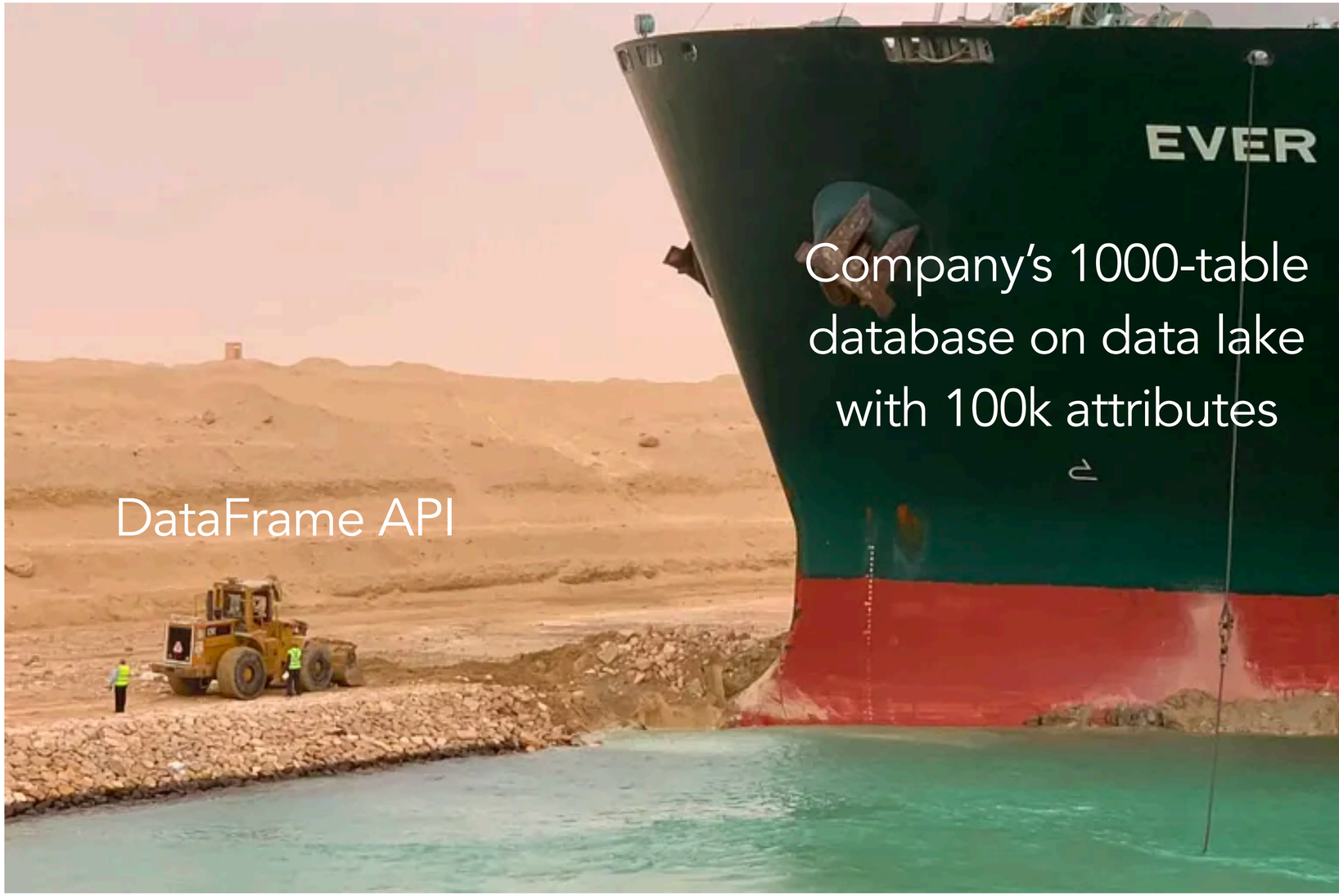
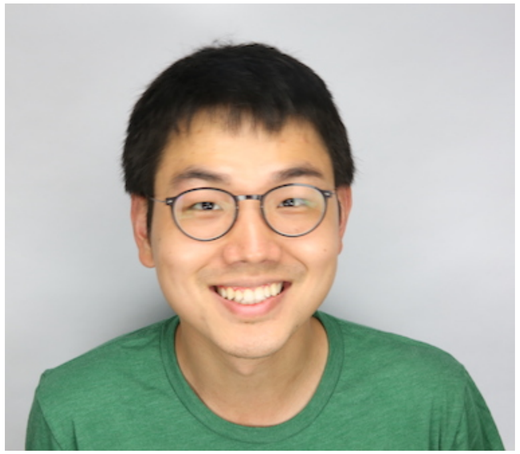
DSC 204a Scalable Data Systems

