W4111
Introduction to Databases
Spring 2022

Computer Science Department

Columbia University

Welcome!

Eugene Wu

B.S.	U.C. Berkeley
Ph.D.	MIT
PostDoc	U.C. Berkeley
Professor	Columbia since Fall 2015

Database systems, vis, data analysis, cleaning, ML systems, crowdsourcing.

www.eugenewu.net
ewu@cs.columbia.edu
421 Mudd

Office hours
Weds 2-3PM virtual.
By appointment by email

Your TAs

Zachary Huang. PhD. Head TA

When I was undergraduate, database was the only CS course I didn't get A.

Jake Fisher. UG. Head TA

There are two other Jake Fishers at Columbia!
(Make sure not to email the other ones accidentally!)

Sughosh V Kaushik. MS. TA

I read a lot of manga and I skateboard :)
PS: one piece is the best manga till date and don't try skateboarding unless you don't want to graduate in one piece.

Your TAs

Ashwathy Menon. MS. TA

The first time that I moved away from home is for MS CS at Columbia. I absolutely love exploring the city and even re-visit the places I have already been to. The one hobby I would never give up is reading.

Rachel Halpern. UG. TA

I lived in Singapore for 6 years!

Twisha Jain. MS. TA

If I could star in any movie series, it would definitely be Harry Potter!

Your TAs

Office hours will be updated on course website later today.

Zoom links for office hours will be shared on the discussion board

Agenda

Big Picture Overview

Data Science and Data Systems

Course Info

Entity-Relationship Modeling

Data

Data

is for serious business

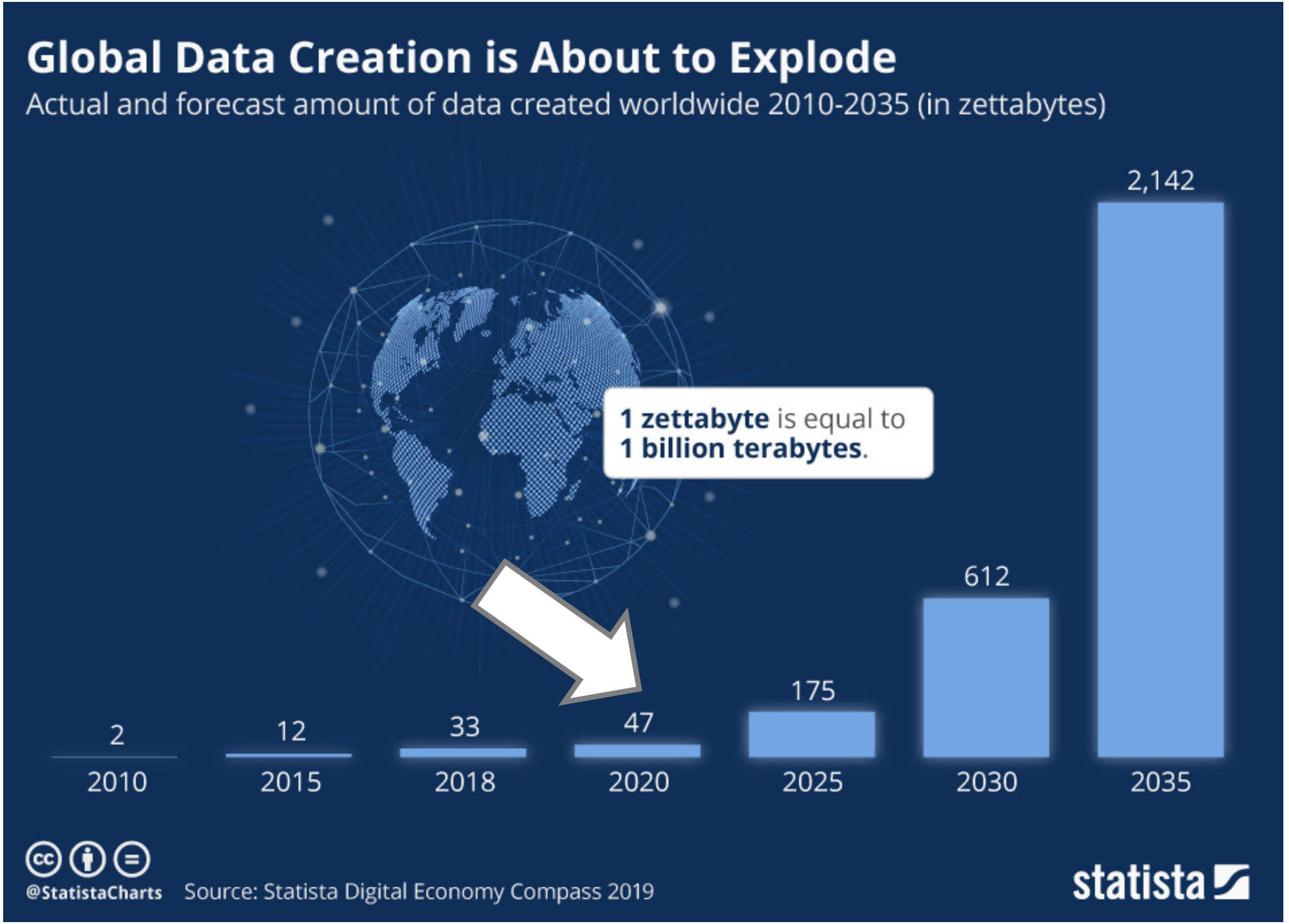
Data

is at the center of most things.

Data

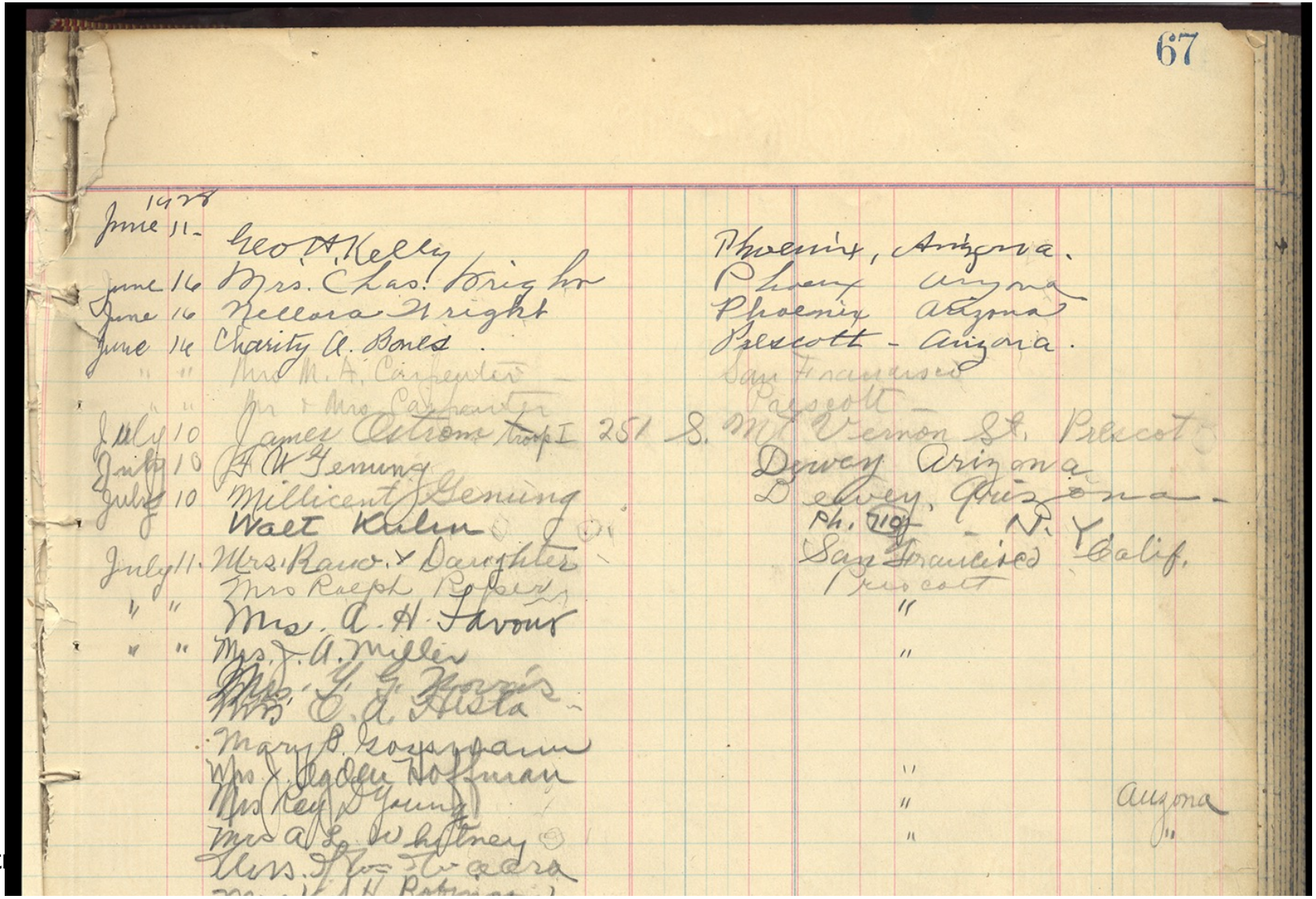
is at the center of *everything*


Data


Global Data Creation is About to Explode
Actual and forecast amount of data created worldwide 2010-2035 (in zettabytes)
1 zettabyte is equal to 1 billion terabytes.
2
12
33
47
175
612
2,142
2010
2015
2018
2020
2025
2030
2035
@StatistaCharts
Source: Statista Digital Economy Compass 2019
statista

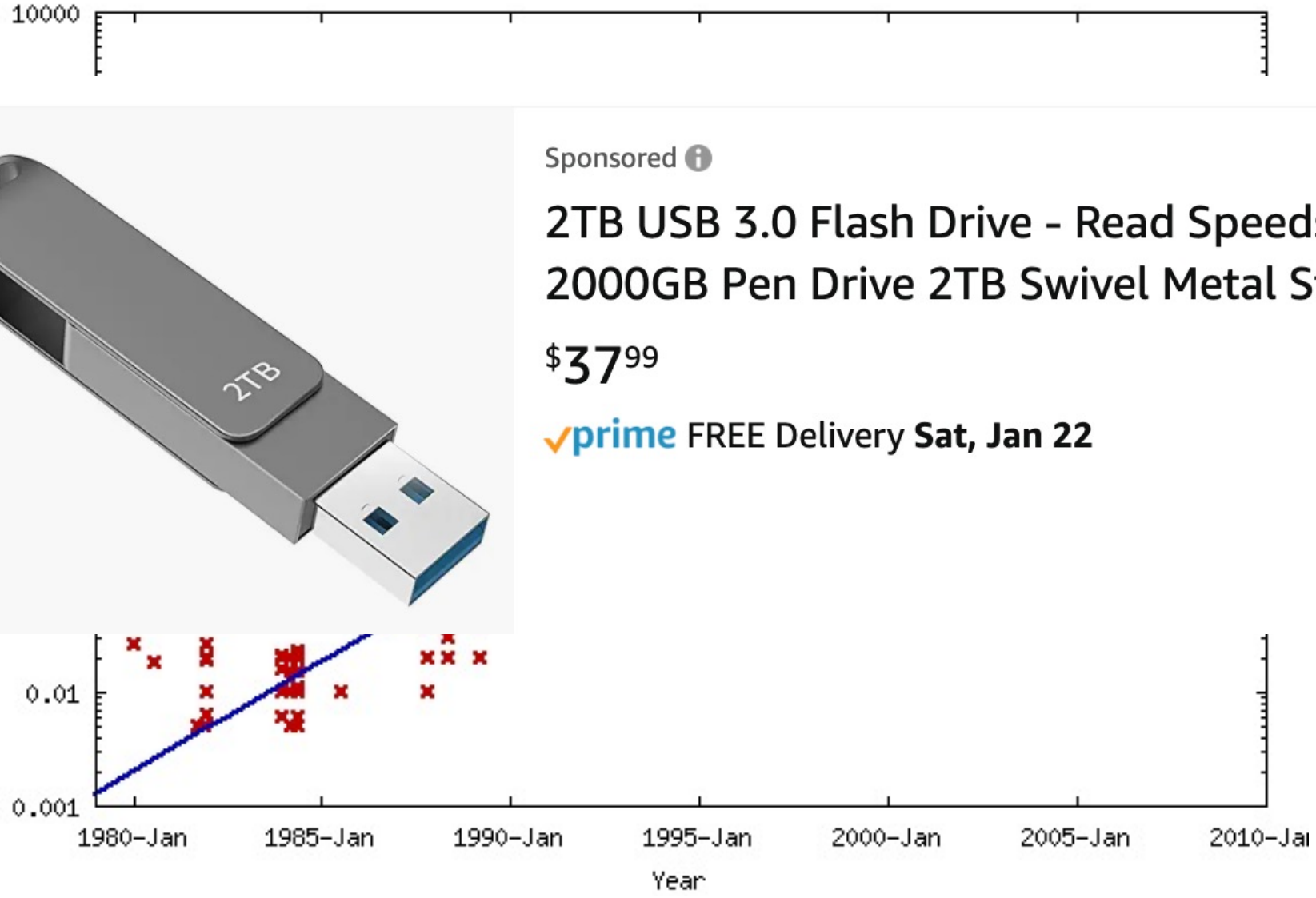
How did we get here?

