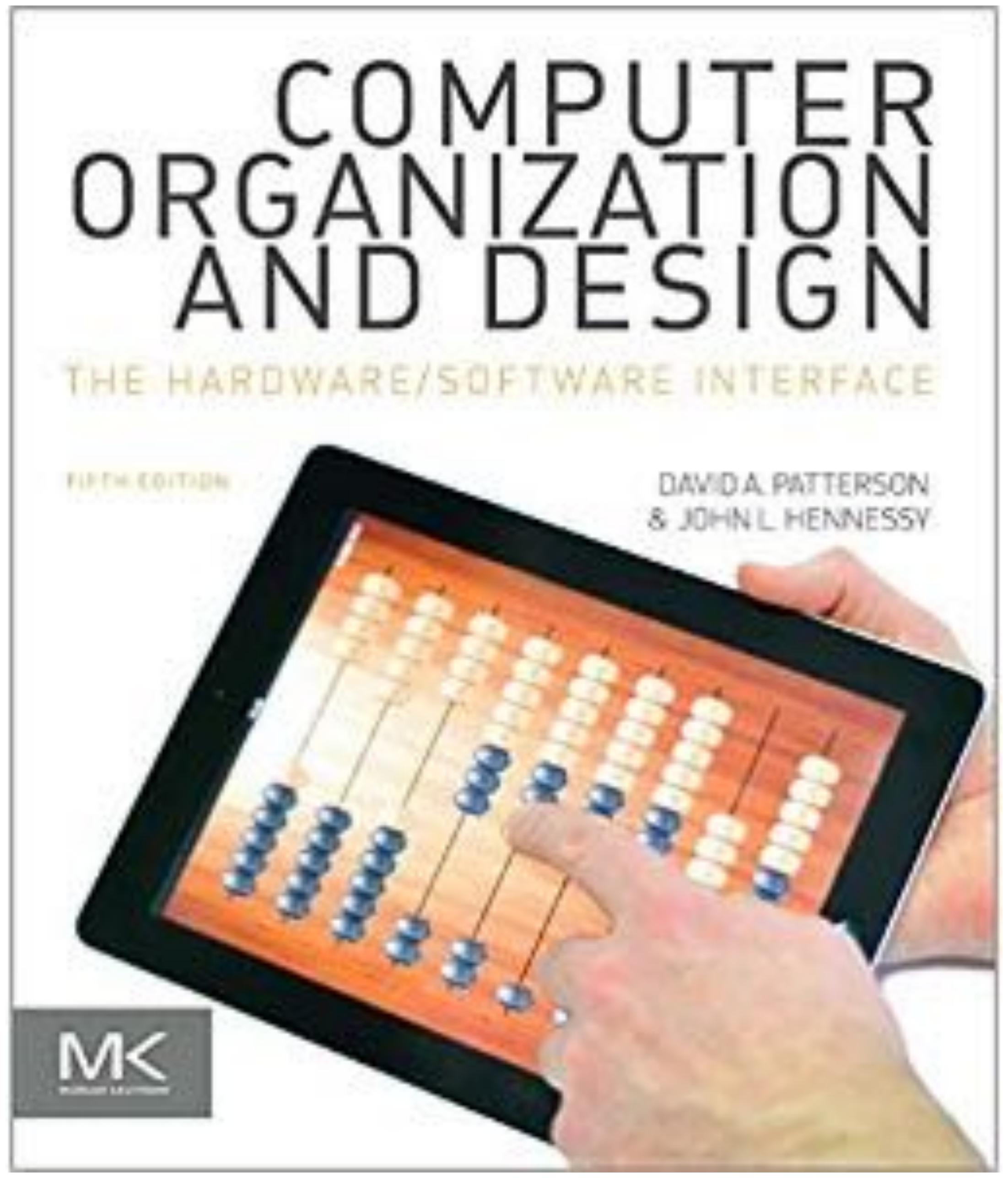
- Computer organization
- System programming
- Networks
- Operating systems
- Distributed systems
- Cloud computing
- + various data sci tricks