- Haojian Jin



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

Bio



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

Asst. Prof @ UCSD-HDSI

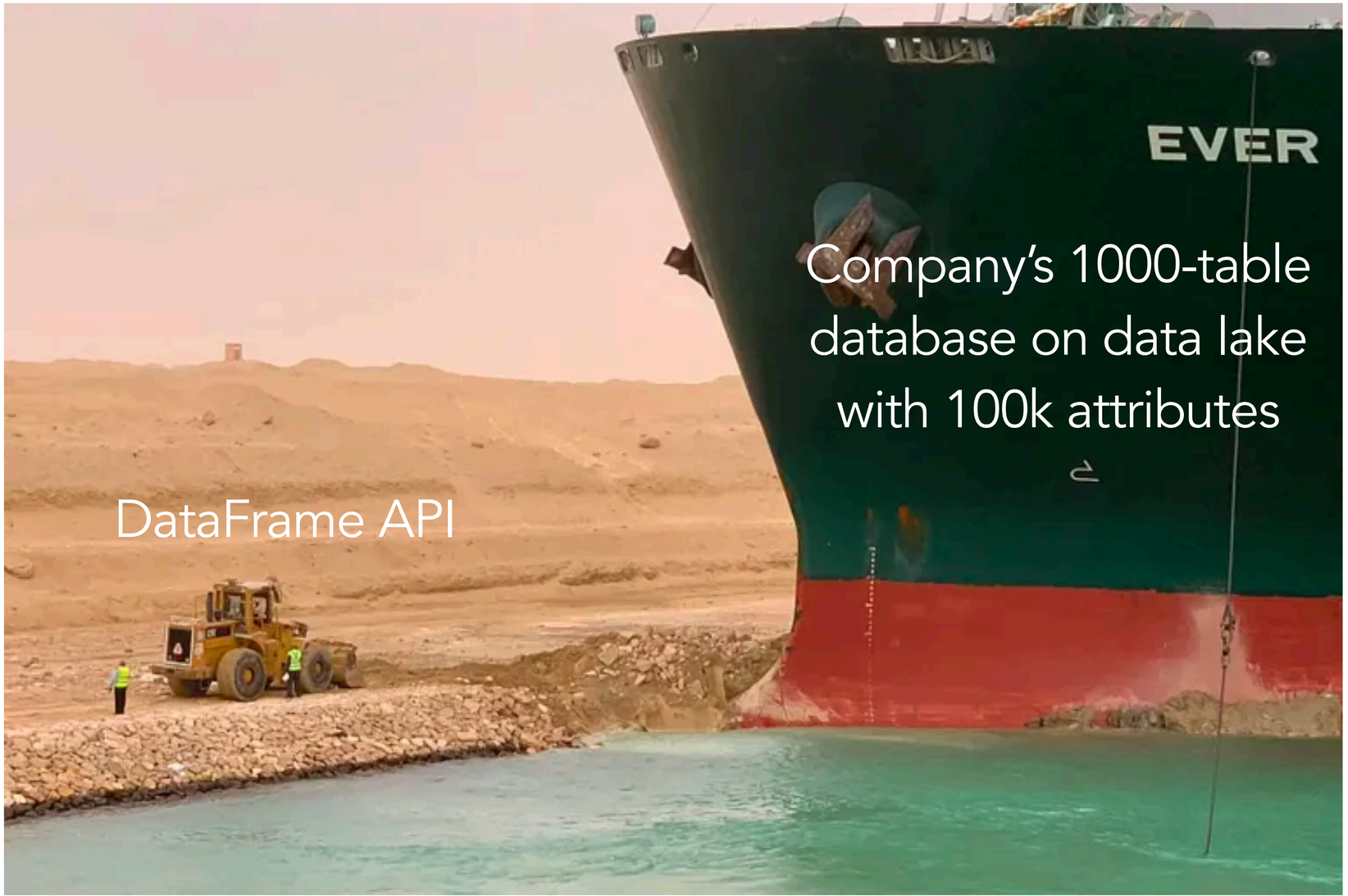
Data Smith Lab:

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

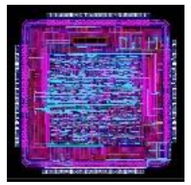
Before Ph.D.: worked at Yahoo Research, ran a startup

What is this course about?