Data was *Manual*



Data was *Expensive*

Data is *Cheap*



Data is *Automated*

Physical devices

Data is *Automated*

Physical devices

Data is *Automated*

Physical devices

Software logs

Data is *Ubiquitous*

Physical devices

Software logs

Phones

Data is *Ubiquitous*

Physical devices

Software logs

Phones

GPS/Cars

sleep number

- Information You provide to Us on a financing application
- Information You provide to Us to allow Your participation in a partner's loyalty program

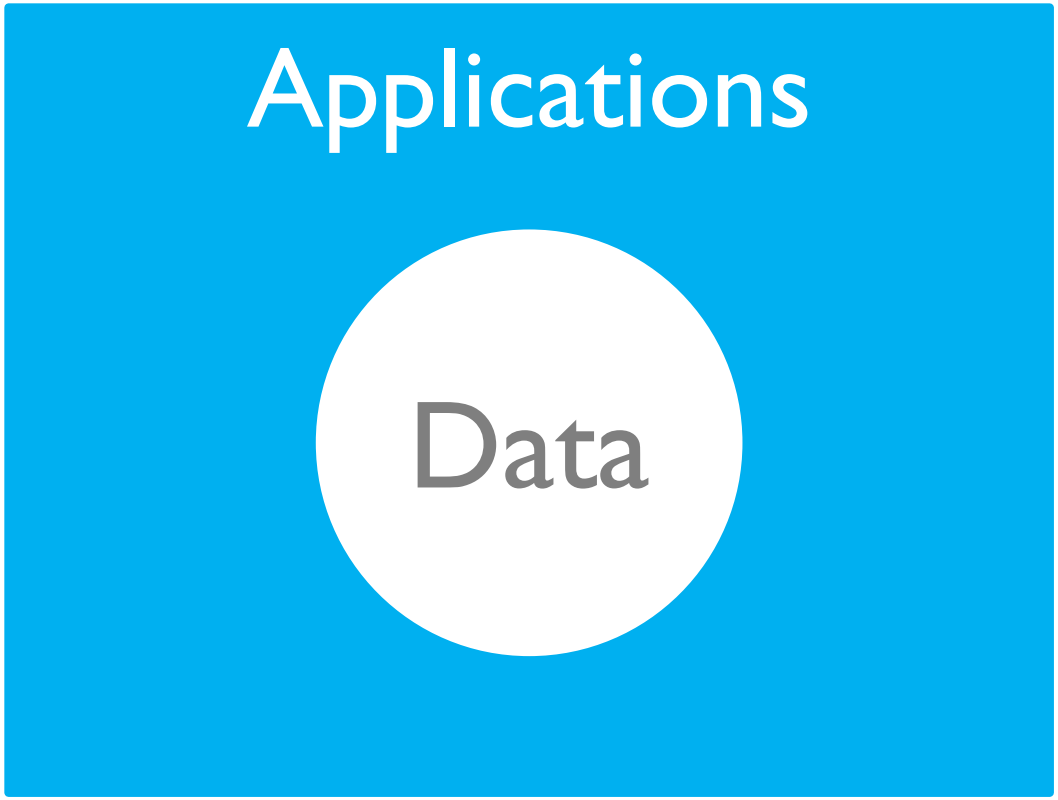
Once You create a User Profile, We also may collect Personal Information, which may include, among other types of information:

- Revised or updated User Profile information
- Biometric and sleep-related data about how You, a Child, and any person that uses the Bed slept, such as that person's movement, positions, respiration, and heart rate while sleeping
- Audio in Your room to detect snoring and similar sleep conditions
- Other information You choose to provide to Us by opting in to additional functionality of Our

Connect to your smart phone
to make & receive calls


Data

What Applications?



What are we doing with data?

Health

What are we doing with data?

Health

What are we doing with data?

Health

Investigative Journalism

MACHINE BIAS

Besieged Facebook Says New Ad Limits Aren't Response to Lawsuits

The social network is removing 5,000 options that regulators say enable advertisers to discriminate.

by Ariana Tobin and Jeremy B. Merrill, Aug. 23, 12:48 p.m. EDT

RETAIL OCTOBER 10, 2018 / 7:04 PM / UPDATED 3 YEARS AGO

REUTERS

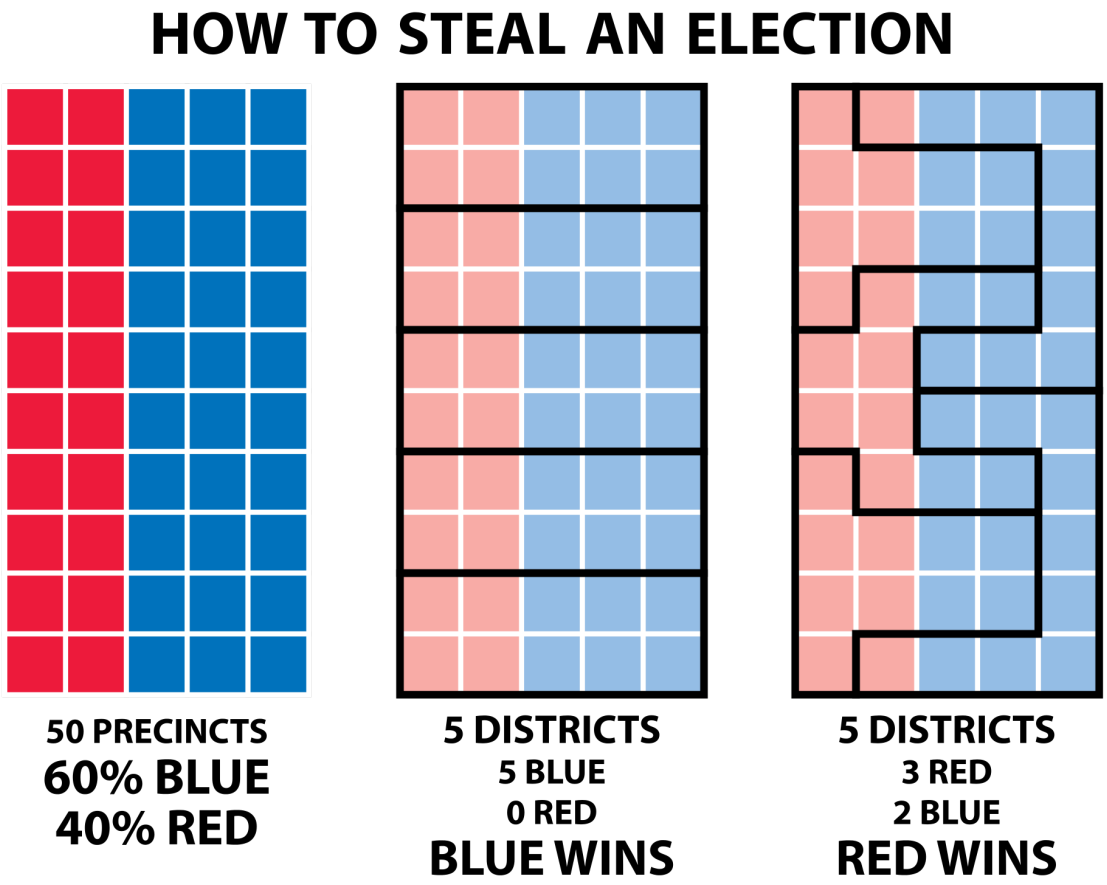
Amazon scraps secret AI recruiting tool that showed bias against women

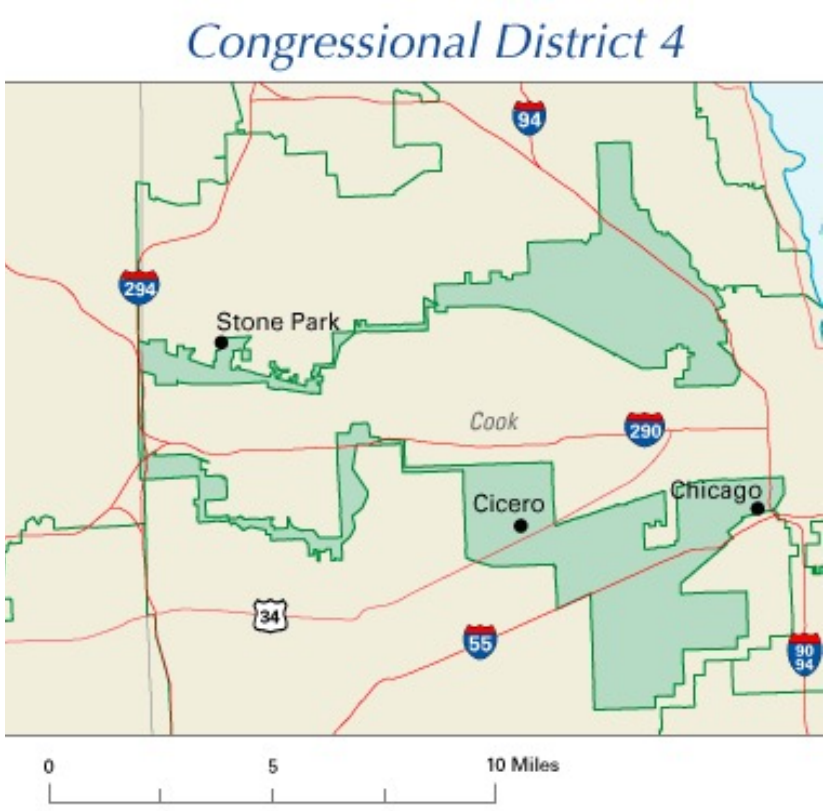
By Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

What are we doing with data?





What are we doing with data?

Health

Investigative Journalism

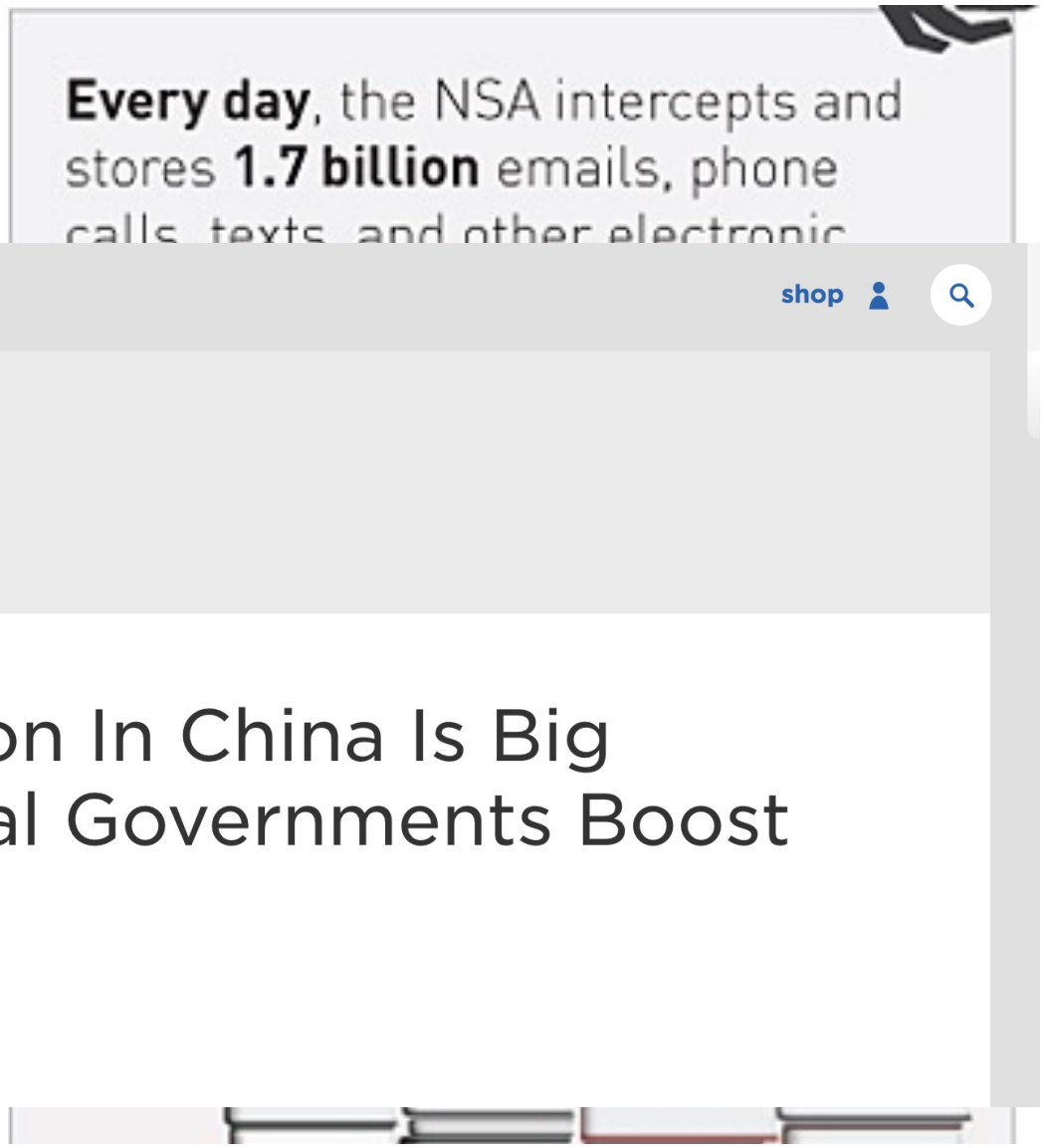
Recommendations

Politics

What are we doing with data?

Health

Investigative Journalism



What are we doing with data?

Health

Investigative Journalism

Recommendations

Politics

Surveillance

Identity

HARVARD | BUSINESS | SCHOOL

30 APR 2012 RESEARCH & IDEAS

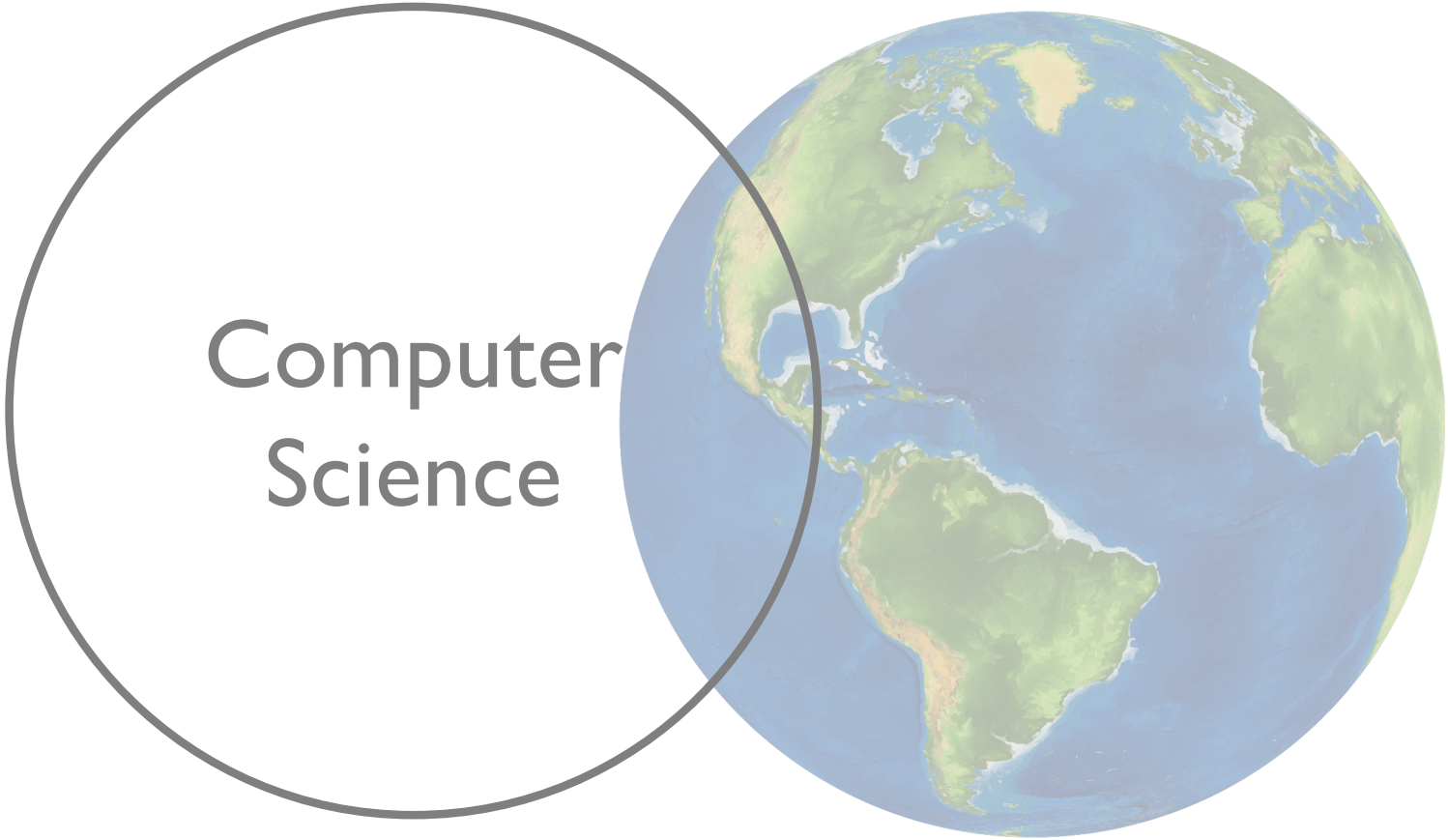
India's Ambitious National Identification Program

Comments 30 Email Print Download Share Recommend Share 92

The Unique Identification Authority of India has been charged with implementing a nationwide program to register and assign a unique 12-digit ID to every Indian resident—some 1.2 billion people—by 2020. In a new case, Professor Tarun Khanna and HBS India Research Center Executive Director Anjali Raina discuss the complexities of this massive data management project.