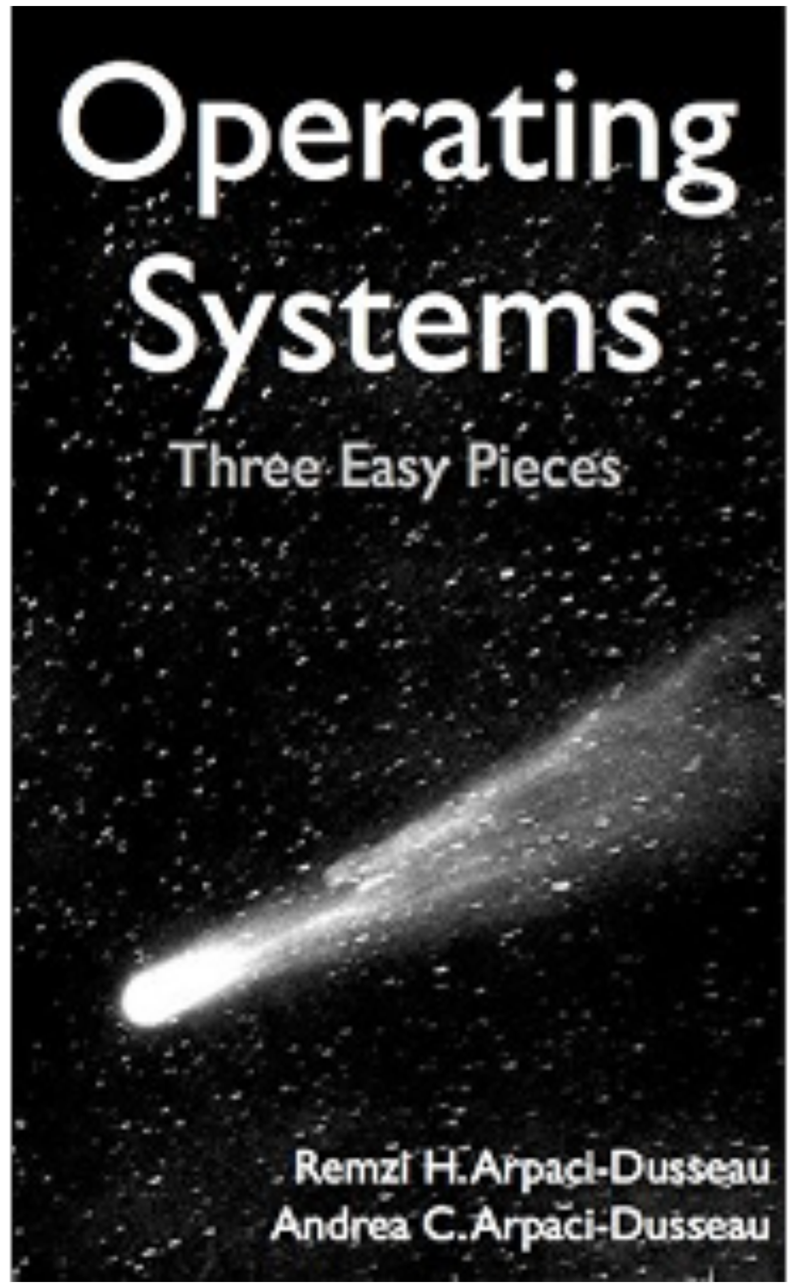
Suggested Textbooks

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

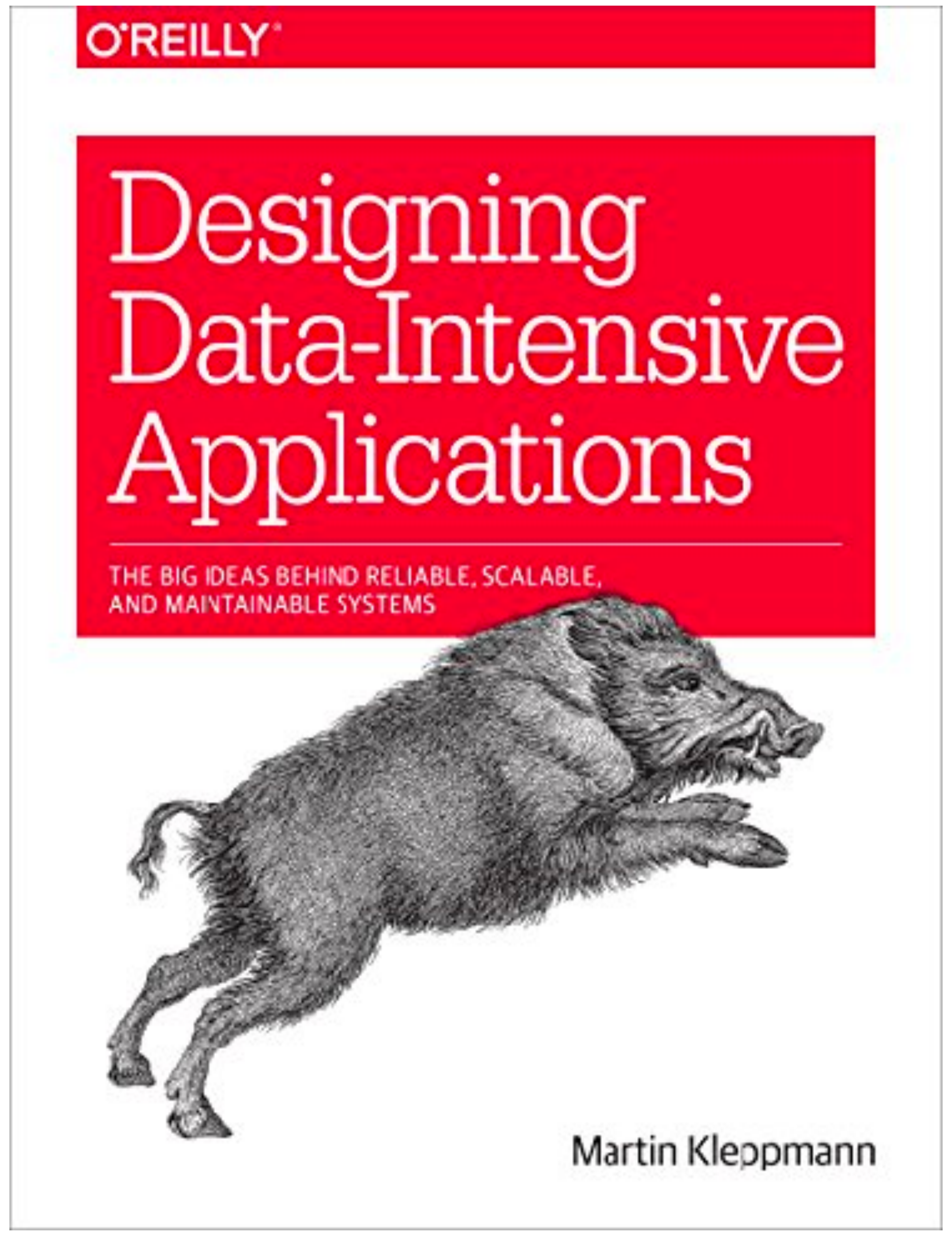

COMPUTER ORGANIZATION AND DESIGN
THE HARDWARE/SOFTWARE INTERFACE
FIFTH EDITION
DAVID A. PATTERSON & JOHN L. HENNESSY
MK



Scalable data flows

Low-level system software

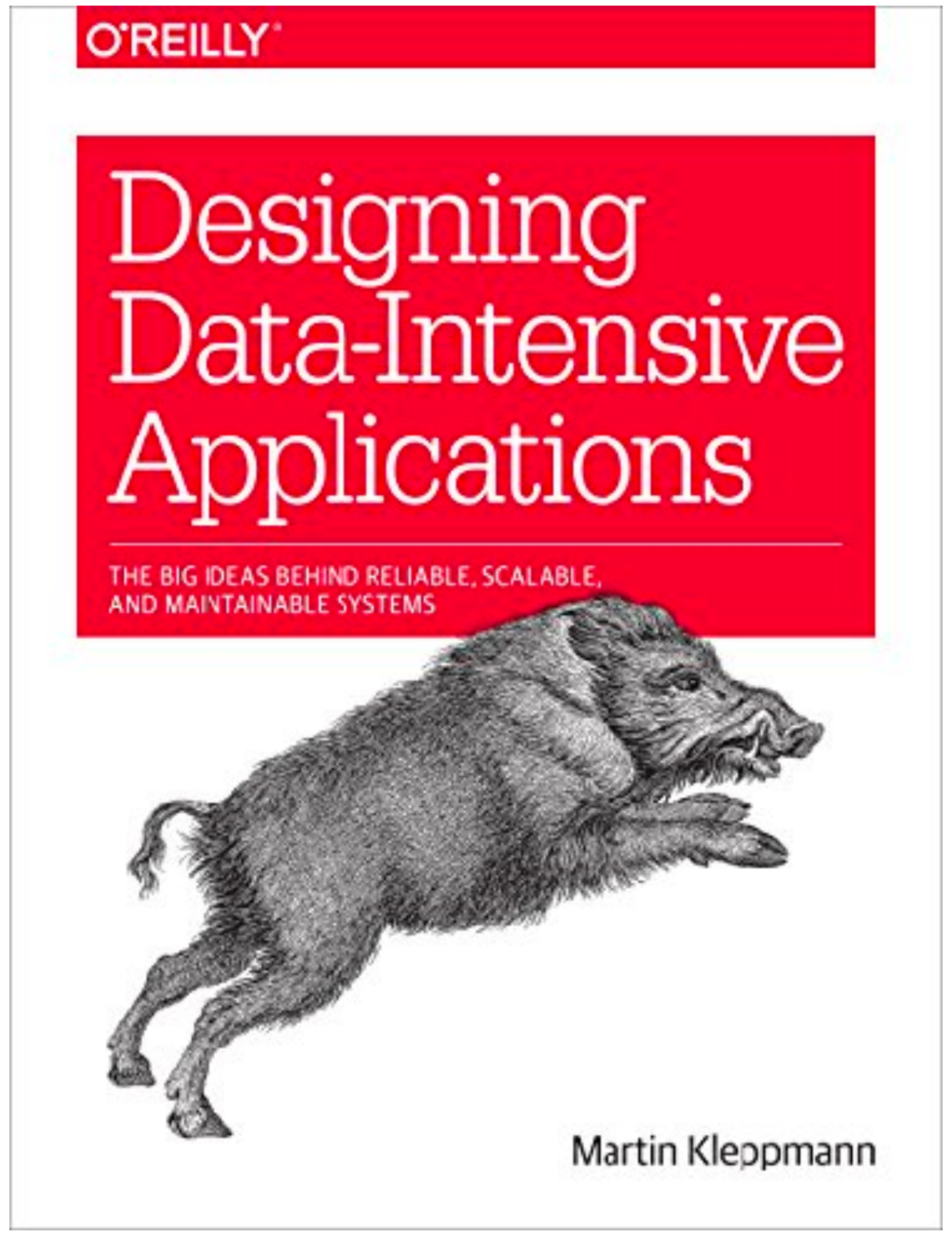
Suggested Textbooks

