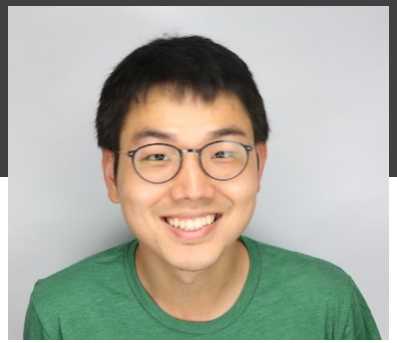
UC San Diego

DSC 102
Systems for Scalable Analytics

Fall 2024

Haojian Jin

About Myself

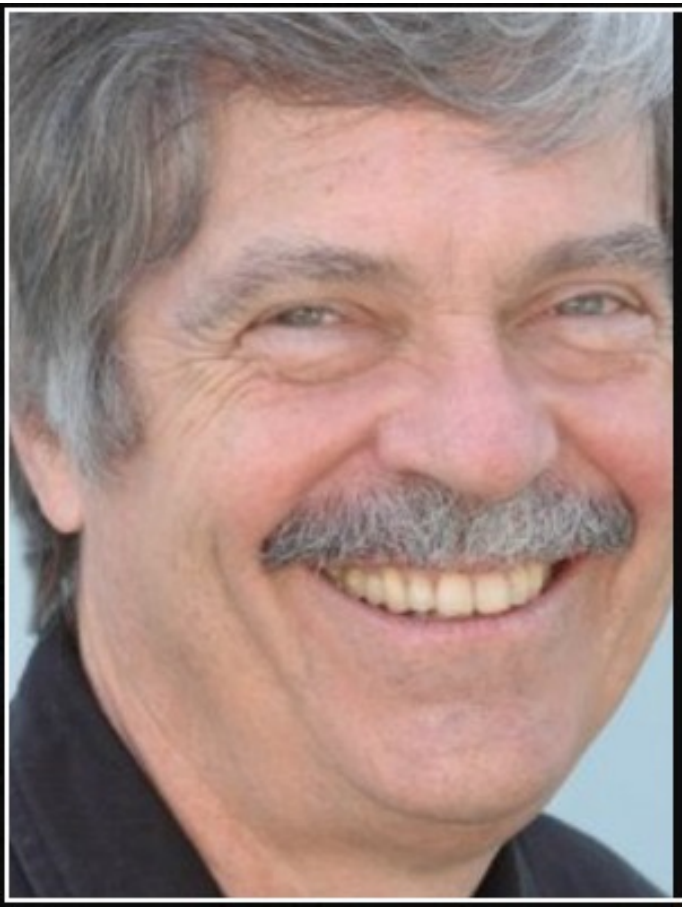


Haojian Jin (http://haojianj.in/)
Asst. Prof @ UCSD-HDSI

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
Ph.D. Thesis: Modular Privacy Flow

Before Ph.D.: worked at Yahoo Research, ran a startup
HCI, Software Engineering, Mobile Computing, AI.