"YOU ARE BASICALLY DENIED ALMOST EVERYTHING IF YOU CAN'T PROVE WHO YOU ARE."

Computer Science


Computer
Science


Computer
Science


Data
Science

EU Parliament
Draft Report on AI in the Digital Age

Argues that artificial intelligence (AI) is the key emerging technology within the fourth industrial revolution; notes that AI is the control centre of the new data layer that surrounds us and which can be thought of as the fifth element after air, earth, water and fire; states that by 2030, AI is expected to contribute more than EUR 11 billion to the global economy, an amount that almost matches China's GDP in 2020;

Conventional View of AI/Data Science

Lone data scientist uses a static, clean table,
applies statistics or fits an ML model
to increase a well-defined score

See popular ML articles, Kaggle competitions, etc

Conventional View of AI/Data Science

Lone data scientist uses a static, clean table,
applies statistics or fits an ML model
to increase a well-defined score

See popular ML articles, Kaggle competitions, etc

Conventional View of AI/Data Science

Team

Dynamic Messy Non-tabular

Lone data scientist uses a static, clean table,

applies statistics or fits an ML model

Unavailable

to increase a well-defined score

unclear, ill-defined

Huge amount of "unseen labor" (data engineering)

in order to support real-world data science & ML

In Reality…

An on-call engineer's biggest nightmare

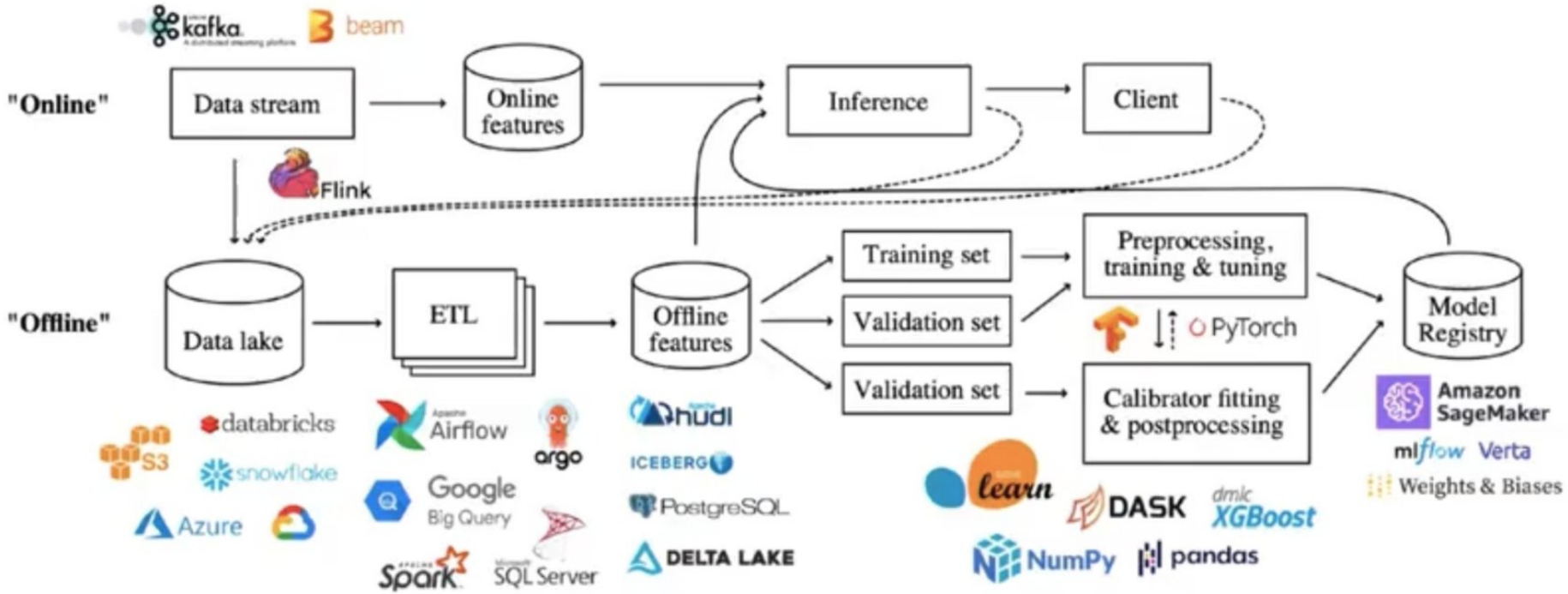


Figure 1: High-level architecture of a generic end-to-end machine learning pipeline. Logos represent a sample of tools used to construct components of the pipeline, illustrating heterogeneity in the tool stack. *Shankar et al. 2021*

https://www.facebook.com/Engineering/videos/1578607659138164/

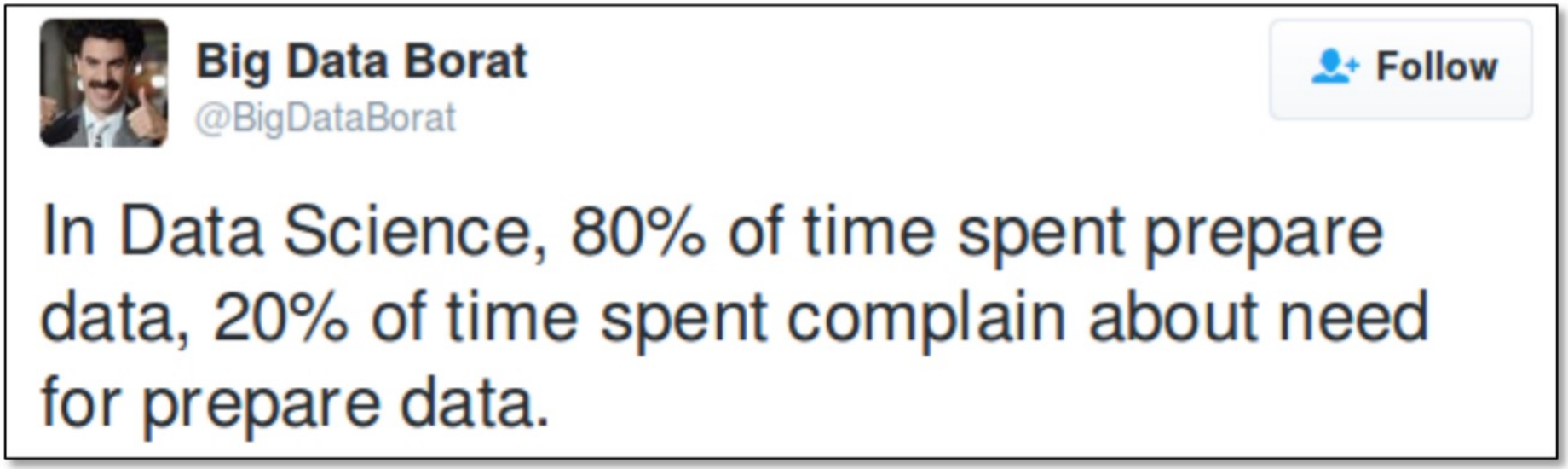
In Reality…

Data engineering dominates data science projects

Data engineer work >> data scientist work

Data engineering key to ML/AI/data science

Data Eng Dominates Data Science Projects



Most time spent on data engineering

Often not viewed as sexy, but critical

Data Eng Dominates Data Science Projects

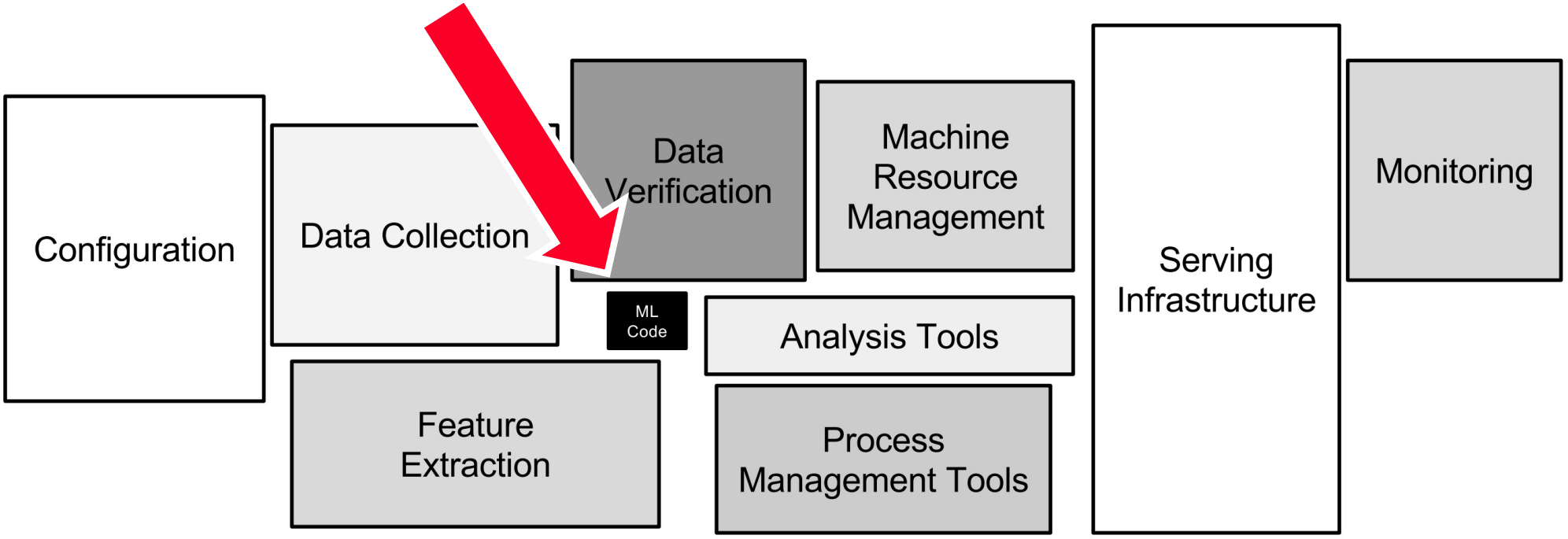
Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.

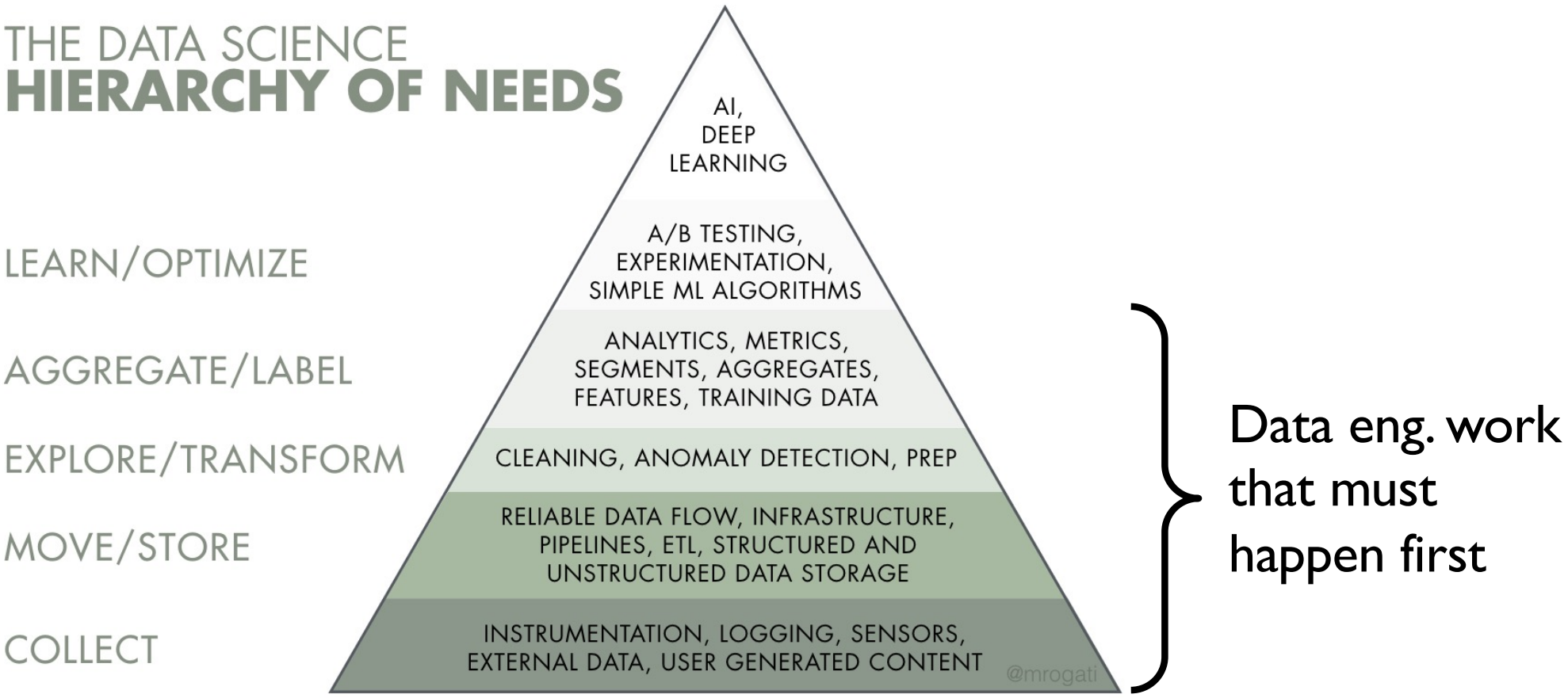
Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, Dan Dennison
{ebner,vchaudhary,mwyoung,jfcrespo,dennison}@google.com
Google, Inc.