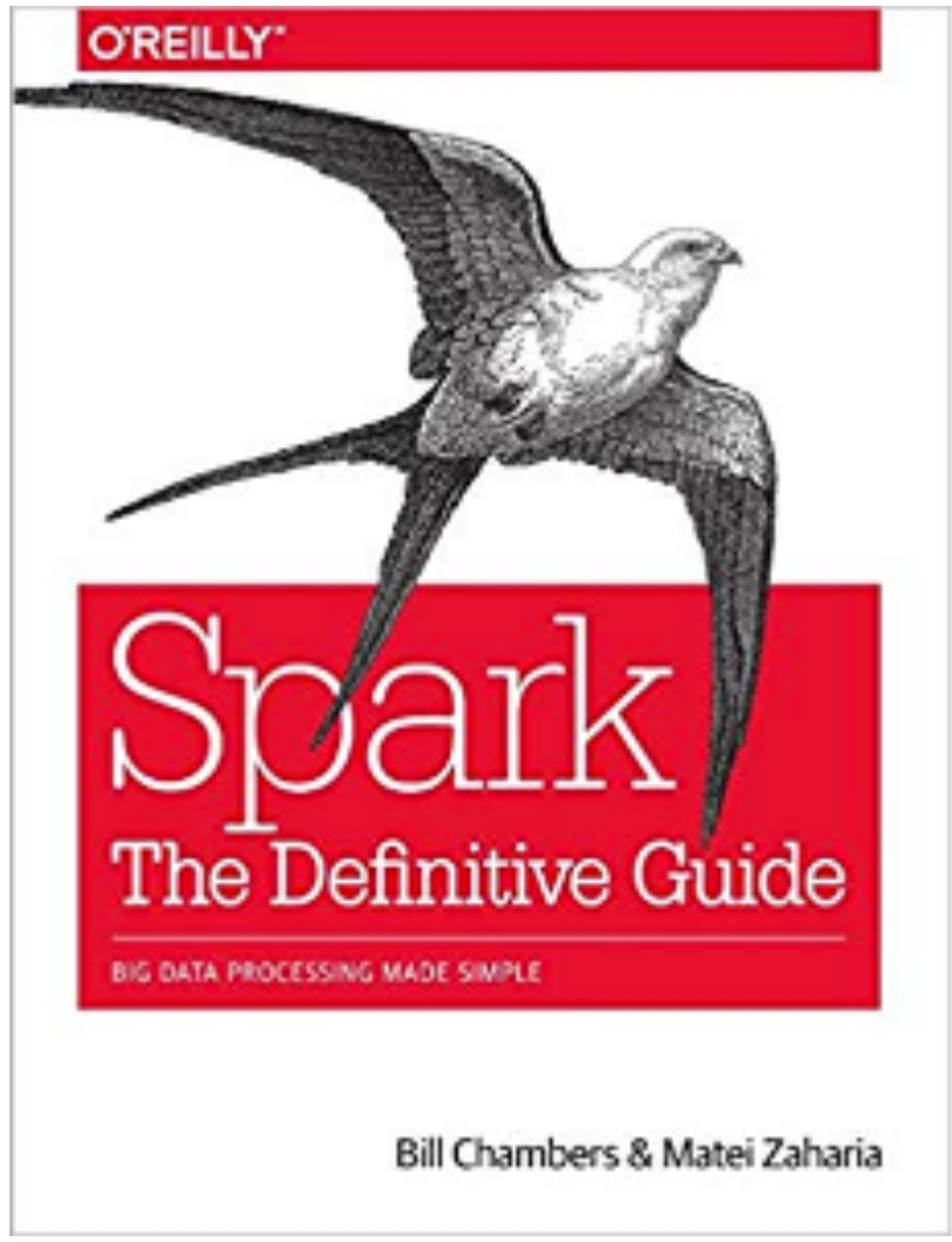

O'REILLY
Designing Data-Intensive Applications
THE BIG IDEAS BEHIND RELIABLE, SCALABLE, AND MAINTAINABLE SYSTEMS
Martin Kleppmann

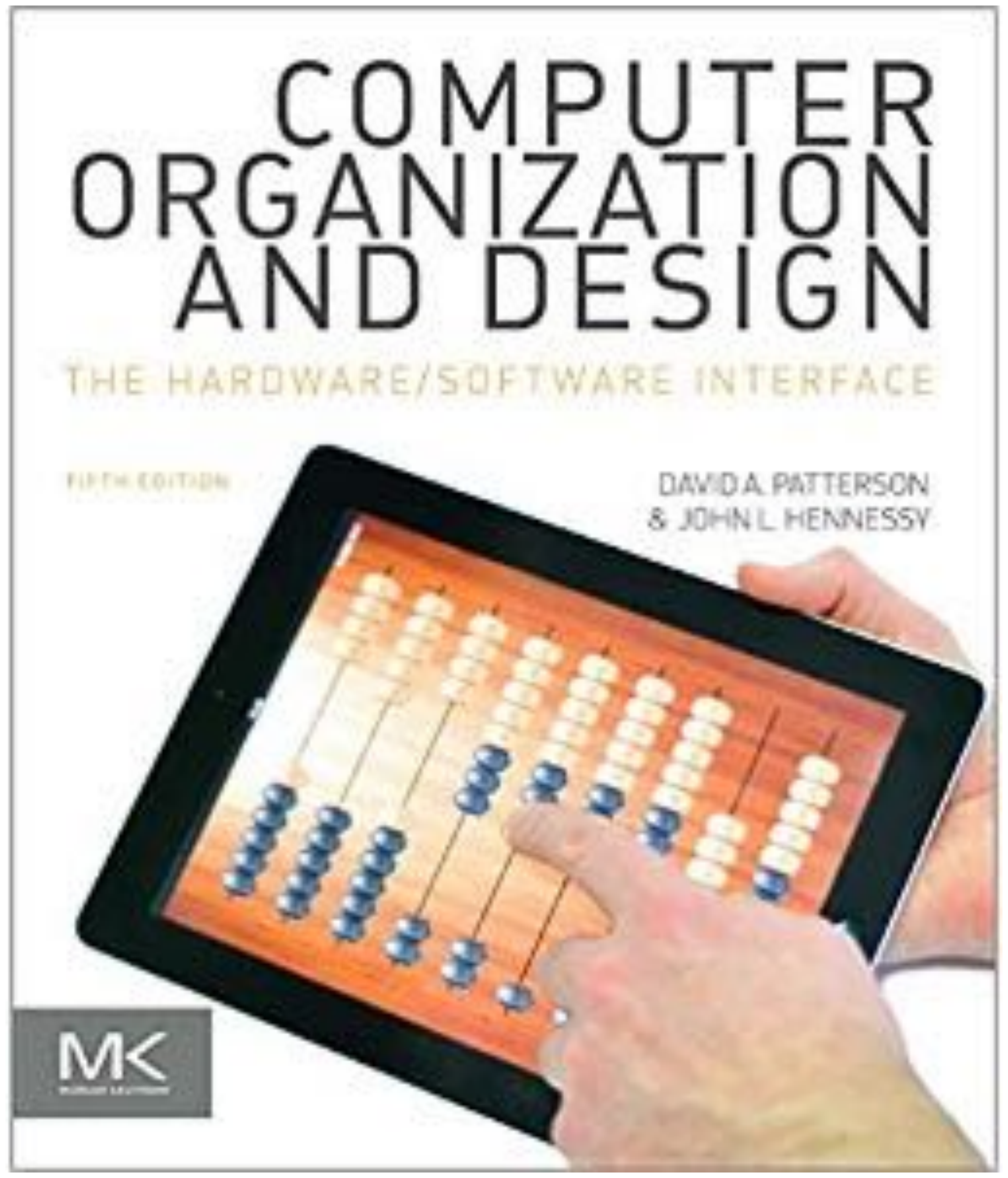
- Chapter 3. Storage and retrieval
- Chapter 4. Encoding and evolution
- Chapter 10. Batch processing
- Chapter 11. Stream processing
- Chapter 12. The future of data systems
- ~~The other chapters~~