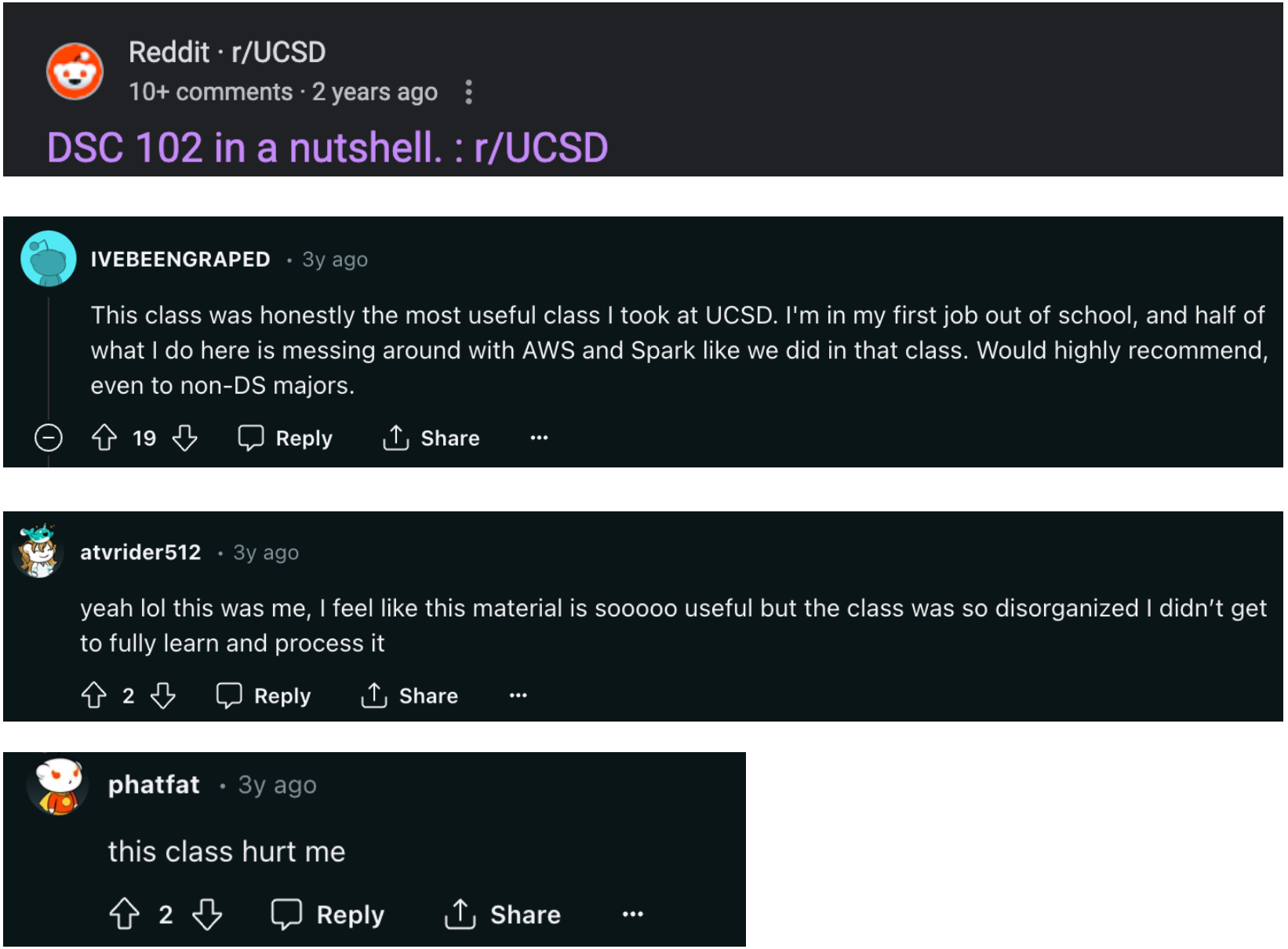


The best way to predict the future is to invent it.

— Alan Kay —

AZ QUOTES

What is this course about? Why take it?