Data Eng Dominates Data Science Projects

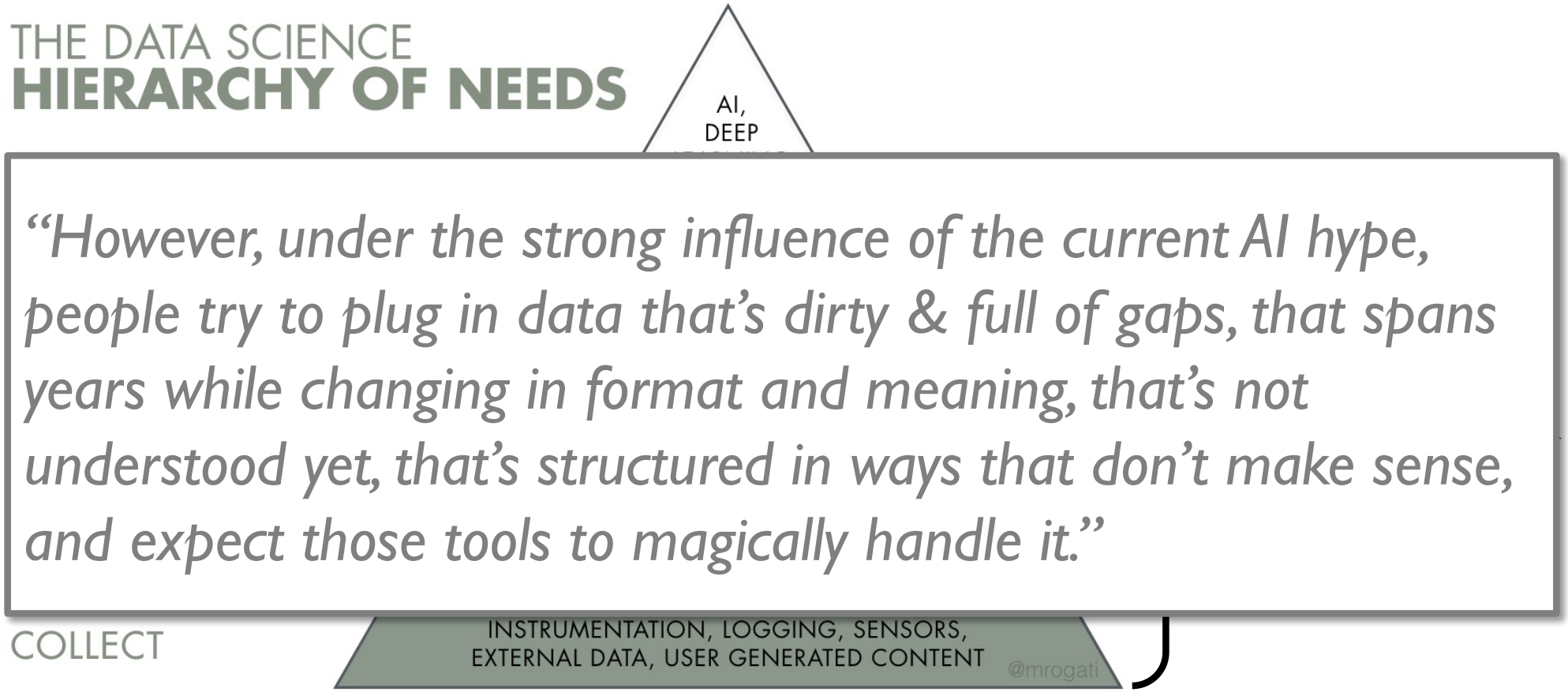
Hidden Technical Debt in Machine Learning Systems



Data Eng Dominates Data Science Projects



https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007



https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007

"*Everyone wants to do the model work, not the data work*": Data Cascades in High-Stakes AI

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, Lora Aroyo
[nithyasamba,kapania,hhighfill,dakrong,pkp,loraa]@google.com
Google Research
Mountain View, CA

Incentives and currency in AI An overall lack of recognition for the invisible, arduous, and taken-for-granted data work in AI led to poor data practices, resulting in the data cascades below. Care of, and improvements to data are not easily 'tracked' or rewarded, as opposed to models. Models were reported to be the means for prestige and upward mobility in the field [112] with ML publications that generated citations, making practitioners competitive for AI/ML jobs and residencies. "*Everyone wants to do the model work, not the data work*" (P4, healthcare, India). Many practitioners described data work as time-consuming, invisible to track, and of-

Data Engineering as a Job Category

"..70% more open roles at companies in data engineering as compared to data science…."

Data Engineering as a Job Category

Job Opening Estimates from Zippia

Data scientist:	79K	16% growth rate
Data engineer:	170K	21% growth rate

Overview of Data Engineering Concerns

ETL and Data Warehouses

Data lakes

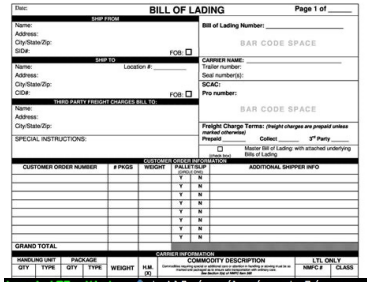
Data Quality

Metadata

Preparation

Python Scripts,
Spark/Databricks, etc

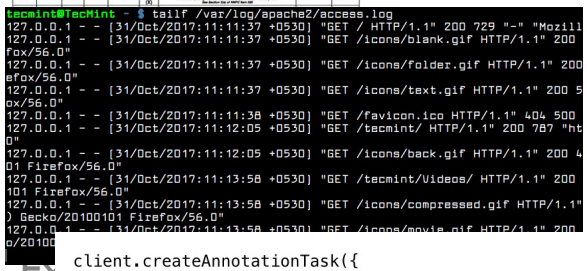
PDF



Data Preparation

Transform/clean data into useful form

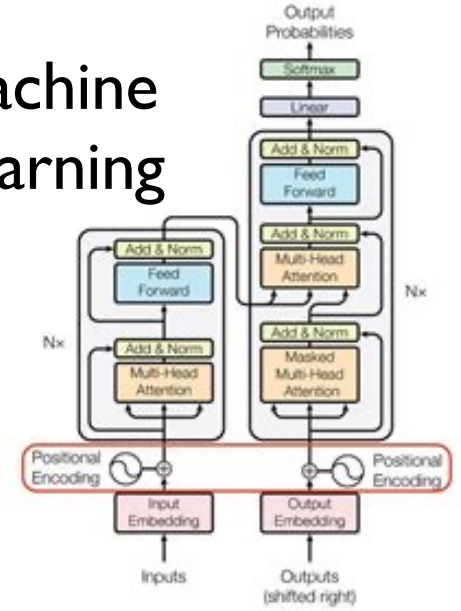
Logs



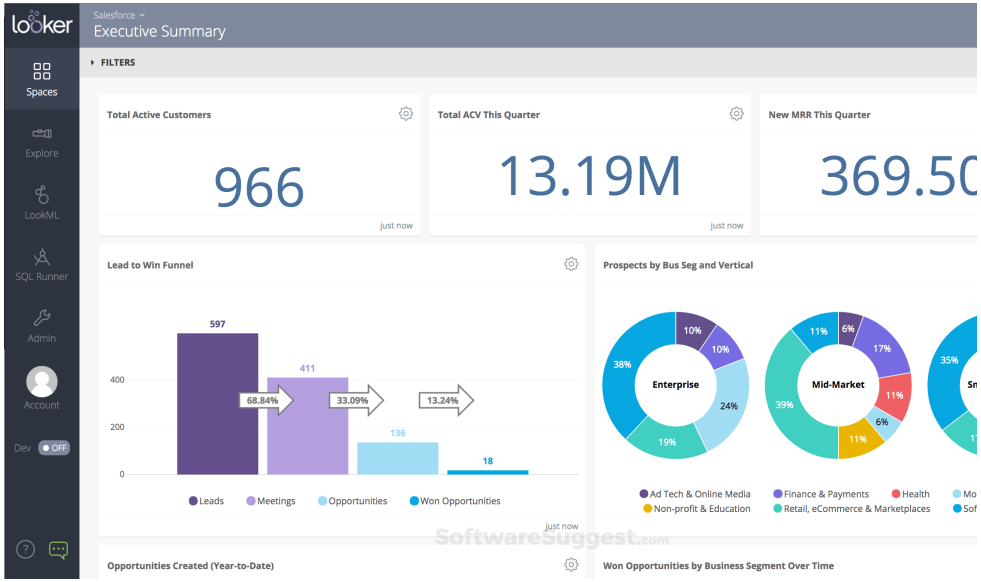
JSON

```
client.createAnnotationTask({
  callback_url: 'http://www.example.com/callback',
  instruction: 'Draw a box around each rooftop and pool.',
  attachment: 'http://i.imgur.com/XOJbalC.jpg',
  objects_to_annotate: ['pool', 'rooftop'],
  with_labels: true,
  min_width: 30,
  min_height: 30
}, (err, task) => {
    // do something with task
});
```

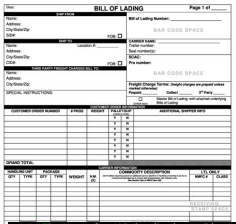
Machine Learning

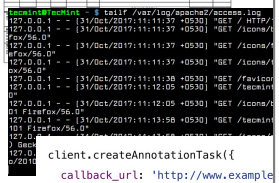


Visualization

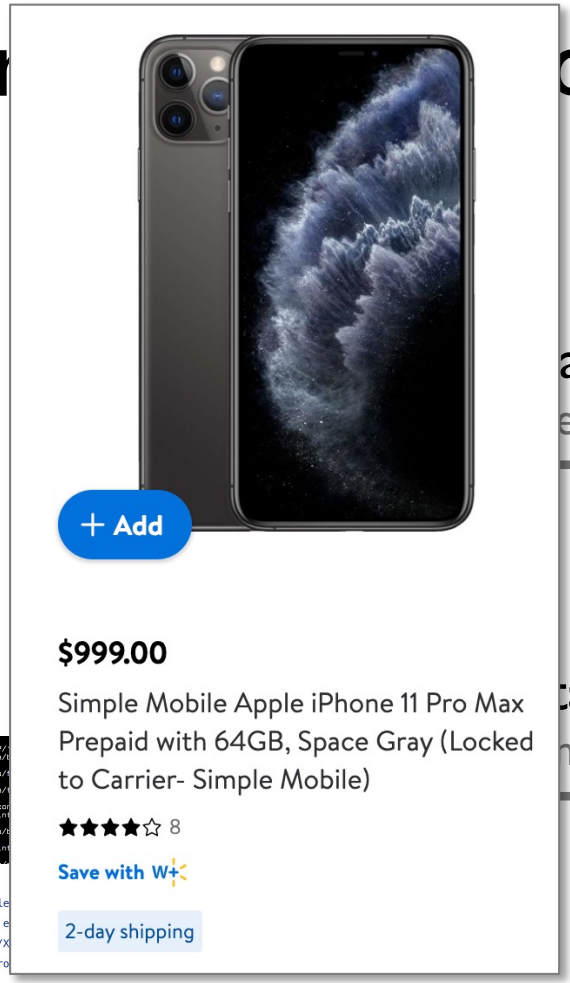


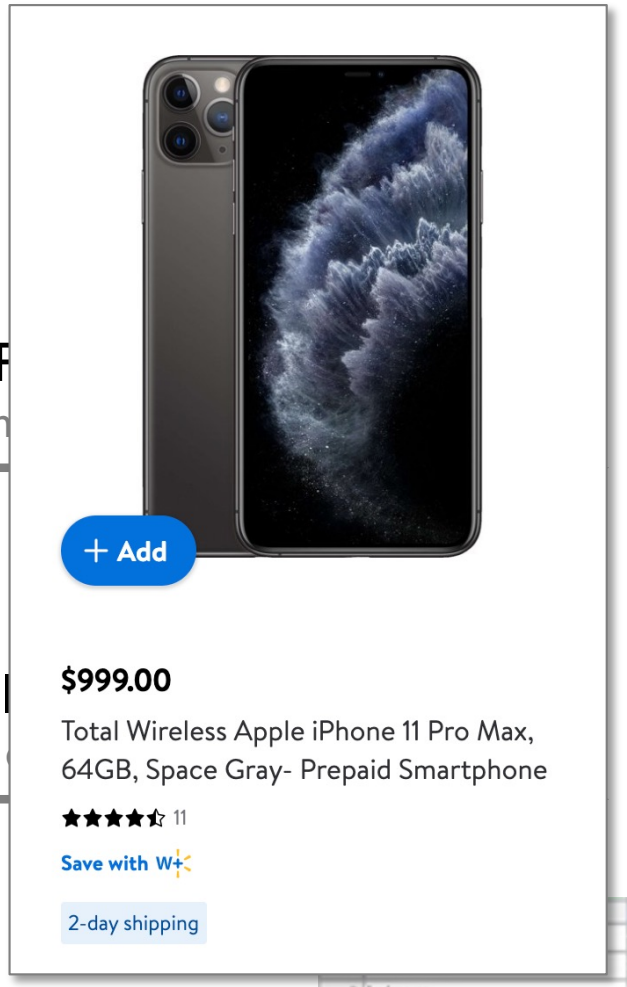
Sources





```
client.createAnnotationTask({
  callback_url: 'http://www.example
  instruction: 'Draw a box around e
  attachment: 'http://i.imgur.com/X
  objects_to_annotate: ['pool', 'ro
  with_labels: true,
  min_width: 30,
  min_height: 30
}, (err, task) => {
  // do something with task
}
```