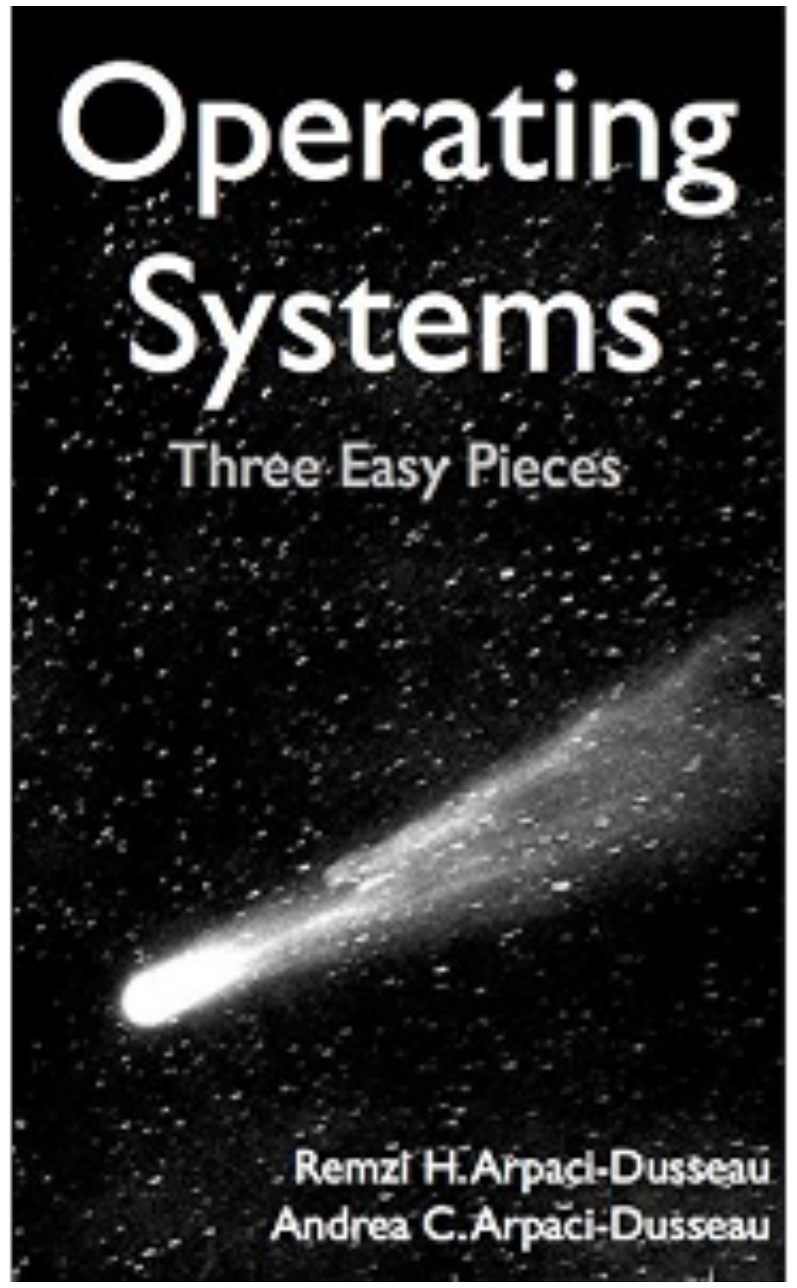
Suggested Textbooks

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



Hands on experience

Background

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?

- Scaling distributed systems
  - Cluster, cloud, edge, network, replication, partition, consistency, ACID, etc.
  - ~~RPC, Caching, Fault tolerance, Paxos, Concurrency~~

What is this course about?

- **Data Processing and Programming model.**
    - Batch processing, stream processing, MapReduce, Hadoop, Spark, Kafka, etc.

Learning outcomes of this corse

- **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 model selection on large-scale datasets.
- **Reason** critically about practical tradeoffs between accuracy, runtimes, scalability, usability, and total cost.

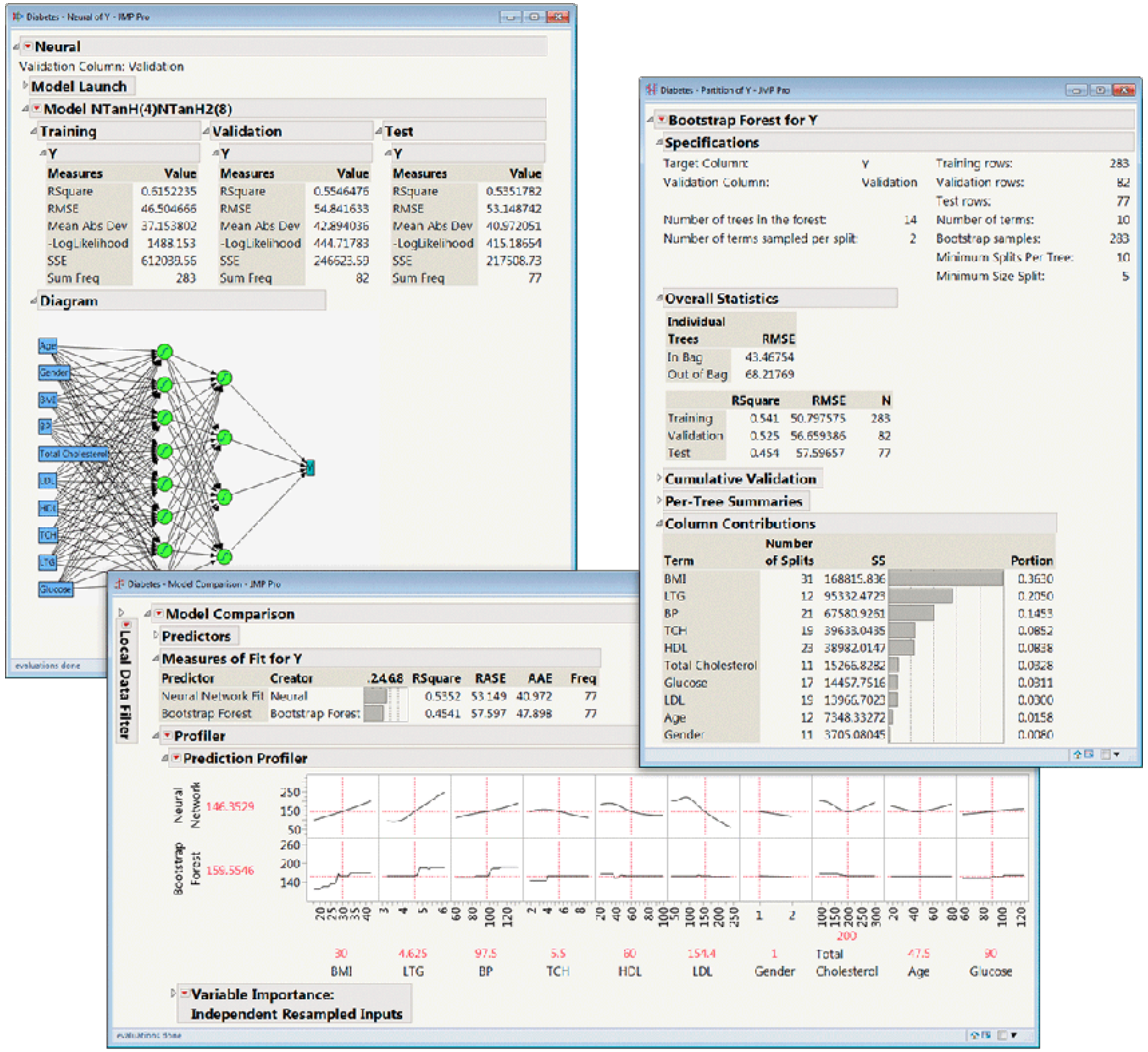
What this course is **NOT** about

- 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
  - We focus more on principles.
- If you have taken DSC 102 and look for a graduate version
  - Take DSC 204A next Winter.

Why bother learning such low-level computer sciencey stuff in Data Science?