https://www.reddit.com/r/UCSD/comments/npqcdm/dsc_102_in_a_nutshell/

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

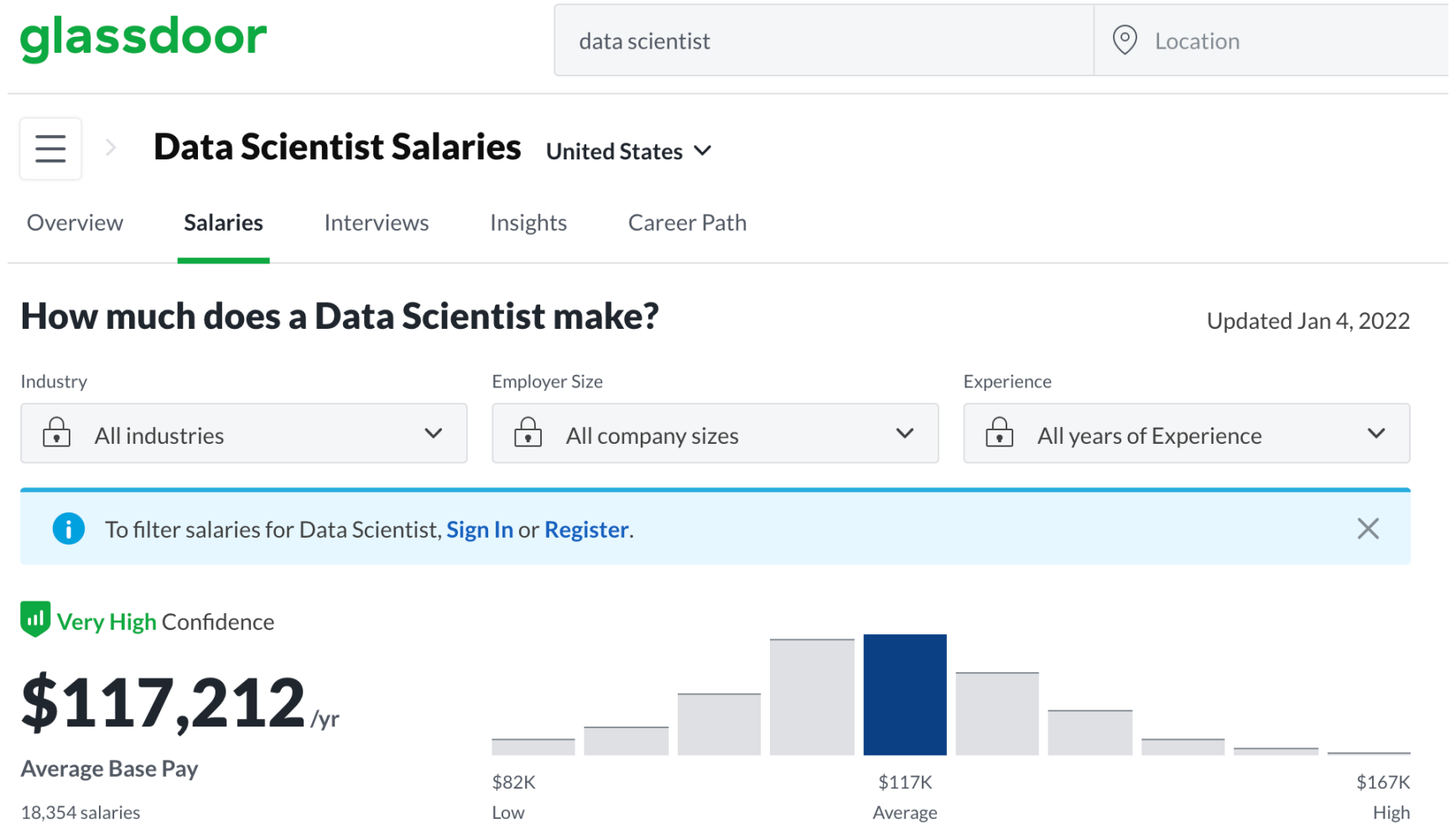
Very High Confidence

$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
$82K
Low
$117K
Average
$167K
High

— 88,989

= 28,223!

Software systems for data analytics and ML over large and complex datasets are now critical for digital applications in many domains

The Age of “Big Data”/“Data Science”

The New York Times

SundayReview | NEWS ANALYSIS

The Age

By STEVE LOHR

Email

Share

Tweet

Save

Forbes / Entrepreneurs

Forbes

MAR 25, 2015 @ 7:33 PM 4,407 VIEWS

Drowning In Big Data - Finding Insight In A Digital

Josh Steimle, CONT

For roughly a deca
information about
Big Data. The IDC
industry will exper
by 2018. What this