Use Cases

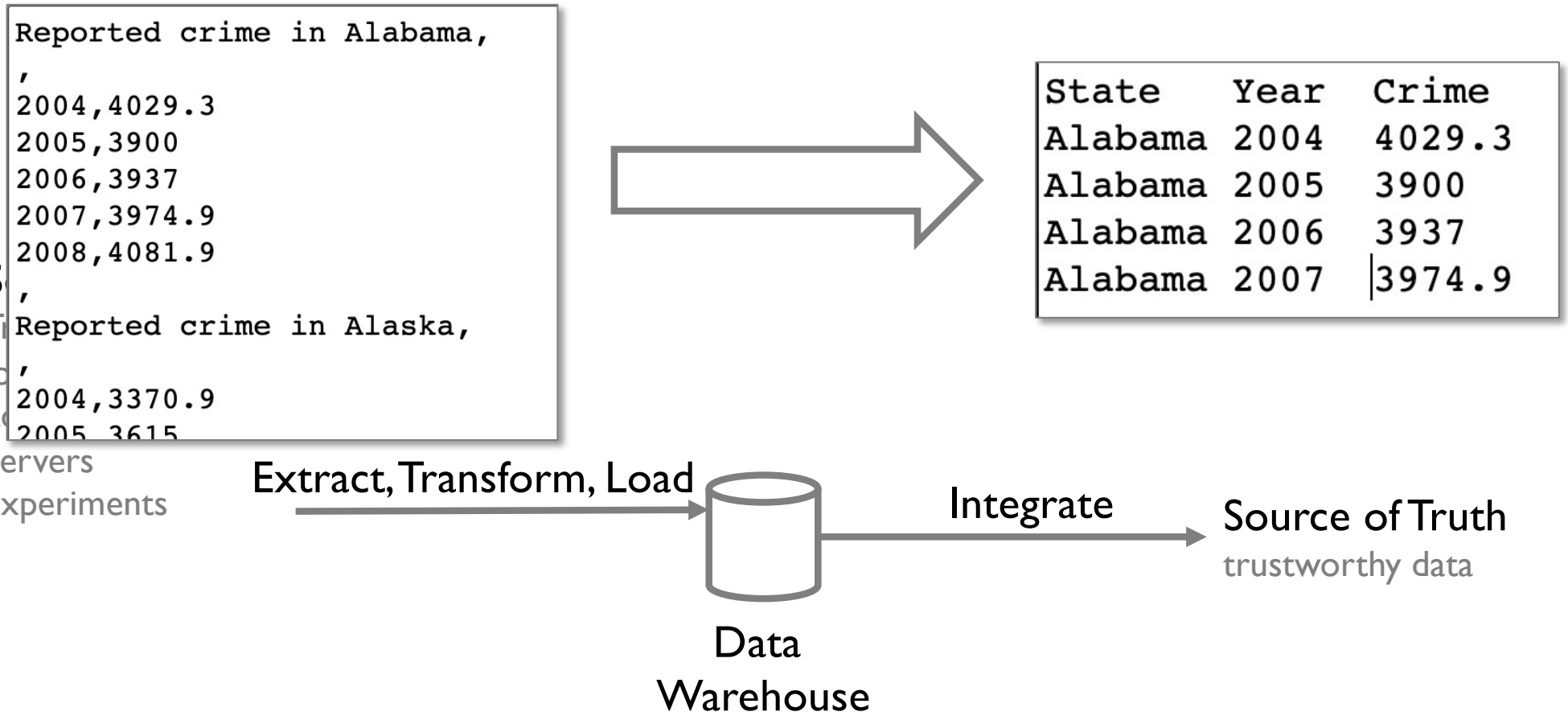
AI, ML, Data science, Apps, Webservices

Source of Truth

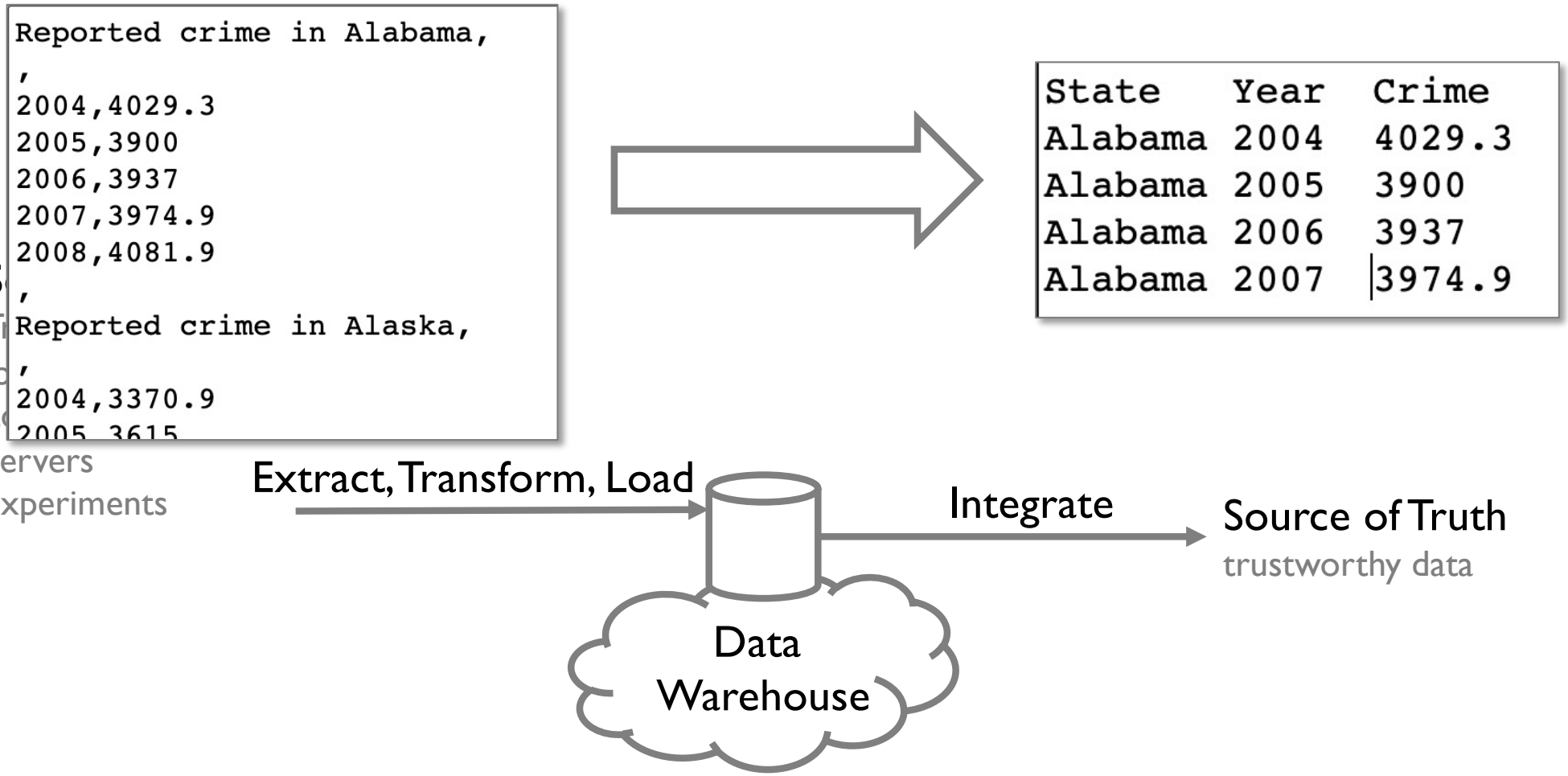
trustworthy data

		2004	2005
		4029.3	3900
		3370.9	3615
2	Arizona	5073.3	4827
3	Arkansas	4033.1	4068
4	California	3423.9	3321
5	Colorado	3918.5	4041
6	Connecticut	2684.9	2579
7	Delaware	3283.6	3110
8	District of Columbia	4852.8	4490
9	Florida	4182.5	4013

ETL: Extract Transform Load (1980s)



ELT: Extract Load Transform (2010s)



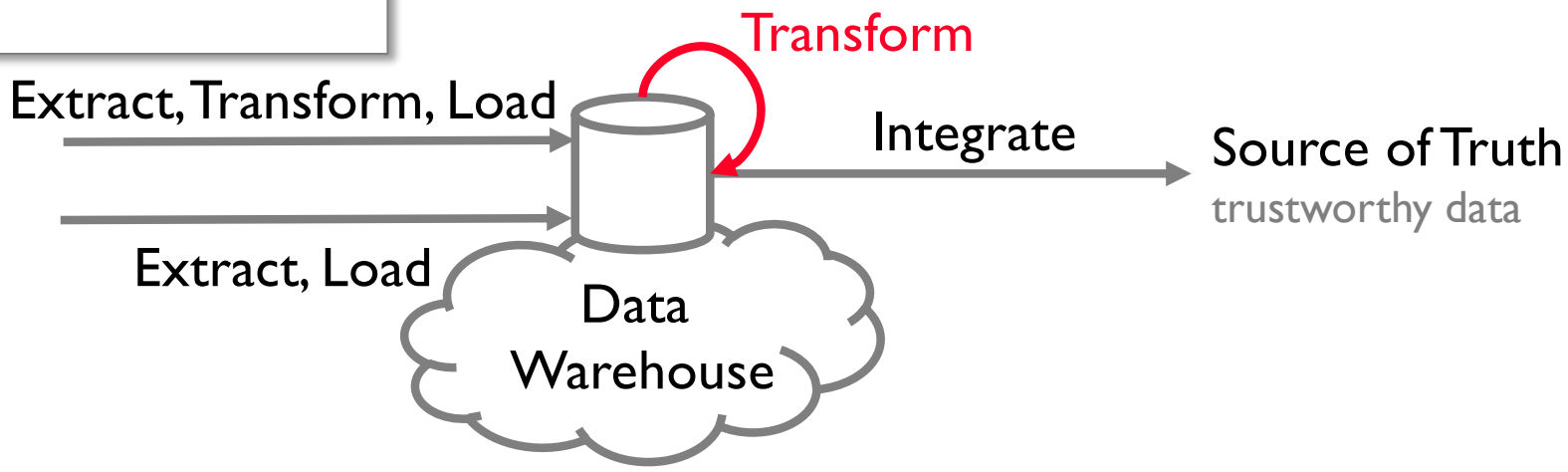
ELT: Extract Load Transform (2010s)

```
Reported crime in Alabama,
,
2004,4029.3
2005,3900
2006,3937
2007,3974.9
2008,4081.9
,
Reported crime in Alaska,
,
2004,3370.9
2005,3615
```

SQL Transform

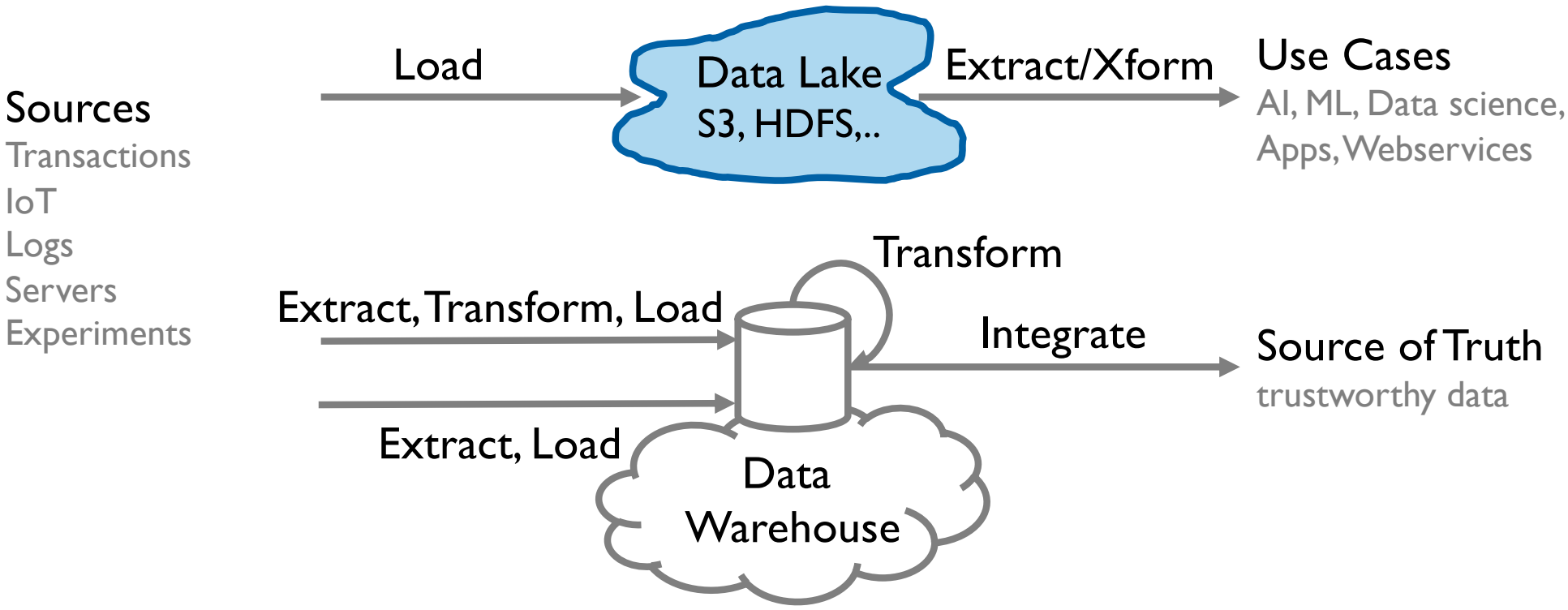
State	Year	Crime
Alabama	2004	4029.3
Alabama	2005	3900
Alabama	2006	3937
Alabama	2007	3974.9