Luxury of “Statisticians”/“Analysts”

- 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



Diabetes - Neural of Y - JMP Pro

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	246523.59
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

Diabetes - Partition of Y - JMP Pro

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

Diabetes - Model Comparison - JMP Pro

Local Data Filter

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

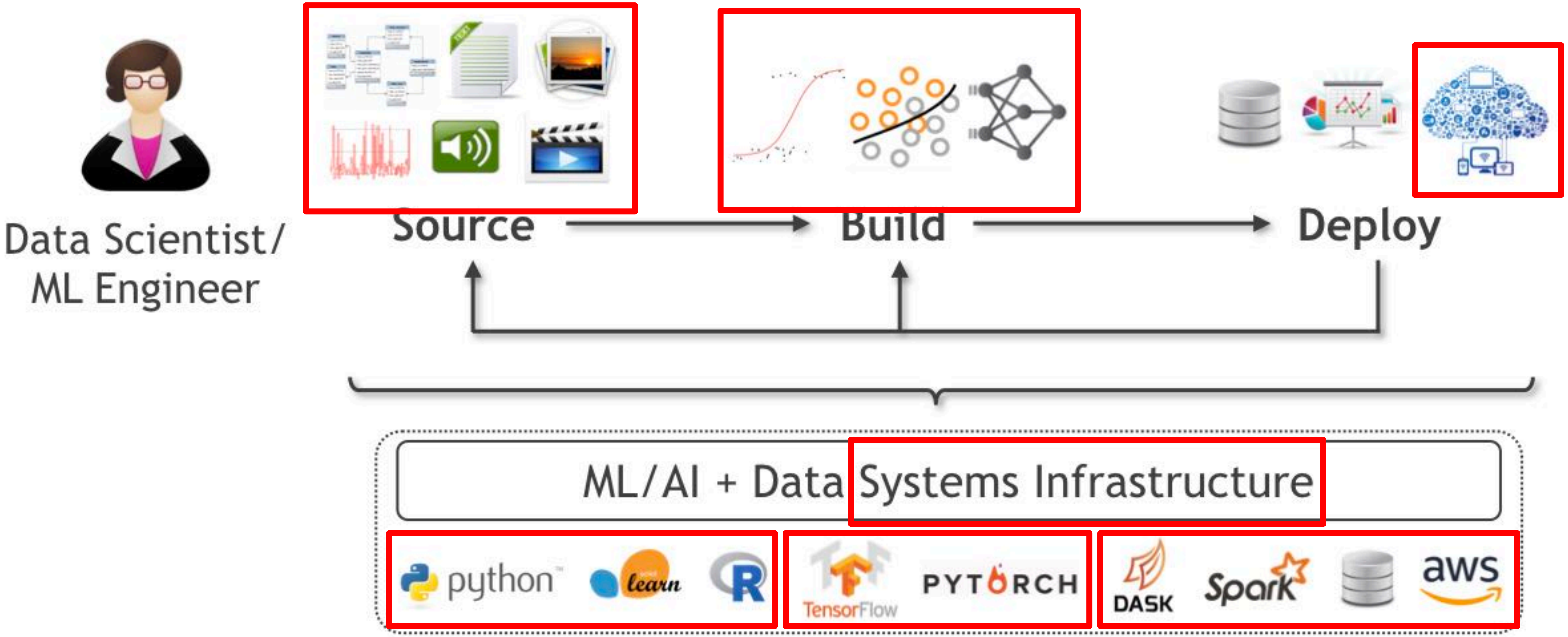
Profiler

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

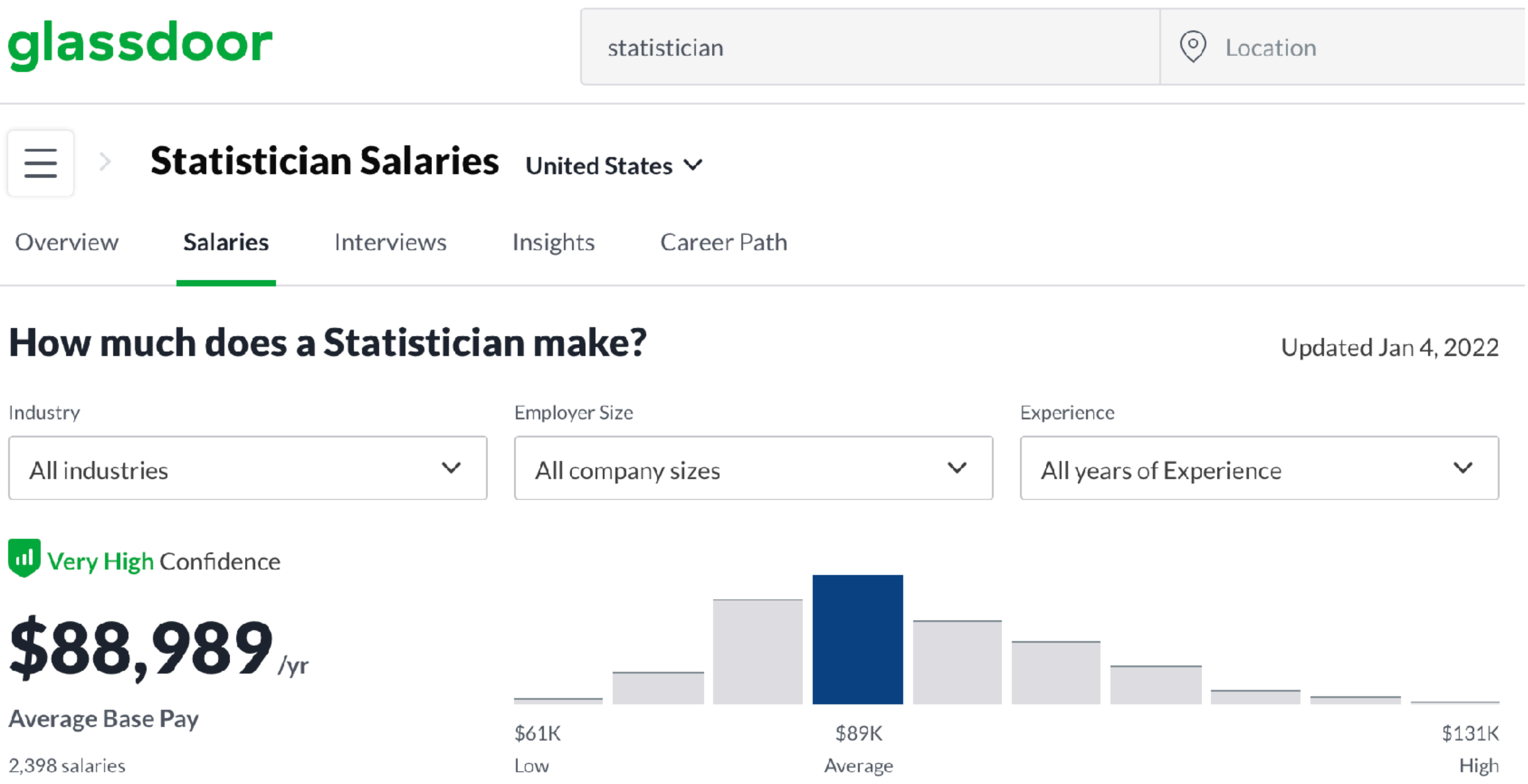
Reality of Today’s "Data Scientists”