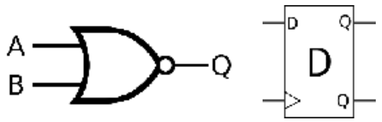


Levels of Abstraction

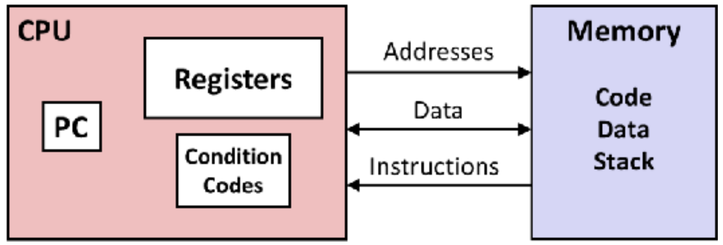
Computer Designer



Gates, clocks, circuit layout, …



Assembly programmer



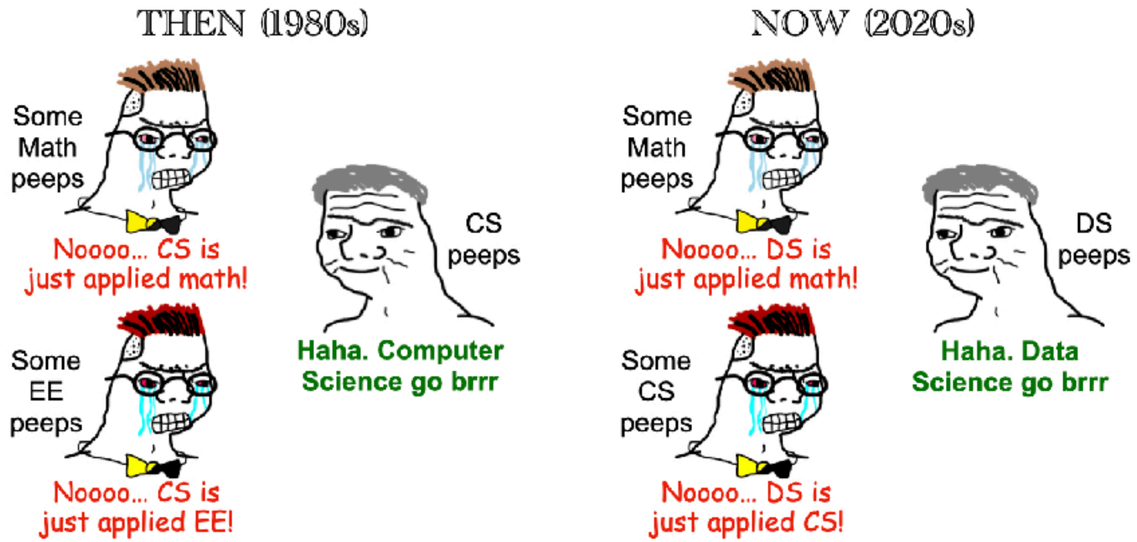
C programmer

```
#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**.



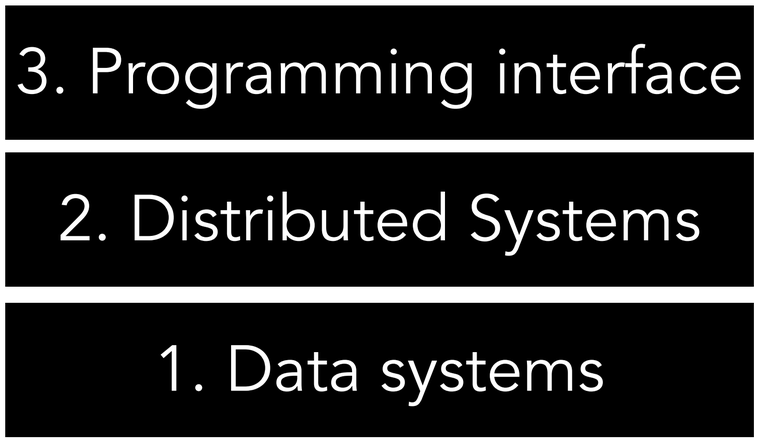
What is this course about?

- 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.



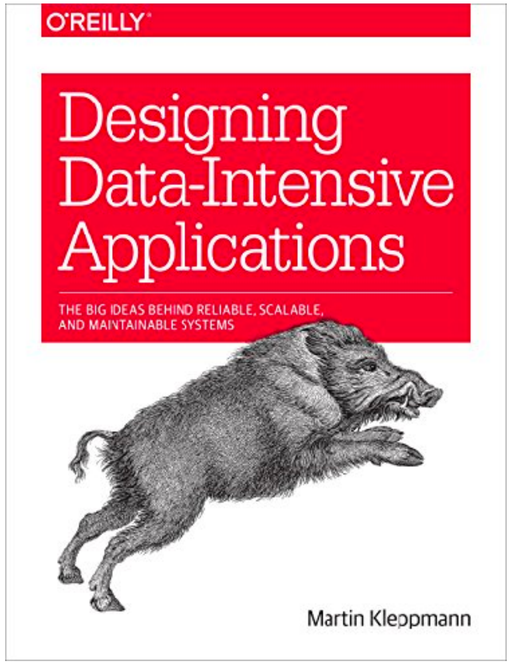
What is this course about?