S
T
Io
L
Servers
Experiments

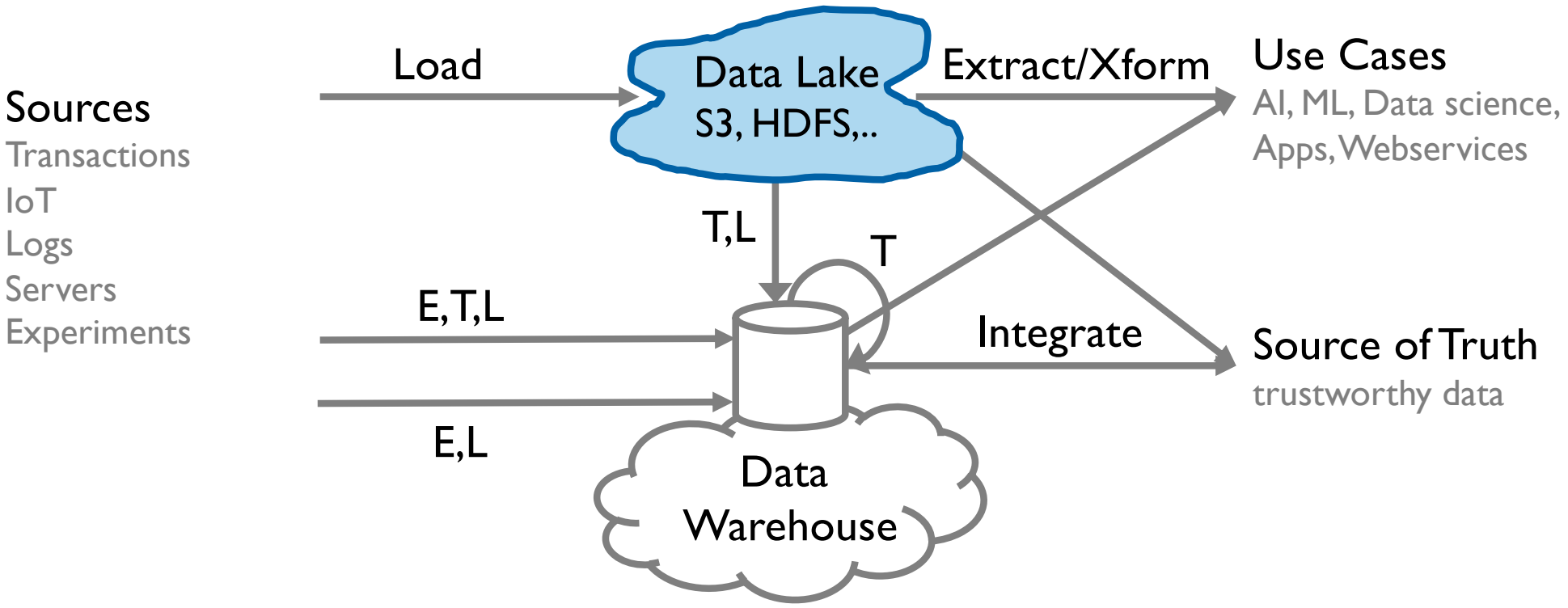


Data Lakes (2000s)

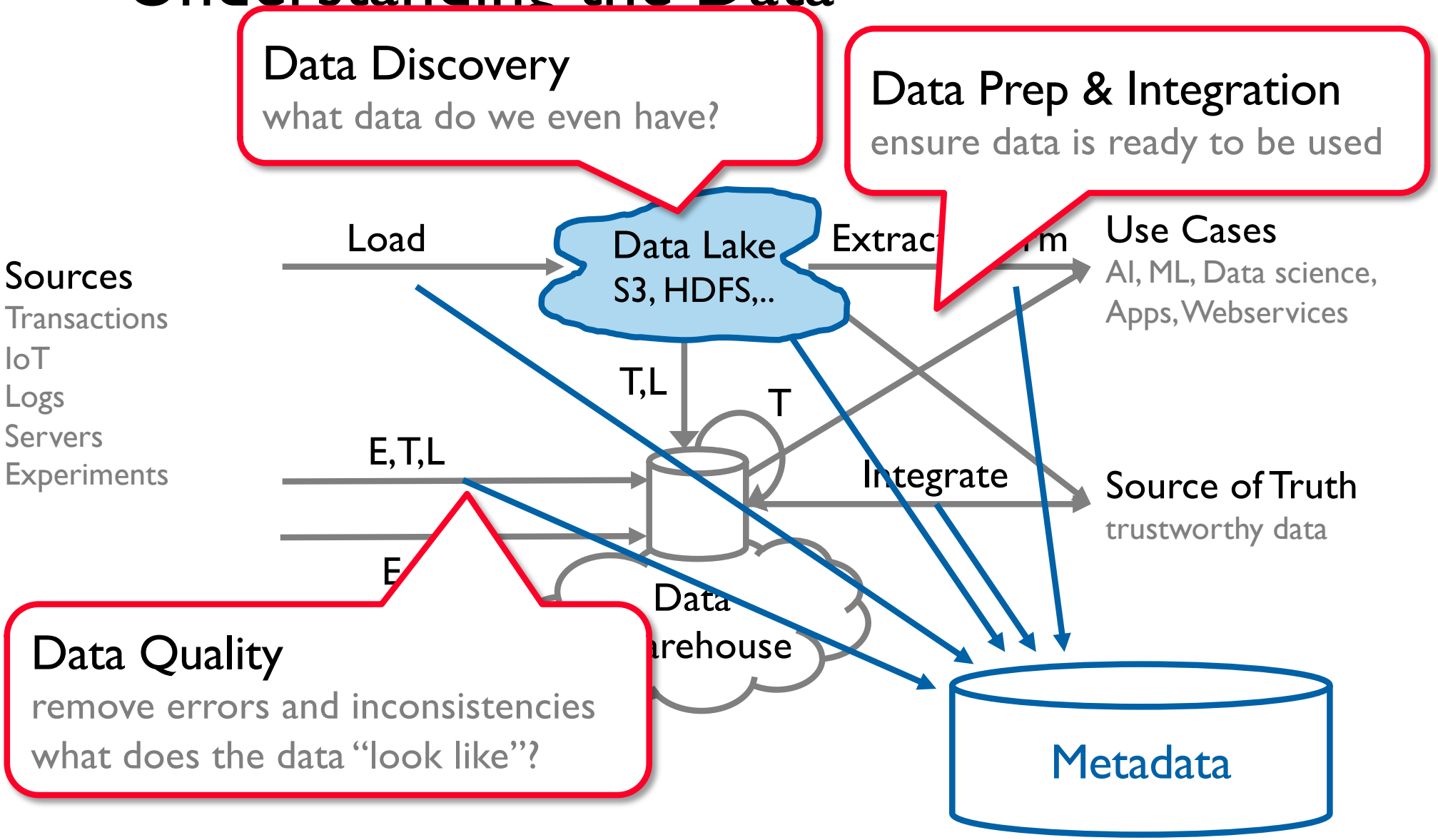
Store as files, transform when needed

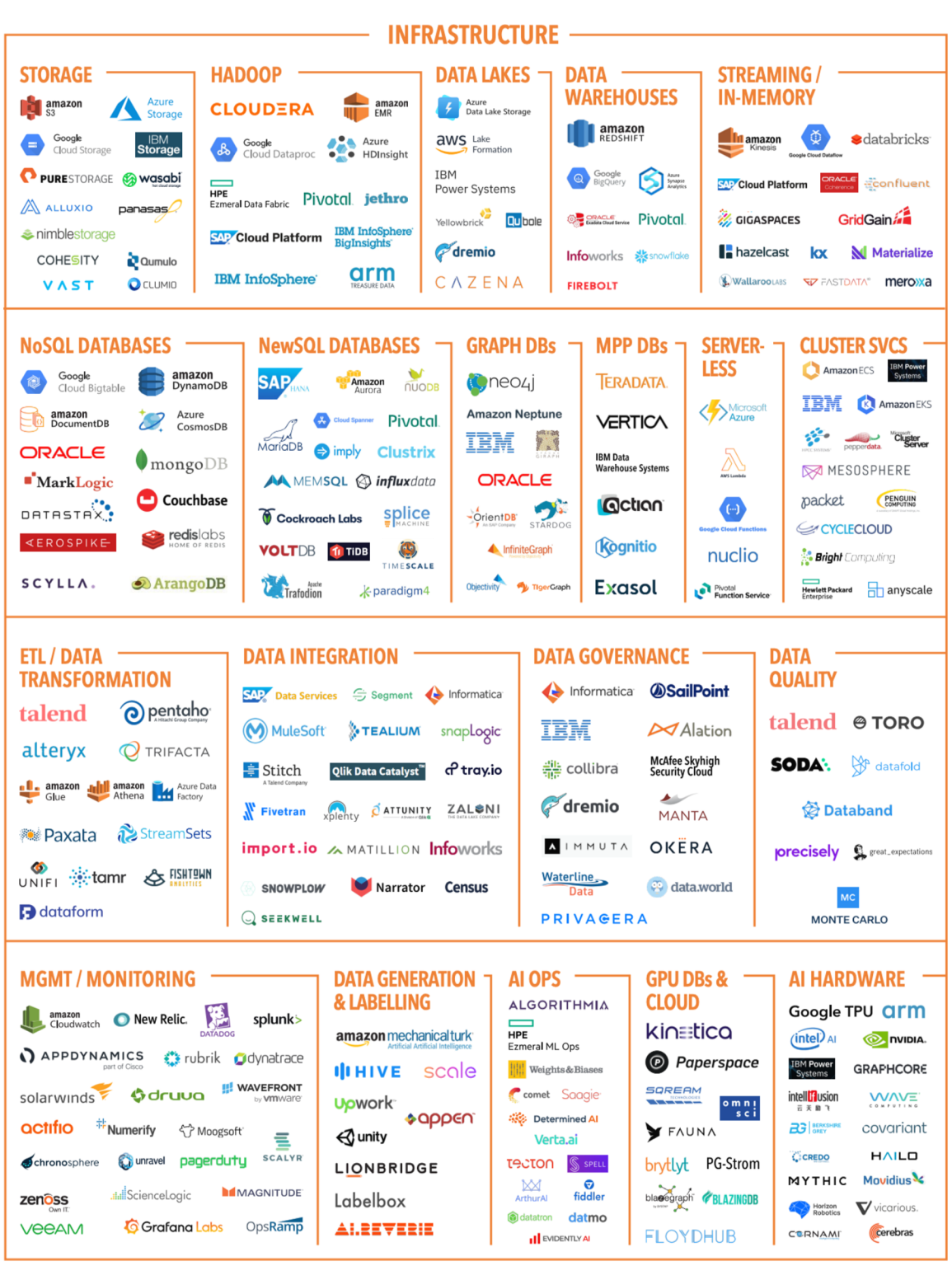


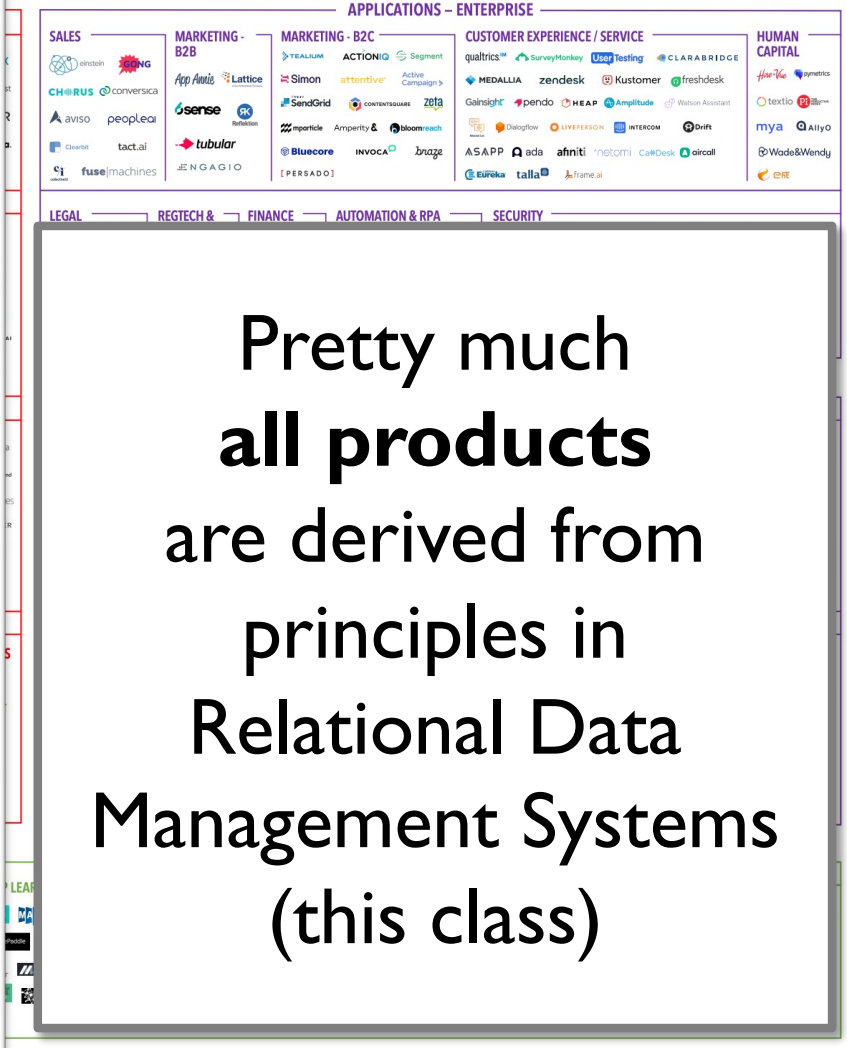
Everything All At Once (2010s)



Understanding the Data







Pretty much
all products
are derived from
principles in
Relational Data
Management Systems
(this class)

tturck.com/data2020

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

https://mattturck.com/data2020/

4111: Intro to Relational Data Management Systems

What's a database?