Data acquisition
Data preparation

Feature Engineering
Training & Inference
Model Selection

Serving
Monitoring



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

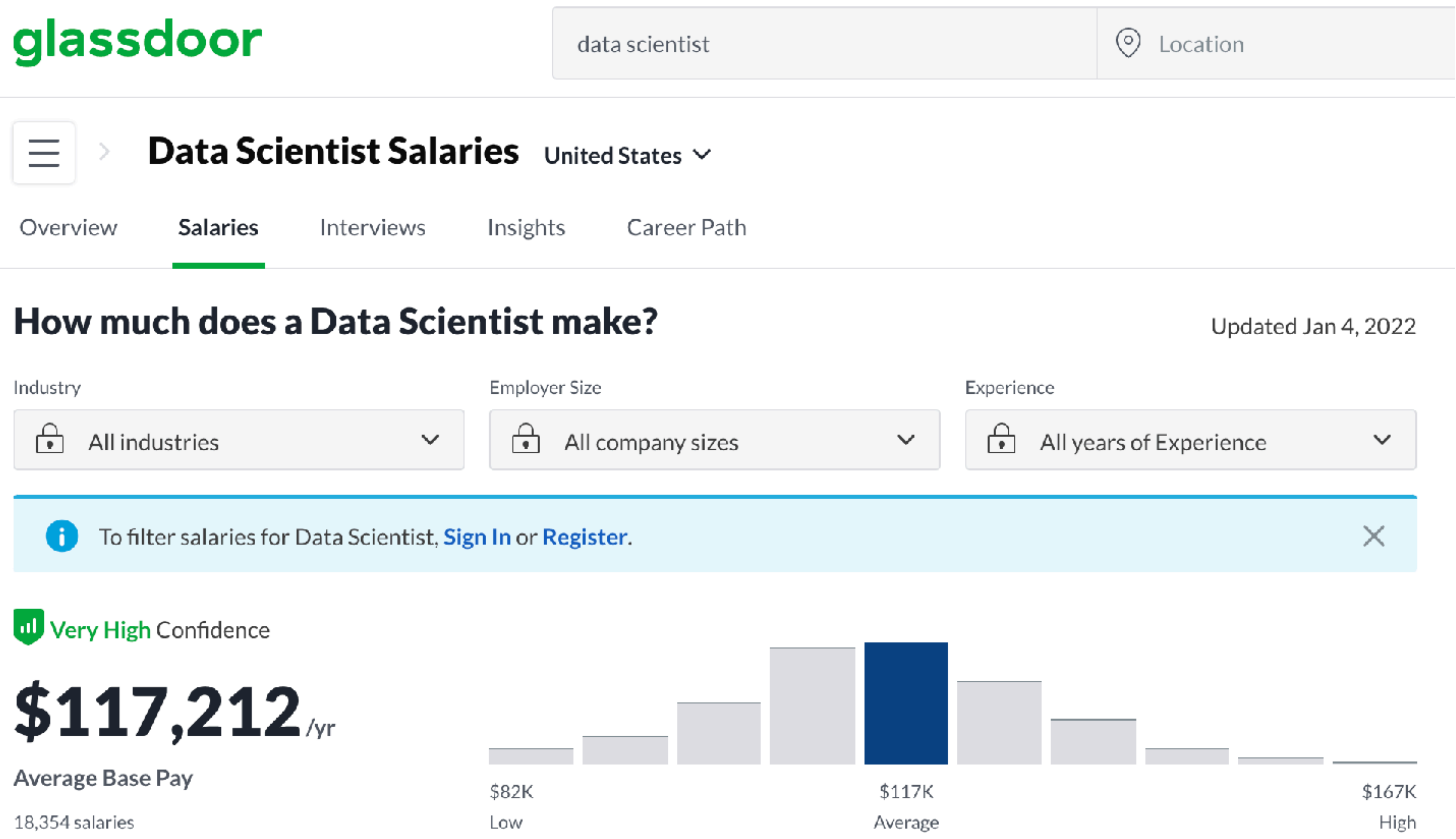
Experience: All years of Experience

Very High Confidence

$88,989 /yr

Average Base Pay

2,398 salaries



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!

Questions?

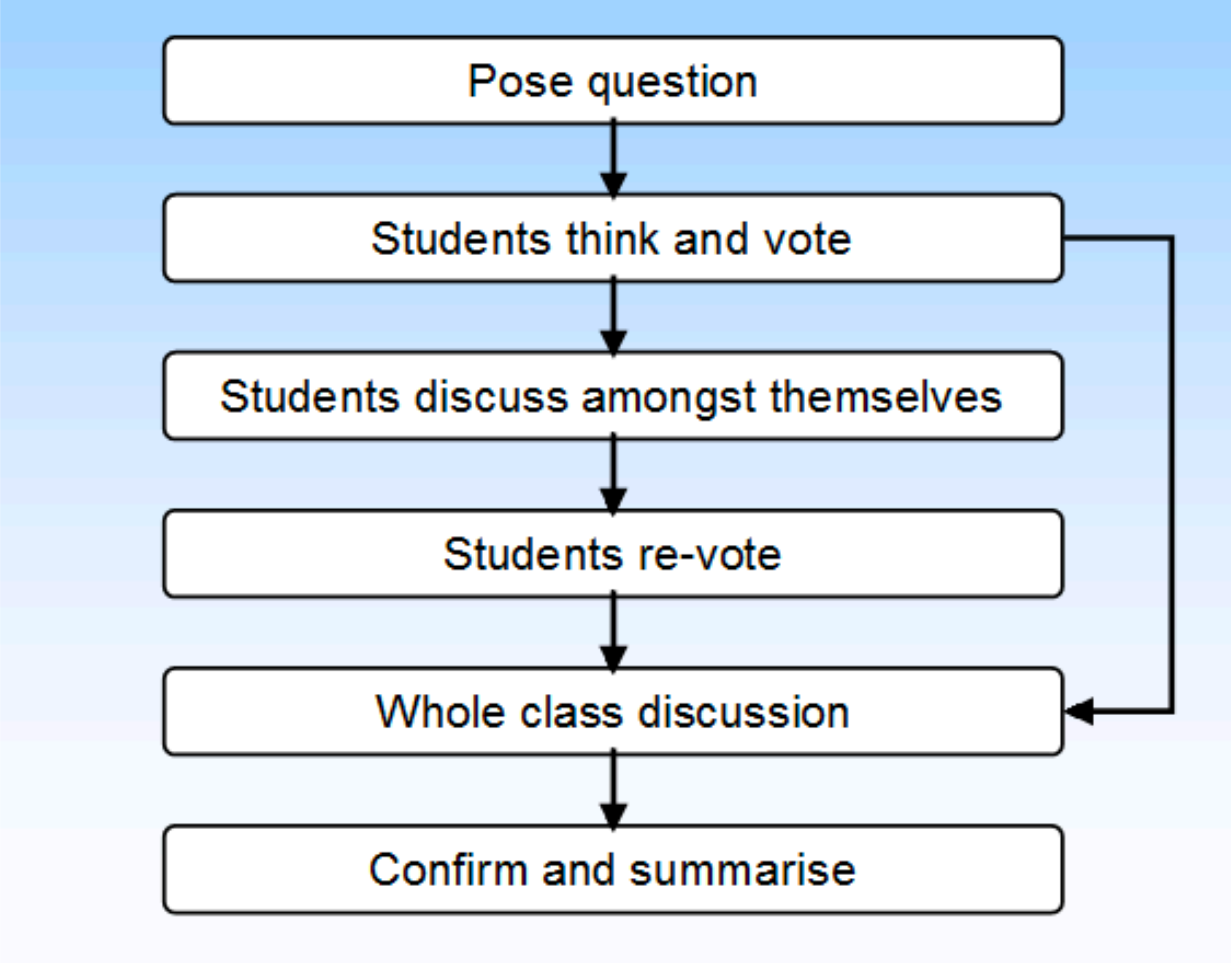
Prerequisites

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

Components and Grading

- 3 Programming Assignments: **40%** (8% + 16% + 16%)
  - No late days! Plan your work well ahead.
- Final Exam (06/14/2023 3pm-6pm): **40% ?**
- Peer Instruction Activities: **20%**
- Extra Credit Peer Evaluation Activities: **4%** (likely)

Peer instruction activity



Pose question

Students think and vote

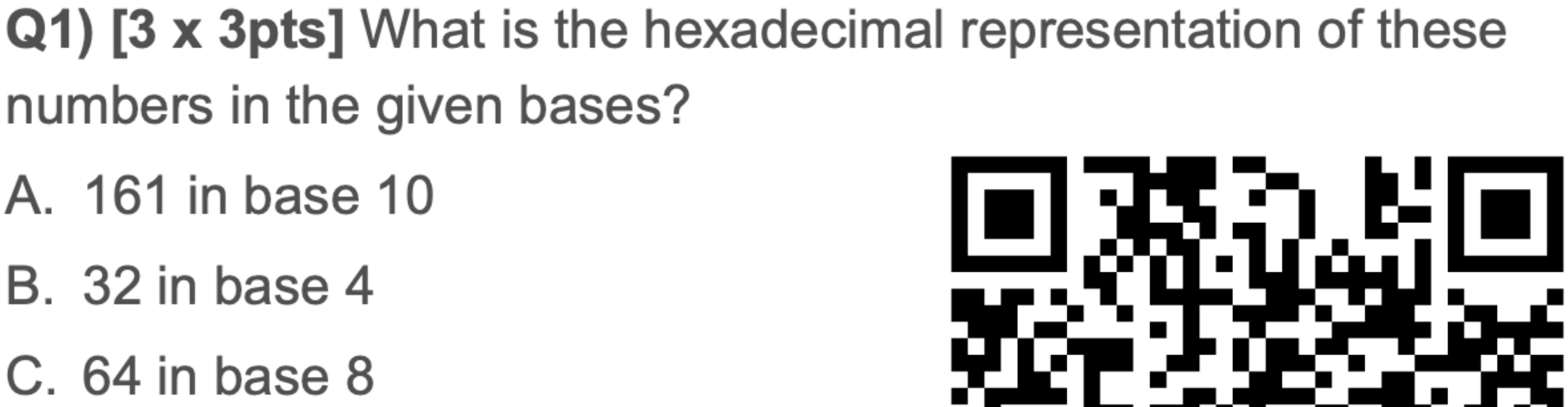
Students discuss amongst themselves

Students re-vote

Whole class discussion

Confirm and summarise

Example flow



Q1) [3 x 3pts] What is the hexadecimal representation of these numbers in the given bases?