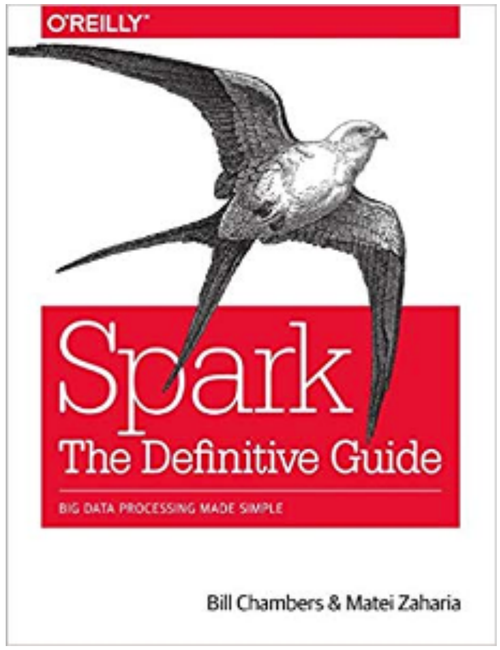
The data sci relevant components in the following course

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

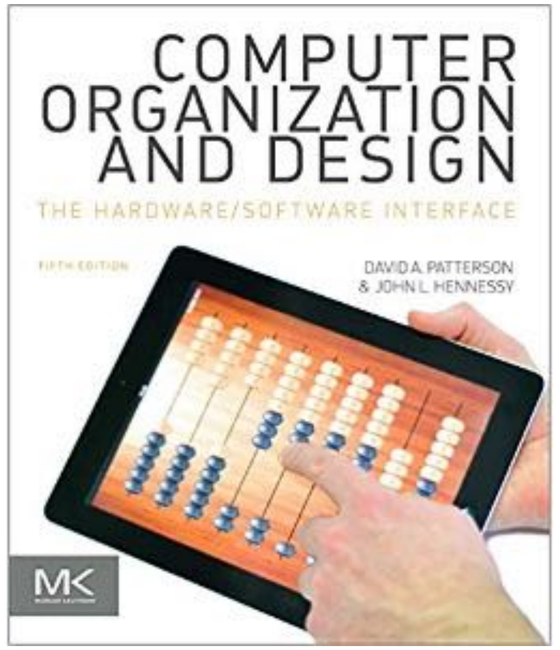
Suggested Textbooks

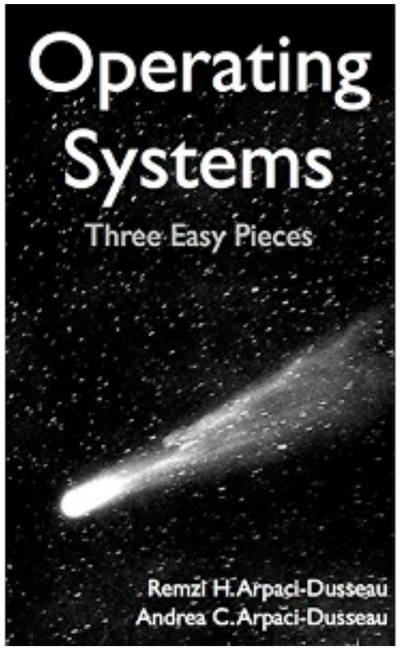
Computer systems are about carefully layering levels of abstraction.





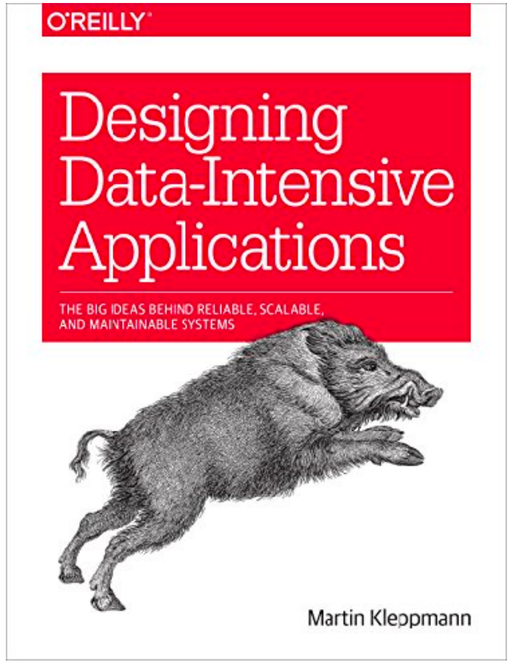
Scalable data flows





Low-level system software

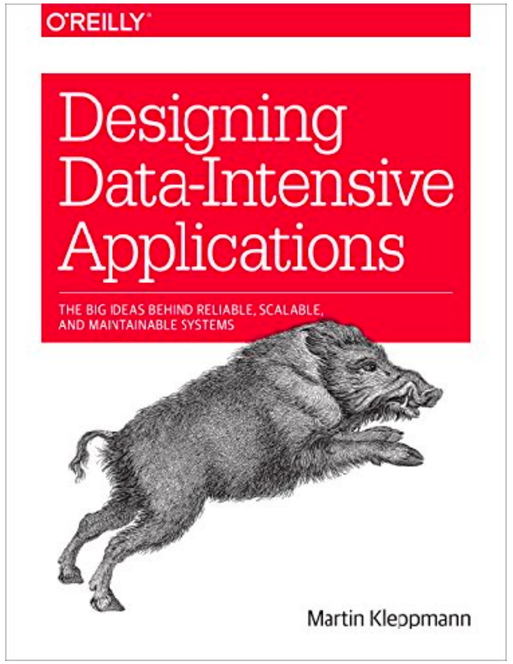
Suggested Textbooks