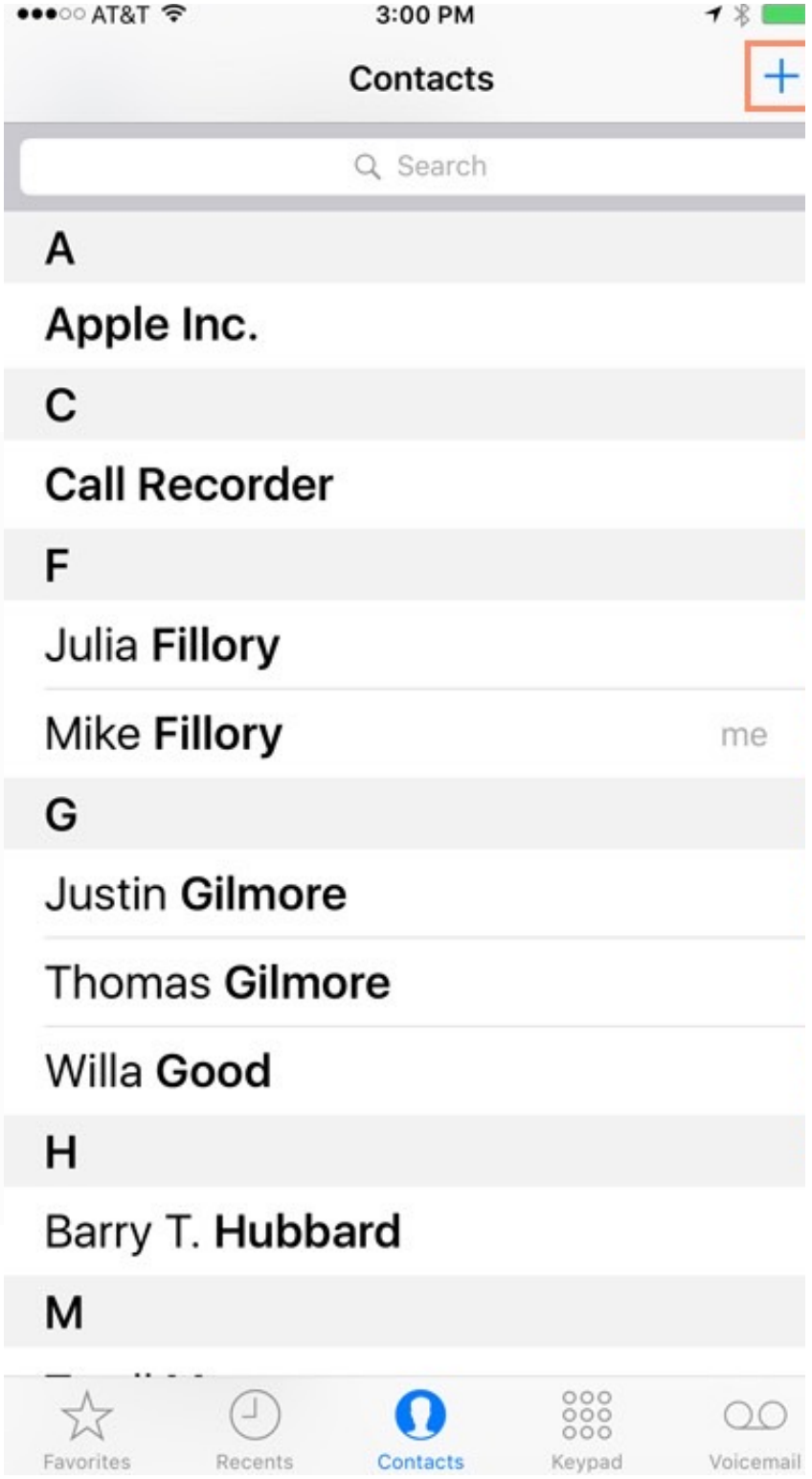
What's a database management system (DBMS)?

What are the core ideas?

What is a Database?

	A	B	C	D	E	F	G	H	I	J	K
1	color	date	slug	title	lshow	link	readings	optional	assigned	ashow	due
2	white	21-Jan	Intro + ER Models			https://githu	Ch 1, 2		HW 0		
3	#e7f8ff	28-Jan	ER Models				Ch 2		<a href="https://github.com/w4111/hw1-s22">H	0	HW 0
4	#e7f8ff	4-Feb	Data Models				Ch 3	optional: <a		0	HW1 Part1
5	#e7f8ff	11-Feb	Data Models + ER->Relational				Ch 3	optional: <a		0	Project 1 Part 1 approval phase
6	#f2f9ed	18-Feb	Relational Algebra				Ch 4		<a href="https://github.com/w4111/project1">Pr	0	Project 1 Part 1 approval phase
7	#f2f9ed	25-Feb	SQL: Basics				Ch 5			0	HW1 Part 2
8	#f2f9ed	4-Mar	SQL: Advanced				Ch 5			0	HW2
9	white	11-Mar	Midterm	one 8x11 page cheat sheet both sides					<a href="https://github.com/w4111/project1/blol	0	
10	white	18-Mar	HOLIDAY							0	Project 1 Part 2
11	#edf3f9	25-Mar	APIs				Ch 6			0	
12	#edf3f9	1-Apr	Data Quality	Normalization and data errors			Ch 19		<a href="https://github.com/w4111/hw4-s22">H	0	HW3
13	#ddf9ff	8-Apr	Physical Design				Ch 8		<a href="https://github.com/w4111/project2_s22	0	
14	#ddf9ff	15-Apr	Query Processing				Ch 12			0	Project 1 Part 3
15	#ddf9ff	22-Apr	Transactions				Ch 16, 18			0	
16	white	29-Apr	Data Pipelines							0	HW 4
17	white	13-May	Exam 2 (Cum	one 8x11 page cheat sheet both sides						0	Project 2
18											
19											
20											

What is a Database?



What is a Database?

```
2012-01-04 00:01:23,180 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving block blk_-2281137920769
010
2012-01-04 00:01:23,184 INFO org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace: src: /127.0.0.1:32981,
 cliID: DFSClient_-603743753, offset: 0, srvID: DS-292194659-127.0.1.1-50010-1324763300176, blockid: blk_-228113
2012-01-04 00:01:23,185 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: PacketResponder 0 for block blk_-2
2012-01-04 00:01:23,291 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving block blk_37660314352523
10
2012-01-04 00:01:23,293 INFO org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace: src: /127.0.0.1:32982,
cliID: DFSClient_-603743753, offset: 0, srvID: DS-292194659-127.0.1.1-50010-1324763300176, blockid: blk_37660314
2012-01-04 00:01:23,293 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: PacketResponder 0 for block blk_37
2012-01-04 00:01:23,324 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving block blk_-8044922265890
010
2012-01-04 00:01:23,326 INFO org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace: src: /127.0.0.1:32983,
liID: DFSClient_-603743753, offset: 0, srvID: DS-292194659-127.0.1.1-50010-1324763300176, blockid: blk_-80449222
2012-01-04 00:01:23,327 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: PacketResponder 0 for block blk_-8
2012-01-04 00:01:23,409 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving block blk_-9657937572621
10
2012-01-04 00:01:23,411 INFO org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace: src: /127.0.0.1:32984,
, cliID: DFSClient_-603743753, offset: 0, srvID: DS-292194659-127.0.1.1-50010-1324763300176, blockid: blk_-96579
2012-01-04 00:01:23,411 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: PacketResponder 0 for block blk_-9
2012-01-04 00:01:23,433 INFO org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace: src: /127.0.0.1:50010,
 cliID: DFSClient_-2054881890, offset: 0, srvID: DS-292194659-127.0.1.1-50010-1324763300176, blockid: blk_-96579
2012-01-04 00:01:23,494 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving block blk_54159109576590
10
2012-01-04 00:01:23,498 INFO org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace: src: /127.0.0.1:32987,
, cliID: DFSClient_-2054881890, offset: 0, srvID: DS-292194659-127.0.1.1-50010-1324763300176, blockid: blk_54159
2012-01-04 00:01:23,498 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: PacketResponder 0 for block blk_54
2012-01-04 00:01:23,523 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving block blk_-5517241460358
```

What is a Database?

What is a Database?

Lots of
Structured data

Database Management System (DBMS)

A system to store, manage and access databases

Database Management System (DBMS)

System to safely and reliably store lots of persistent structured data and is convenient for multiple users to efficiently access and modify.

Is a script a DBMS?

Javascript/Python Script

- Data stored in variables (RAM)
- Very fast access
- Data structures (lists, dicts, tuples)

Is Excel a DBMS?

Microsoft office security

Visually access/modify/compute over data cells

Click save to store persistently

Is the file system a DBMS?

Manages files that are persistently stored on disk

Open/read/seek/write access to files

Access via file names

Access control via permissions

Is the file system a DBMS?

You and a friend edit the same text file

Save at the same time

What happens?

1. Your changes survive
2. Friend's changes survive
3. Both changes survive
4. No changes survive
5. ¯_(ツ)_/¯

Is the file system a DBMS?

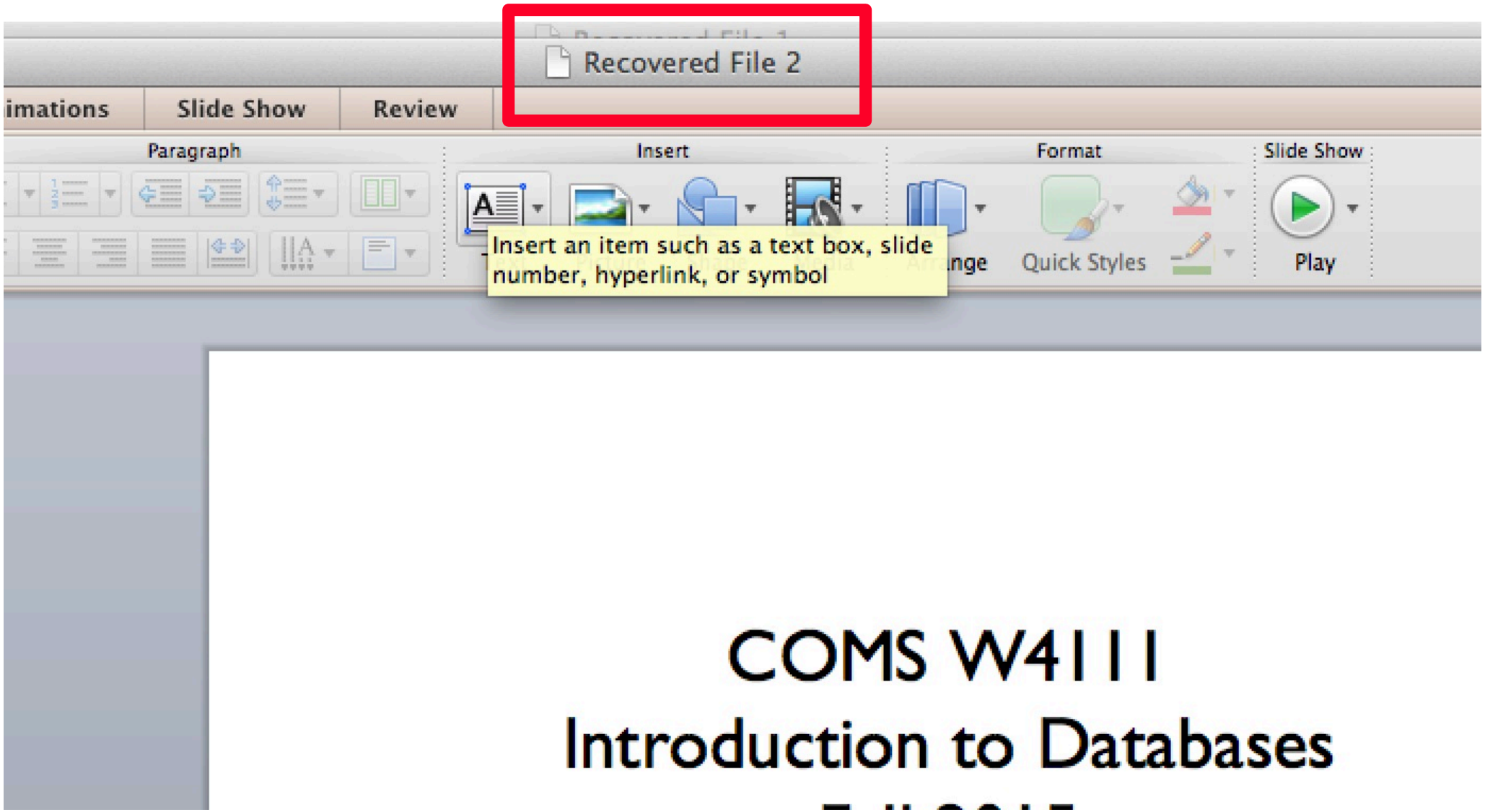
You edit a text file

Computer crashes

What happens?

1. All changes survive
2. No changes survive
3. Changes from last save survive
4. ¯_(ツ)_/¯

Is the file system a DBMS?

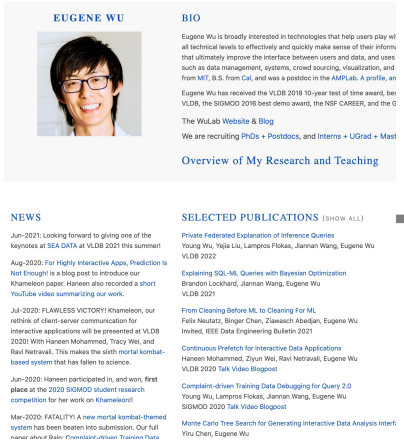


Who… would ever do this?

Real $1B+ Companies…

Store extracted data in a file

Every change → rewrite the file

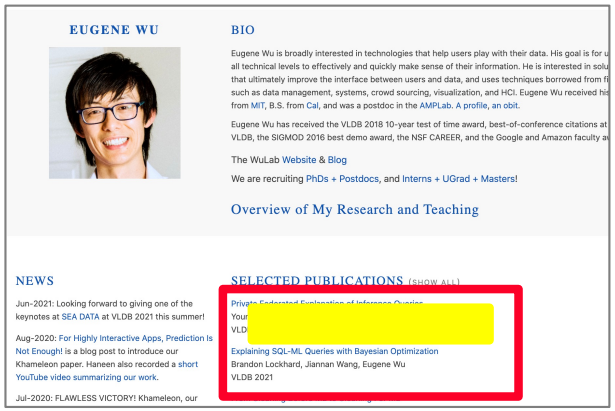


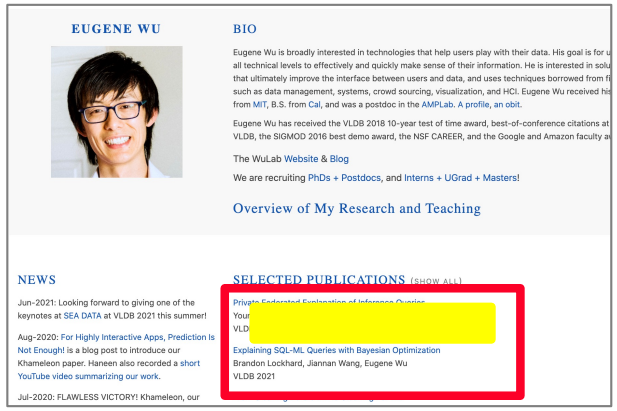
Text extraction and transform tasks

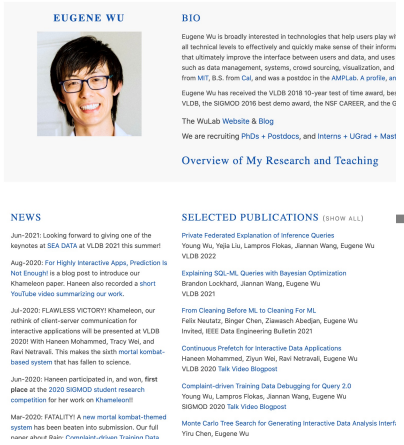
```
[{                              [{
  obj: region                     obj: authorlist
  box: 36,34,100,120              box: 40,40,100,20
  text: "Private F.."             text: "Eugene W.."
 }, { .. }, ..]                  }, { .. }, ..]
```

Browser

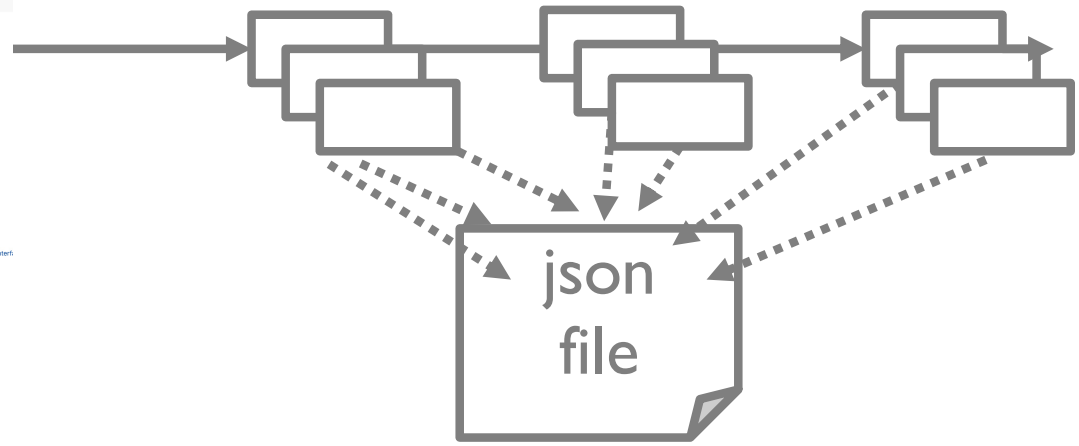
New Tab







Annotation tasks



Want *Guarantees* from DBMS

You want to write a hot new app on a DBMS.
What do you *not* want to worry about?