A. 161 in base 10

B. 32 in base 4

C. 64 in base 8

Answers

Q1) [3 x 3pts] What is the hexadecimal representation of these numbers in the given bases?

A. 161 in base 10

B. 32 in base 4

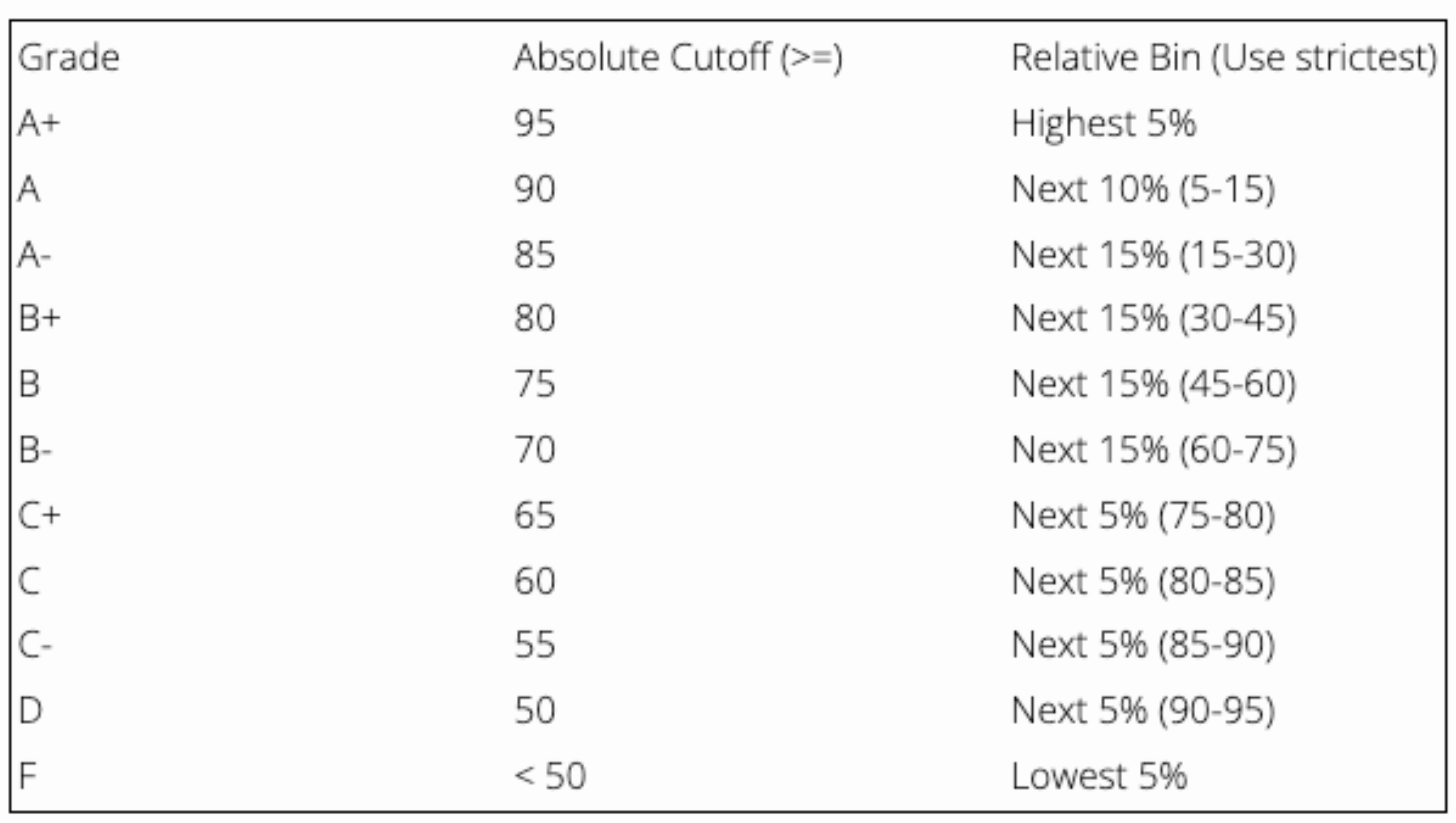
C. 64 in base 8

A. A1$_{16}$ (aka 0xA1)

B. 0xE

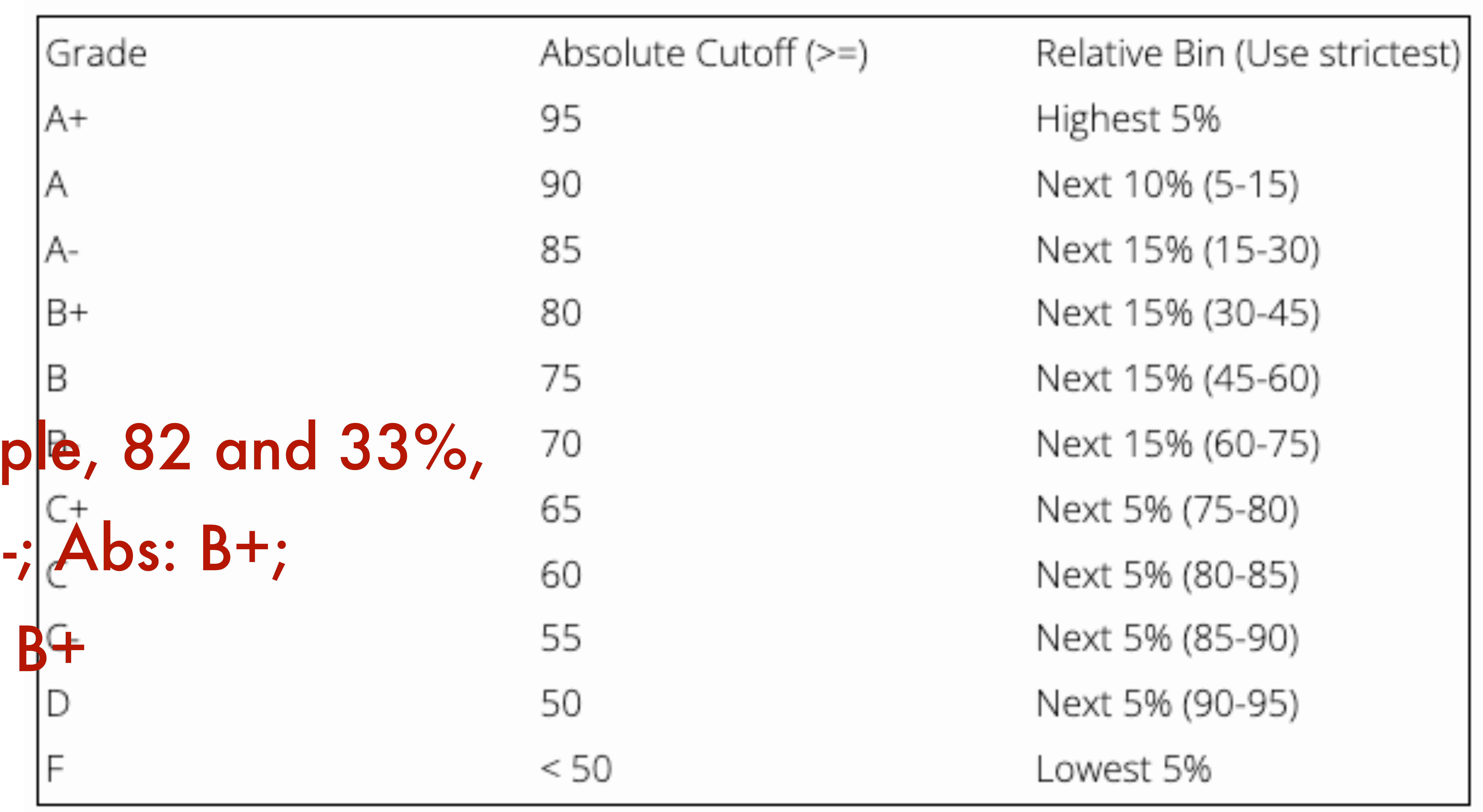
C. 0x34

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 1-3	Foundations of Data Systems	Single Machine: CompOrg -> OS -> Cloud
Week 4-6	Scaling Distributed Systems	Multiple Machine: Storage -> Network
Week 7-10	Data Processing and Programming model	Processing: Batch -> Stream -> Cloud