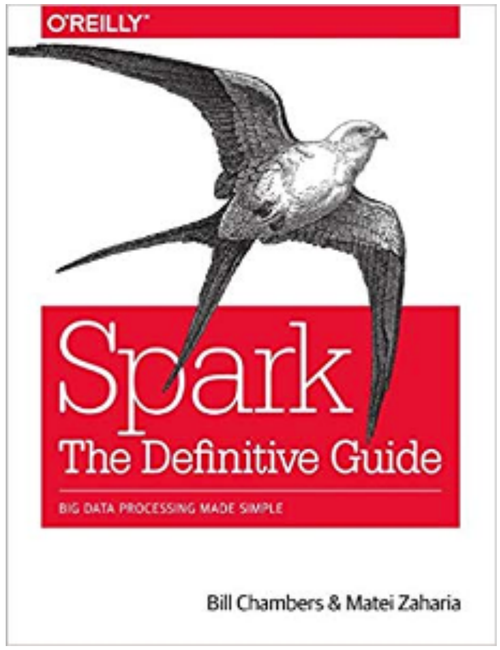


- Chapter 3. Storage and retrieval
- Chapter 4. Encoding and evolution
- Chapter 10. Batch processing
- Chapter 11. Stream processing
- Chapter 12. The future of data systems
- ~~The other chapters~~

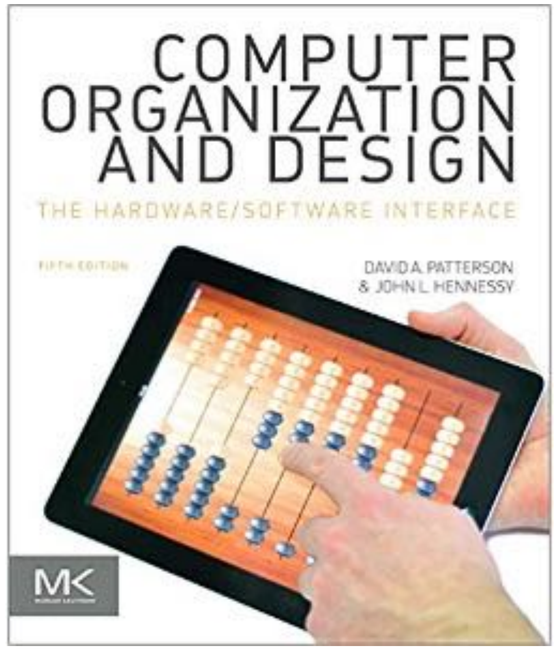
Suggested Textbooks

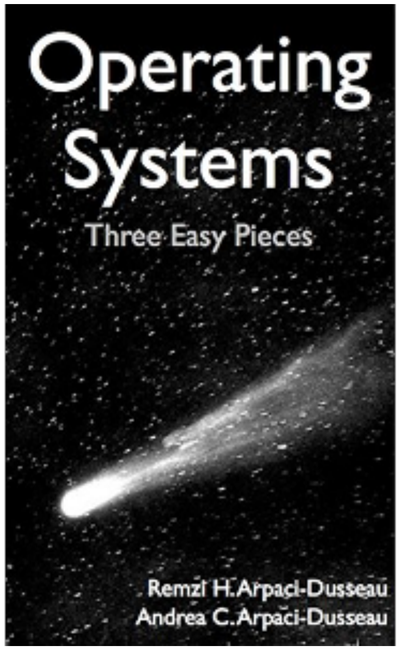
Computer systems are about carefully layering levels of abstraction.





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

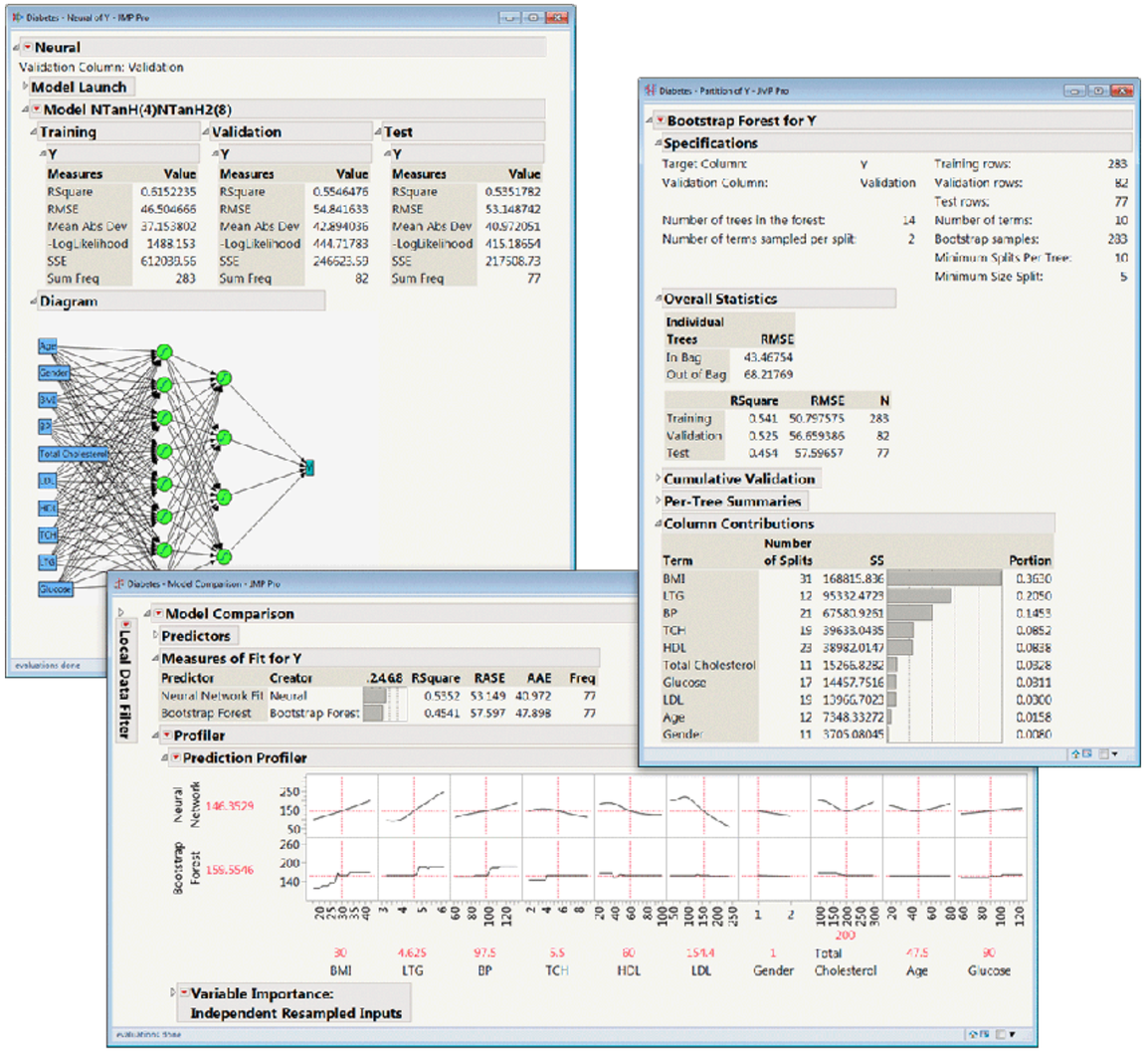