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

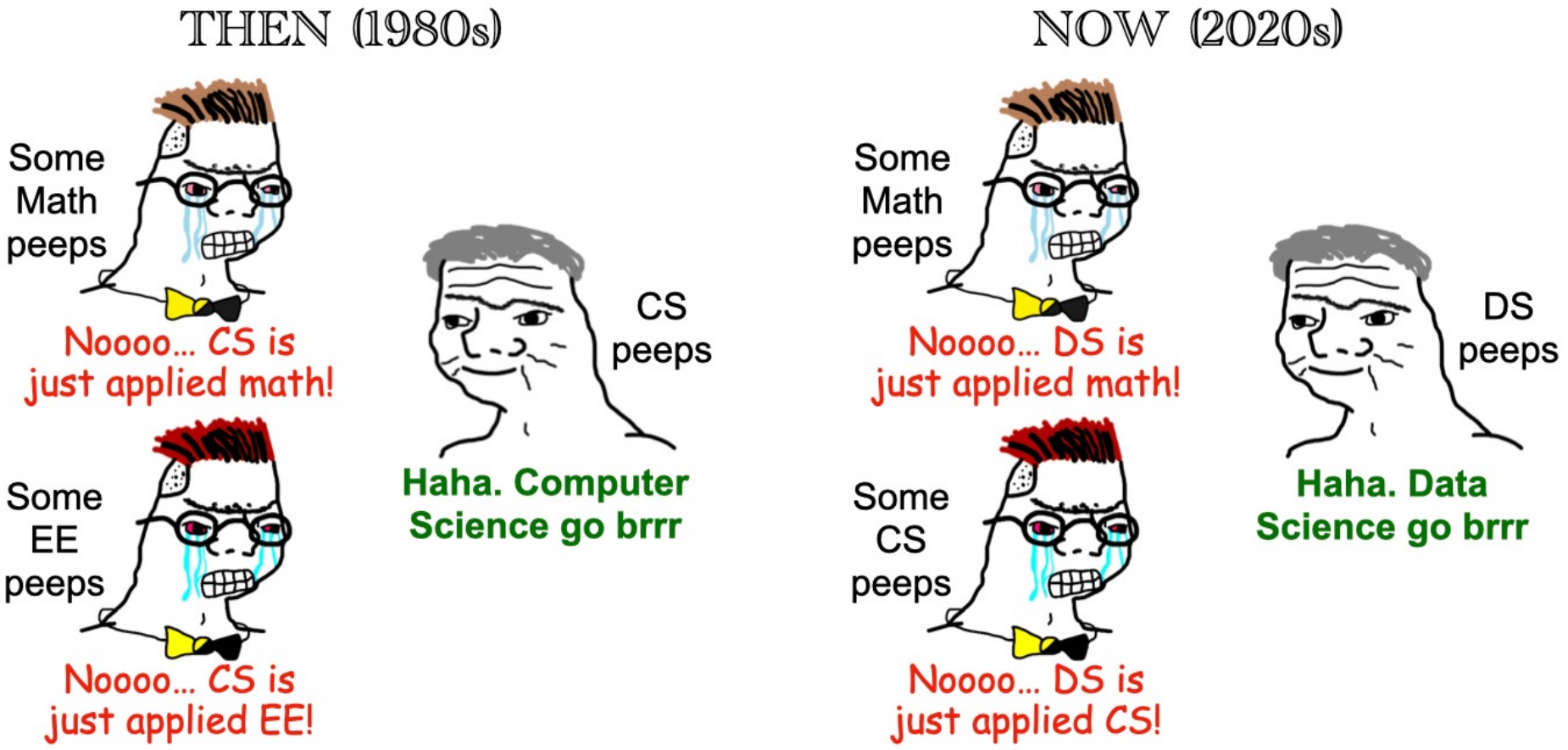
FROM THE OCTOBER 2012 ISSUE

Harvard Business Review

SUMMARY SAVE SHARE COMMENT 5 TEXT SIZE PRINT $8.95 BUY COPIES

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—

Meme from Previous DSC 102



15-213/15-513/14-513 Introduction to Computer Systems (ICS)

Fall 2023

- 15-213 Pittsburgh: Tue, Thu 12:30 PM–01:50 PM, GHC 4401, Brian Railing and Phillip Gibbons
- 14-513 Pittsburgh: Tue, Thu 12:30 PM–01:50 PM, CIC 1202, David Varodayan

12 units

The ICS course provides a programmer's view of how computer systems execute programs, store information, and communicate. It enables students to become more effective programmers, especially in dealing with issues of performance, portability and robustness. It also serves as a foundation for courses on compilers, networks, operating systems, and computer architecture, where a deeper understanding of systems-level issues is required. Topics covered include: machine-level code and its generation by optimizing compilers, performance evaluation and optimization, computer arithmetic, memory organization and management, networking technology and protocols, and supporting concurrent computation.

Course Syllabus

Prerequisites: 15-122

The ICS course provides **a programmer's view** of how computer systems execute programs, store information, and communicate.
It enables students to become **more effective programmers**, especially in dealing with issues of performance, portability and robustness.