https://haojian.github.io/DSC204A23WI/

Programming Assignments

- PA0: Setting up AWS and Dask
  - Apr 10 to Apr 25
- PA1: Data Exploration with Dask
  - Apr 26 to May 10
- PA2: Feature Eng. and Model Selection with Spark
  - May 11 to June 2
- 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
- I will cover the concepts and tools’ tradeoffs in the lectures
- 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

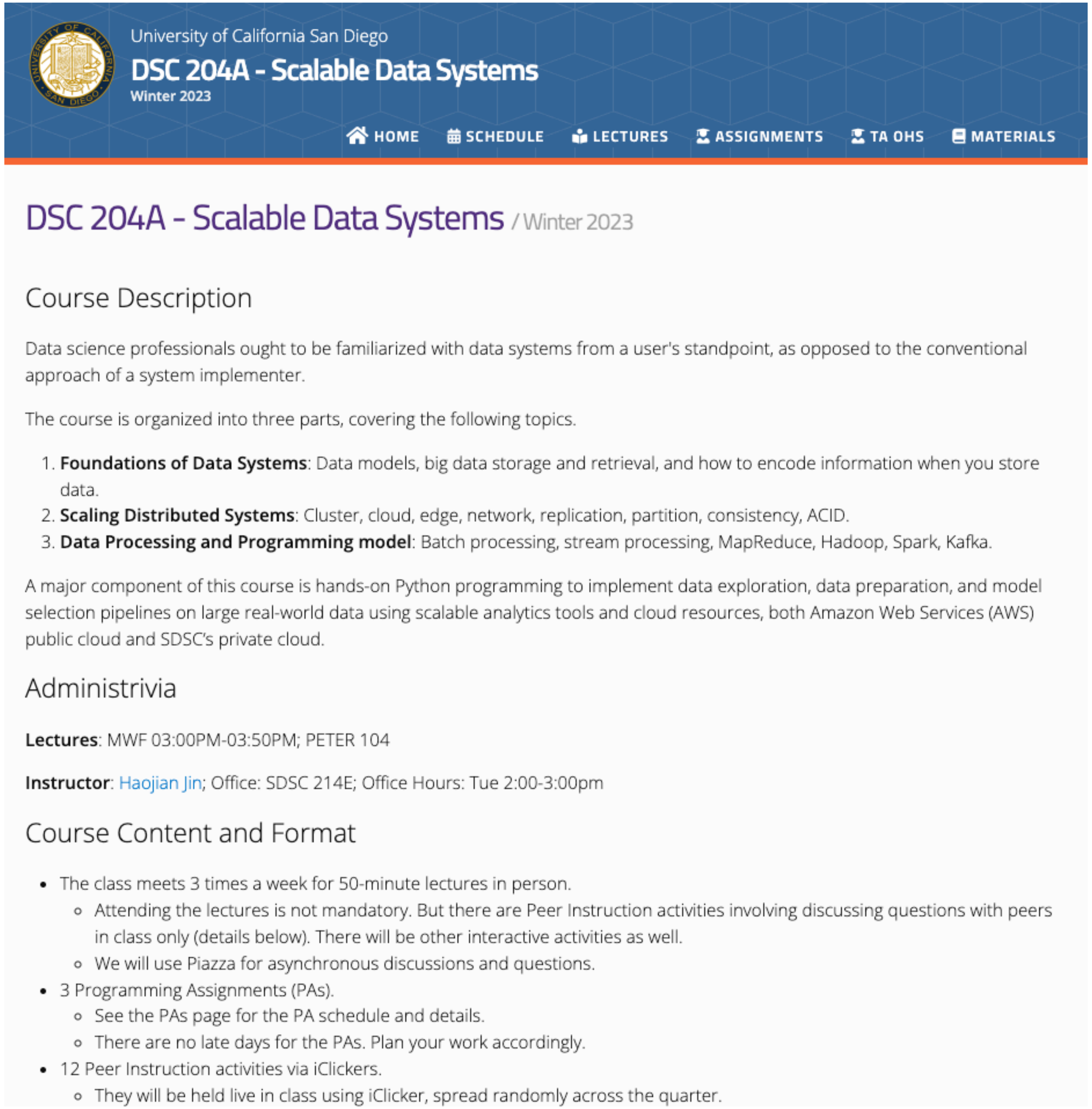
- Office hours are for getting ideas on how to debug or better approach your homework.
- Write a description! Try to narrow down your problem area as much as possible.
- If you don’t have a description, TA can reject your questions.
- Respect TA’s working hours.
  - Respond in 24 hours.
  - Members may send msgs at night or on weekends, but only expect to receive a reply on weekday.

Tentative plan

- Rohit
  - Tuesday 1:30 PM - 2:30 PM.
- Megha
  - Thursday TBD.
- Location:
  - CSE building or HDSI building?

Course administrivia

https://haojian.github.io/DSC204A23WI/



University of California San Diego
DSC 204A - Scalable Data Systems
Winter 2023

DSC 204A - Scalable Data Systems / Winter 2023

Course Description

Data science professionals ought to be familiarized with data systems from a user's standpoint, as opposed to the conventional approach of a system implementer.

The course is organized into three parts, covering the following topics.

1. **Foundations of Data Systems**: Data models, big data storage and retrieval, and how to encode information when you store data.
2. **Scaling Distributed Systems**: Cluster, cloud, edge, network, replication, partition, consistency, ACID.
3. **Data Processing and Programming model**: Batch processing, stream processing, MapReduce, Hadoop, Spark, Kafka.

A major component of this course is hands-on Python programming to implement data exploration, data preparation, and model selection pipelines on large real-world data using scalable analytics tools and cloud resources, both Amazon Web Services (AWS) public cloud and SDSC's private cloud.

Administrivia

Lectures: MWF 03:00PM-03:50PM; PETER 104

Instructor: Haojian Jin; Office: SDSC 214E; Office Hours: Tue 2:00-3:00pm

Course Content and Format

- The class meets 3 times a week for 50-minute lectures in person.
  - Attending the lectures is not mandatory. But there are Peer Instruction activities involving discussing questions with peers in class only (details below). There will be other interactive activities as well.
  - We will use Piazza for asynchronous discussions and questions.
- 3 Programming Assignments (PAs).
  - See the PAs page for the PA schedule and details.
  - There are no late days for the PAs. Plan your work accordingly.
- 12 Peer Instruction activities via iClickers.
  - They will be held live in class using iClicker, spread randomly across the quarter.

Course administrivia

- Lectures: MWF 3pm-3:50pm PT at PETER 104
- Instructor: Haojian Jin; haojian@ucsd.edu
  - OHs: TBD. See website.
- TAs: Rohit Ramaprasad; Megha Agarwal
  - TA hours see the course web site.
- Slack for all communications (also see Canvas).
- Canvas for PA submission, Peer Evaluation Activities, Grading.

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 Slack 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 am very strict on this matter!

Questions?