- Not a course on database, relational model, or SQL
  - Take DSC 202 instead (pre-requisite)
- Not a course on how to build scalable data systems
  - Take Distributed Systems, Operating Systems, Cloud Computing,

    …
- 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?

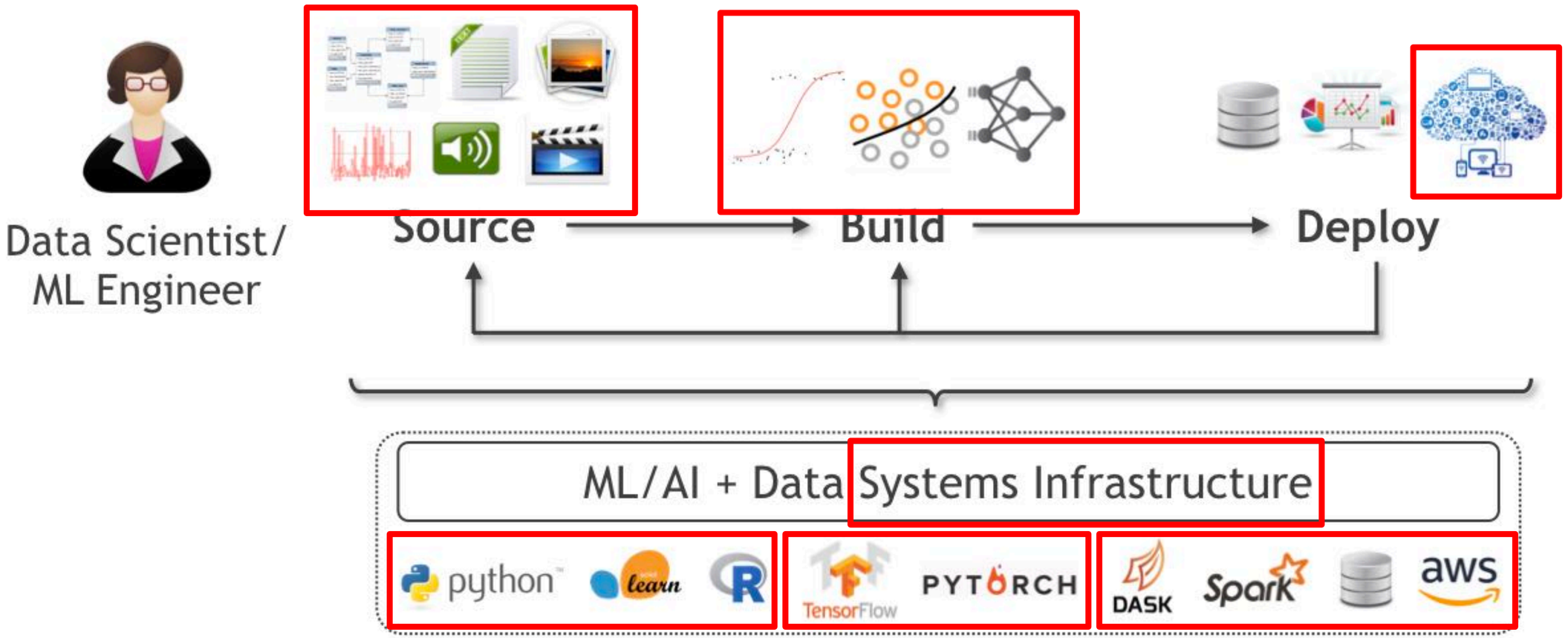
Luxury of "Statisticians"/"Analysts"

- 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



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

$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