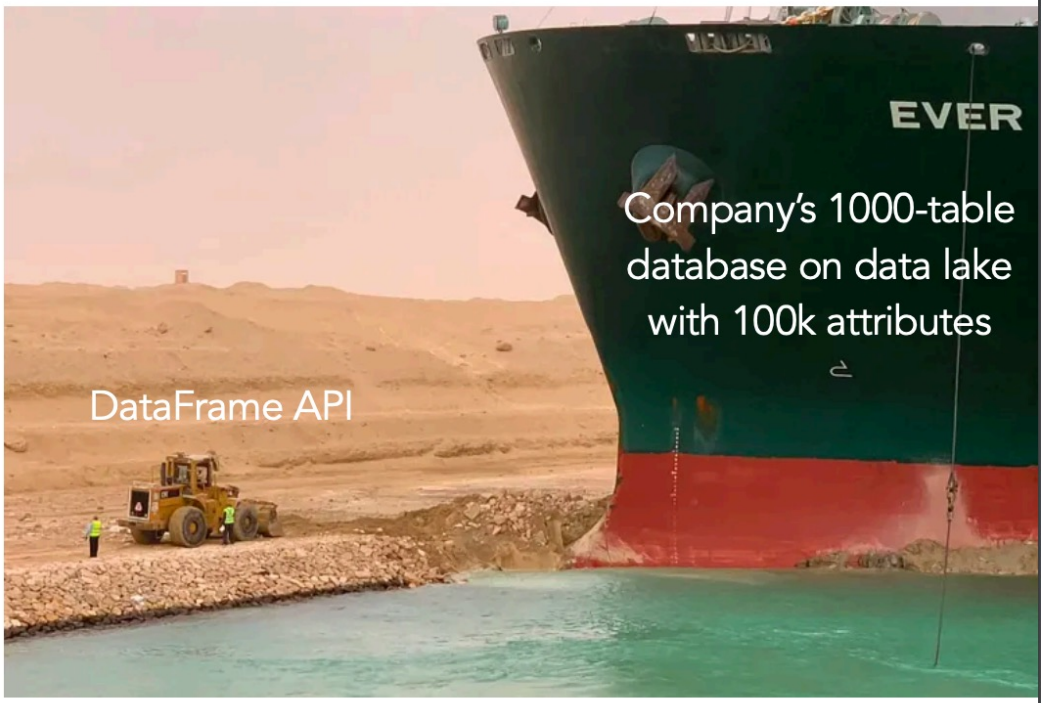
Vision

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.

DSC 204a

DSC 204a Scalable Data Systems

- Haojian Jin

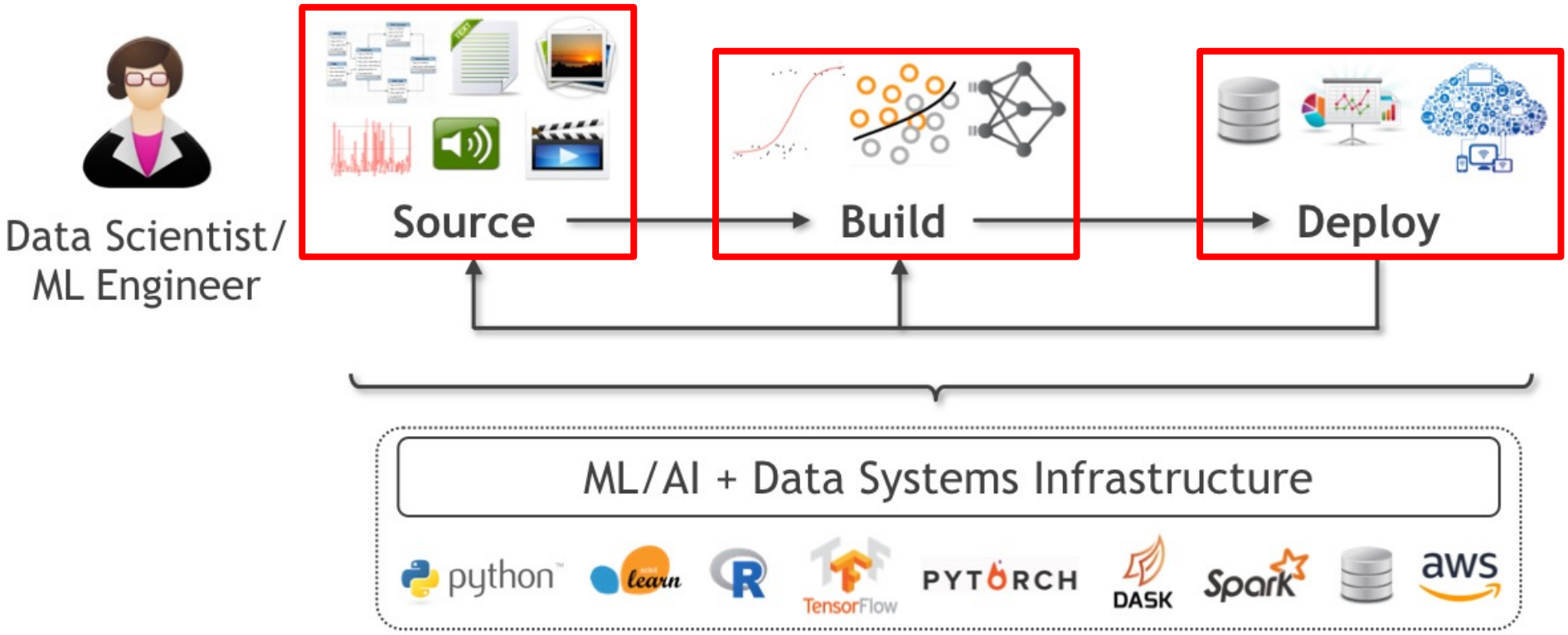


Meme idea credit: https://datasystemsfun.tumblr.com/

DSC 102 will get you thinking about the fundamentals of systems for scalable analytics

1. "**Systems**": What resources does a computer have? How to store and efficiently compute over large data? What is cloud?
2. "**Scalability**": How to scale and parallelize data-intensive computations?
3. **For "Analytics"**:
    1. **Source**: Data acquisition & preparation for ML
    2. **Build**: Model selection & deep learning systems
    3. **Deploying** ML models
4. Hands-on experience with scalable analytics tools

The Lifecycle of ML-based Analytics



Data acquisition
Data preparation

Feature Engineering
Training & Inference
Model Selection

Serving
Monitoring

ML Systems

***Q:** What is a Machine Learning (ML) System?*

- A data processing system (aka *data system*) for mathematically advanced data analysis operations (inferential or predictive):
  - Statistical analysis; ML, deep learning (DL); data mining (domain-specific applied ML + feature eng.)
  - *High-level APIs* to express ML computations over (large) datasets
  - *Execution engine* to run ML computations efficiently

Categorizing ML Systems

- **Orthogonal Dimensions of Categorization**:
  1. **Scalability:** In-memory libraries v. Scalable ML system (works on larger-than-memory datasets)
  2. **Target Workloads:** General ML library v. Decision tree-oriented v. Deep learning, etc.
  3. **Implementation Reuse:** Layered on top of scalable data system v. Custom from-scratch framework

Major Existing ML Systems

General ML libraries:

In-memory:

Disk-based files:

Layered on RDBMS/Spark:

Cloud-native:

Azure Machine Learning

Amazon SageMaker

"AutoML" platforms:

DataRobot

Decision tree-oriented:

Microsoft
LightGBM

Deep learning-oriented:

Data Systems Concerns in ML

Key concerns in ML:

Accuracy

Runtime efficiency (sometimes)

Additional key *practical* concerns in ML Systems:

Scalability (and efficiency at scale)

Usability

Manageability

Developability

Q: How do "ML Systems" relate to ML?

ML Systems : ML :: Computer Systems : TCS

*Long-standing concerns in the **DB systems** world!*

Q: [illegible] ML systems?

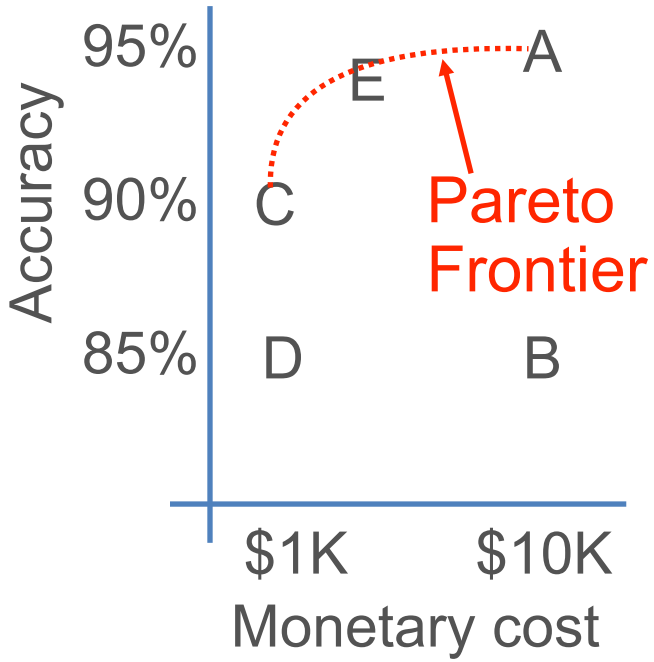
Conceptual System Stack Analogy

	Relational DB Systems	ML Systems
Theory	First-Order Logic Complexity Theory	Learning Theory Optimization Theory
Program Formalism	Relational Algebra	Tensor Algebra Gradient Descent
Program Specification	SQL	TensorFlow? Scikit-learn?
Program Modification	Query Optimization	???
Execution Primitives	Parallel Relational Operator Dataflows	Depends on ML Algorithm
Hardware	CPU, GPU, FPGA, NVM, RDMA, etc.	