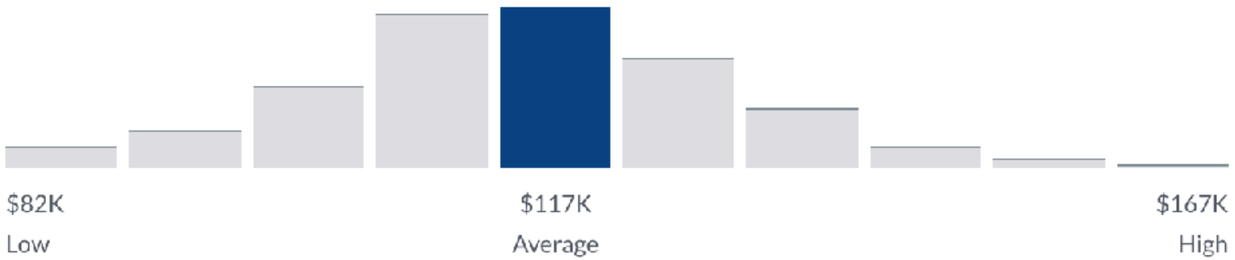
To filter salaries for Data Scientist, Sign In or Register.

Very High Confidence

$117,212 /yr

Average Base Pay

18,354 salaries



— 88,989

= 28,223!

Questions?

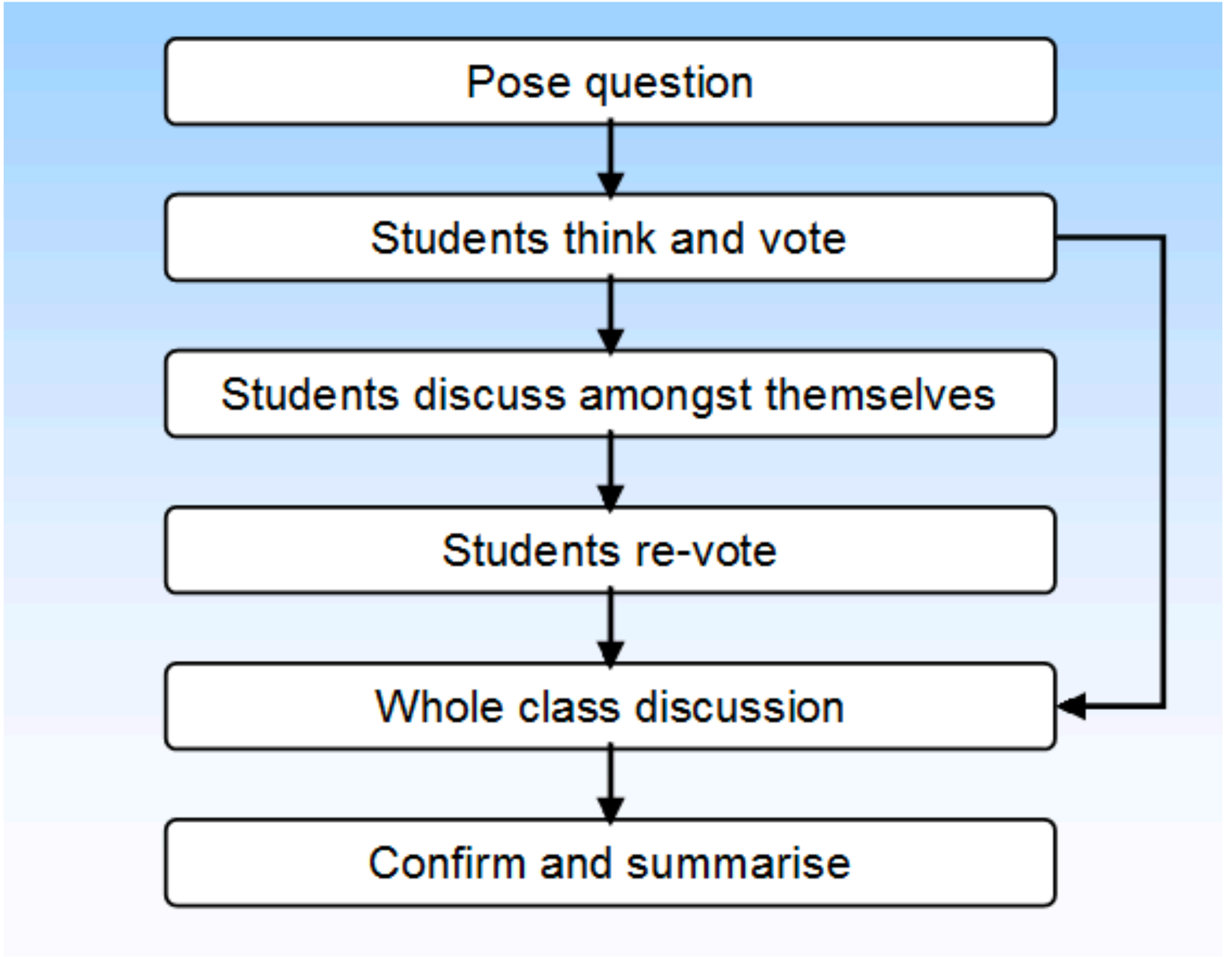
Prerequisites

- 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



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

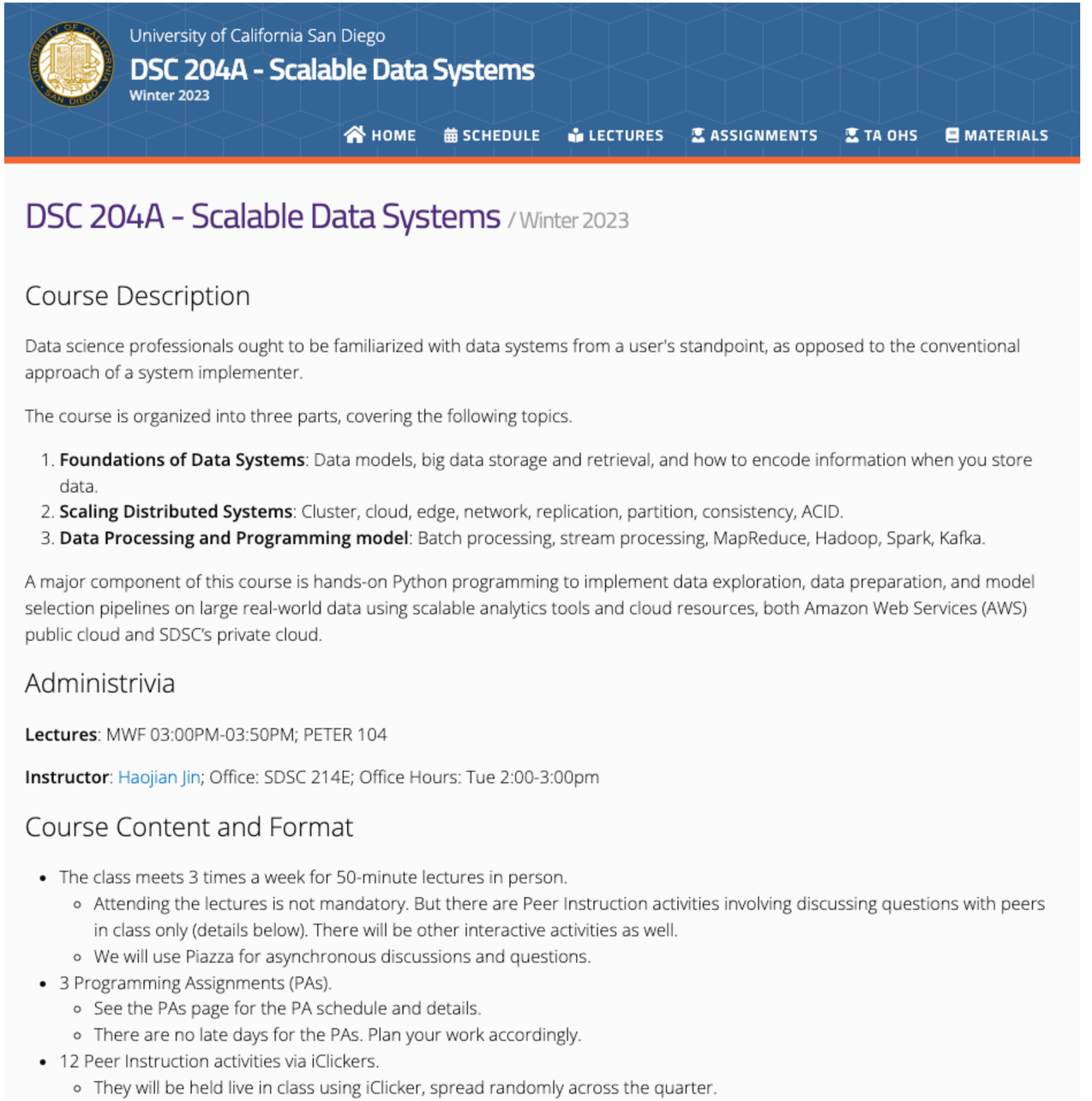
- Office hours are for getting ideas on how to debug or better approach your homework.
- Write a description! Try to narrow down your problem area as much as possible.
- If you don't have a description, TA can reject your questions.
- Respect TA's working hours.
  - Respond in 24 hours.
  - Members may 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/



Course administrivia

- Lectures: MWF 3pm-3:50pm PT at PETER 104
- Instructor: Haojian Jin; haojian@ucsd.edu
  - OHs: Tue 2-3pm PT at SDSC 214E
- 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?