Real-World ML: Pareto Surfaces

Q: *Suppose you are given ad click-through prediction models A, B, C, and D with accuracies of 95%, 85%, 90%, and 85%, respectively. Which one will you pick?*

Q: *What about now?*



- Real-world ML users must grapple with multi-dimensional *Pareto surfaces*: accuracy, monetary cost, training time, scalability, inference latency, tool availability, interpretability, fairness, etc.
- *Multi-objective optimization* criteria set by application needs / business policies.

Learning Outcomes of this course

- **Explain** the basic principles of the memory hierarchy, parallelism paradigms, scalable data systems, and cloud computing.
- **Identify** the abstract data access patterns of, and opportunities for parallelism and efficiency gains in, data processing and ML algorithms at scale.
- **Outline** how to use cluster and cloud services, dataflow ("Big Data") programming with MapReduce and Spark, and ML tools at scale.
- **Apply** the above programming skills to create 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 databases, relational model, or SQL
  - Take DSC 100 instead (pre-requisite)
- NOT a course on internal details of RDBMSs
  - Take CSE 132C instead
- NOT a training module for how to use Spark
- NOT a course on ML or data mining *algorithmics*; instead, we focus on ML *systems*

Now for the course logistics …

Prerequisites

- **DSC 100** (or equivalent) is necessary
- Transitively **DSC 80**; a mainstream ML algorithmics course is necessary
- Proficiency in Python programming
- 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.
  - **Plan your credit as well!**
- **Midterm Exam: 15%**
  - **TBD**; in-class only (50min)
- **Cumulative Final Exam: 35%**
  - 3hrs long but 4hrs limit
- **10 (of 12) Peer Instruction Activities: 10%**
- **Extra Credit Evaluation Activities:** 2% (likely)
- LMK ahead of time if you need makeup exam slot

https://haojian.github.io/DSC102SP24/

Grading Scheme

Hybrid of relative and absolute; grade is better of the two

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

Example: Score 82 but 33%le; Rel.: B-; Abs.: B+; so, B+

Programming Assignments

- **PA0: Setting up AWS and Dask**
- **PA1: Data Exploration with Dask**
- **PA2: Feature Eng. and Model Selection with Spark**
- **Expectations on the PAs:**
  - Teams of 1-3; 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!

https://haojian.github.io/DSC102SP24/

Course Administrivia

- **Lectures**: **MWF 3pm-3:50pm PT** at **Mandeville Center - B-202**
  - Attendance optional but encouraged; podcast available
  - No need for clickers.
- **Discussions:**
  - Only for talks on PAs by TAs, for pre-exam review by me
- **Instructor**: Haojian Jin; haojian@ucsd.edu
  - OHs: **Wednesday 4-5 pm PT** at **HDSI 341**
- **Slack** for all communications
- **Canvas** for PA submission, Peer Evaluation Activities, Final Exam

https://haojian.github.io/DSC102SP24/

Office hours

- Haojian Jin's OHs: Wednesday 4:00 PM - 5:00 PM
- Course content.
- Qiyu Li's OHs: TBD
- Ariane Yu's OHs: TBD
- Assignments, HDSI 3rd floor. Near conference rooms.
- Post questions to the ta-public channel.
- Avoid asking repetitive questions.

General Dos and Do NOTs

Do:

- Follow all announcements on Piazza
- Try to join the lectures/discussions live
- Raise your hand before speaking
- View/review podcast videos asynchronously by yourself
- To contact me/TAs, use private Slack; if you really need to email, use “DSC 102:” as subject prefix

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!

Reasonable person.

(1) Everyone will be reasonable.
(2) Everyone expects everyone else to be reasonable.
(3) No one is special.
(4) Do not be offended if someone suggests you are not being reasonable.