Failures disk, machine, human, corruption, deity

Lots of users concurrency, scaling, responsiveness

Ad-hoc data access arbitrary queries

Data formats csv? tsv? custom format?

Database Management System (DBMS)

System to safely and reliably store lots of persistent structured data and is convenient for multiple users to efficiently access and modify.

Database Management System (DBMS)

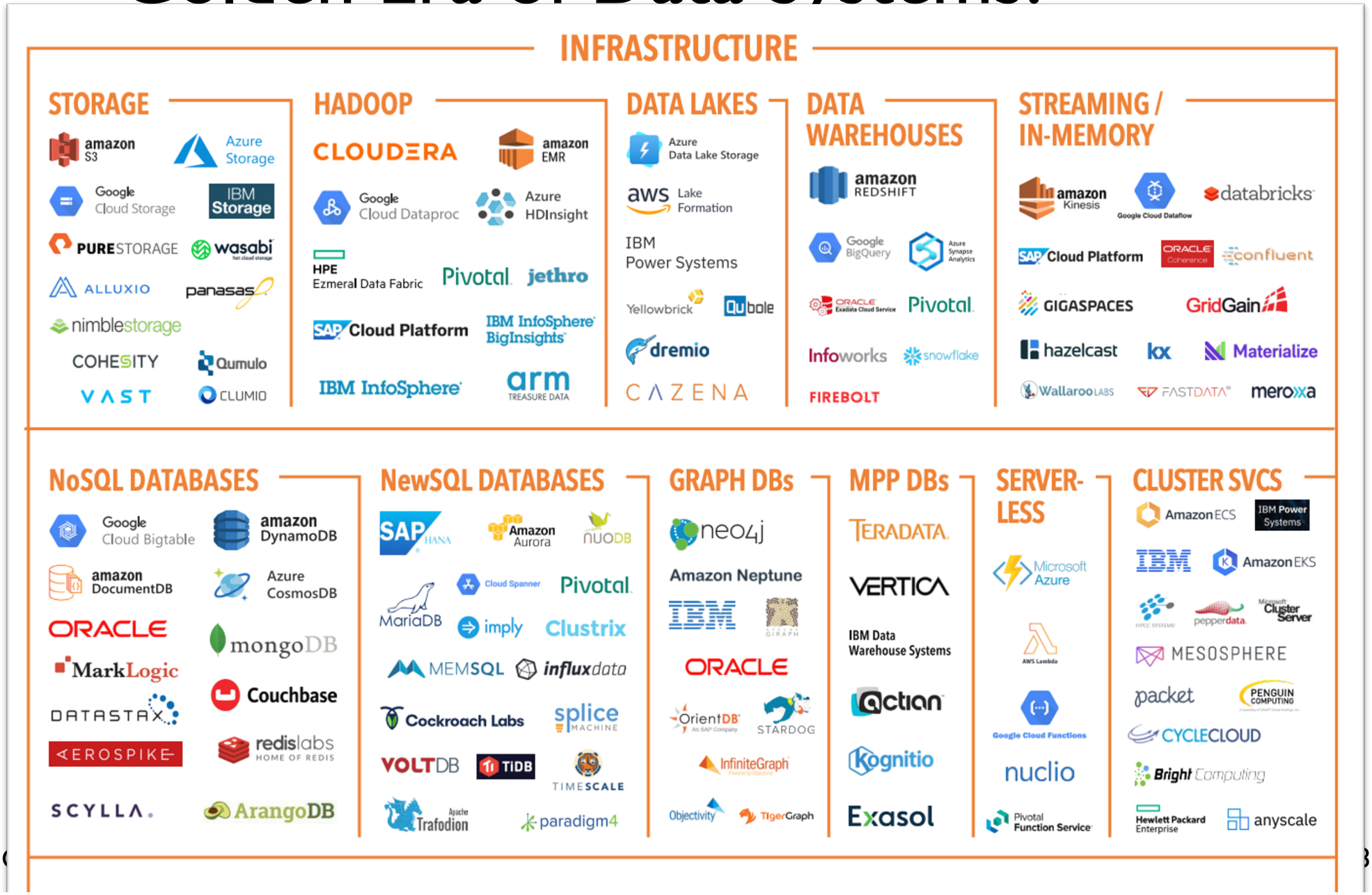
Safe	Consistent and correct data after failures
Reliable	99.99+% Uptime
Lots	>>RAM (terabytes)
Persistent	Lives longer than DBMS application
Convenient	Physical Independence. Declarative.
Multiple Users	Concurrent access. Access control.
Efficient	*Fast*: 100k+ queries / sec

Encompasses most of CS

OS	DBMS directly manages hardware
Languages	SQL is a domain specific language
Theory	Algorithms, models, NP-complete
AI/ML	Knowledge Discovery, KDD
Logic	Relational Algebra = 1st order logic

Scalable Computer Science

Golden Era of Data Systems!



2 Key Concepts

Data Independence

Declarative Languages

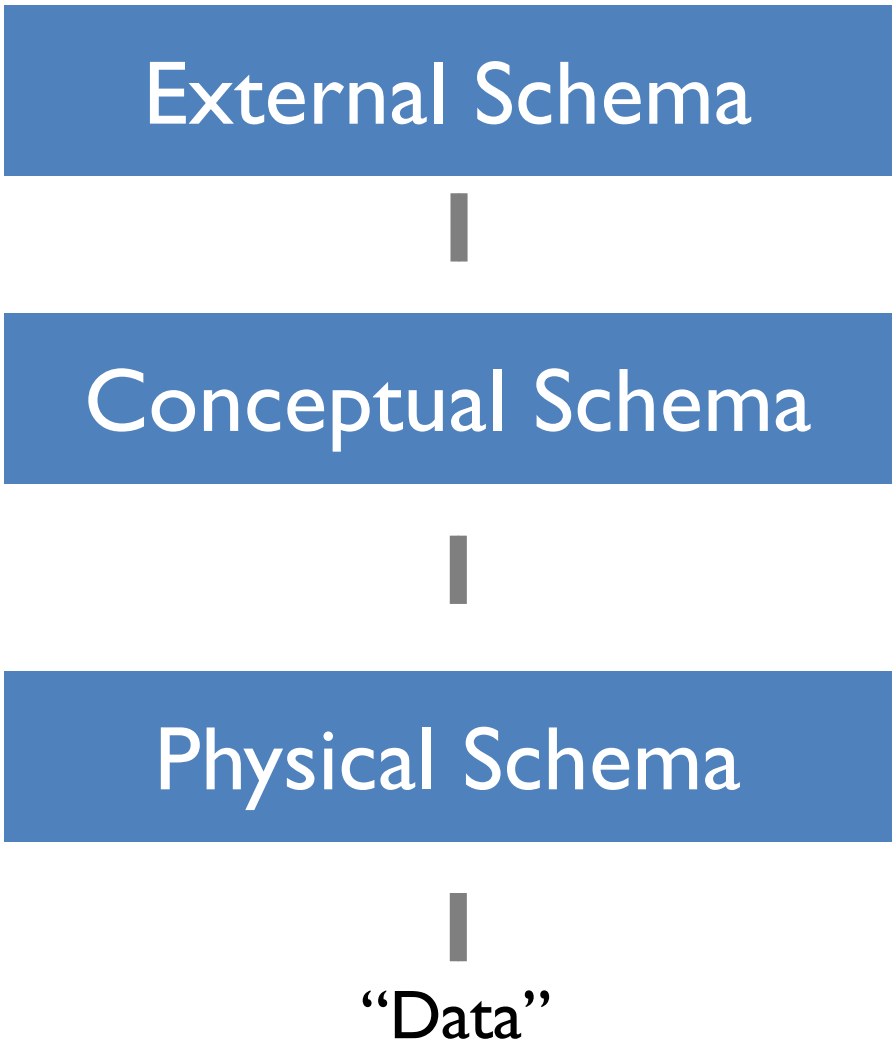
Serve to insulate application programmers from the system implementation

Data Independence

External Schema
Describe how users see data

Conceptual Schema
Describes logical structure

Physical Schema
Describes files, formats, indexes



Example App: Guuber

Users(uid int, name str, age int)

Drivers(did int, name str)

Rides(uid int, did int, distance float, drive_time float)

Data Independence

UID	Name	Age
0	Eugene	17
1	Luis	20
2	Ken	30

```
0,Eugene,17
1,Luis,20
2,Ken,30
```

CSV File

What is the number of adults?

Data Independence

UID	Name	Age
0	Eugene	17
1	Luis	20
2	Ken	30

0,Eugene,17
1,Luis,20
2,Ken,30

CSV File

```
n = 0
for line in csv_file:
  attributes = line.split(“,”)
  if attributes[2] >= 18:
    n += 1
```

Data Independence

UID	Name	Age
0	Eugene	17
1	Luis	20
2	Ken	30

0 Eugene 17
1 Luis 20
2 Ken 30

TSV File

```
n = 0
for line in csv_file:
  attributes = line.split(“,”)
  if attributes[2] >= 18:
    n += 1
```

Data Independence

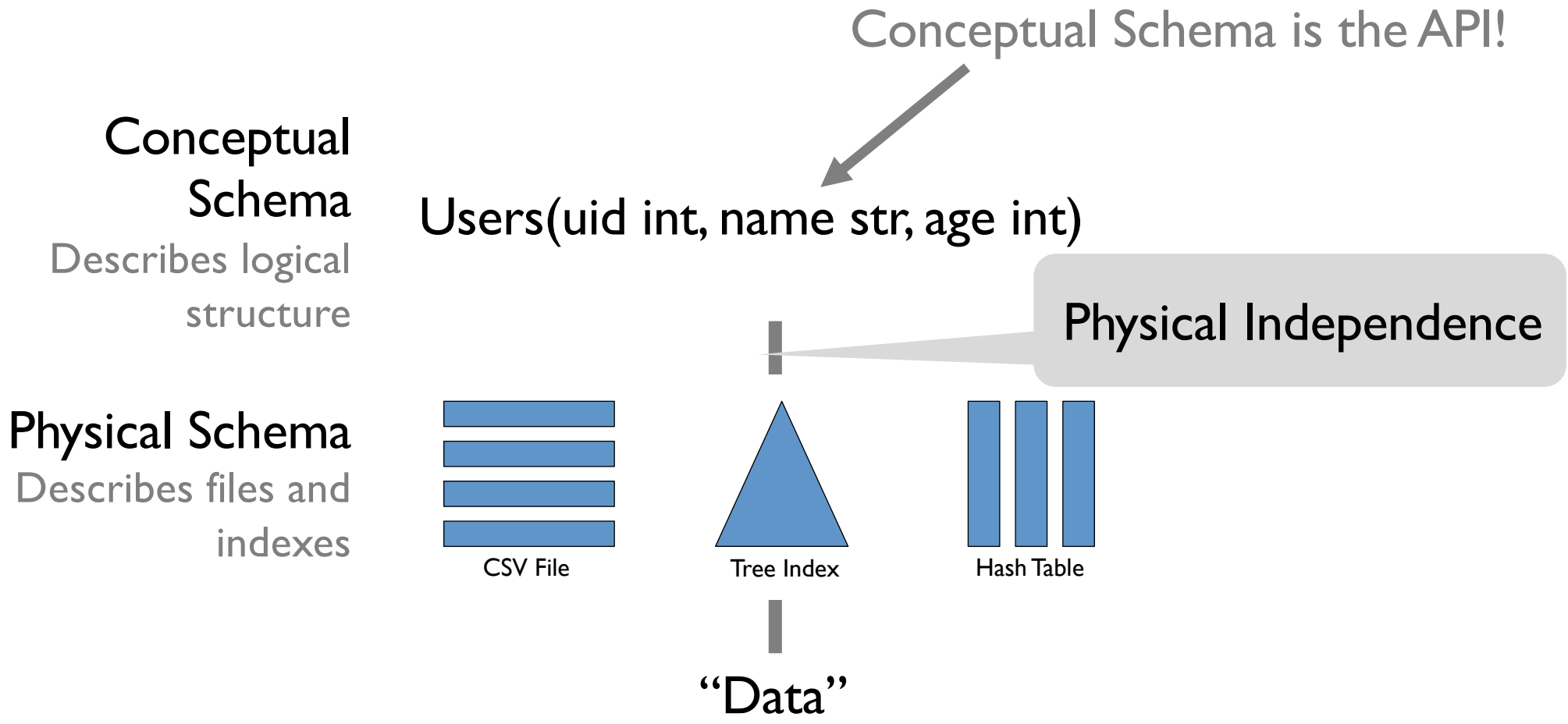
UID	Name	Age
0	Eugene	17
1	Luis	20
2	Ken	30

0,1,2
Eugene,Luis,Ken
17,20,30

Columnar File

```
n = 0
For line in csv_file:
  attributes = line.split(“,”)
  if attributes[2] >= 18:
    n += 1
```

Data Independence



Data Independence

Users(uid int, name str, age int)

"Welcome back Mr. Wu"

Data Independence

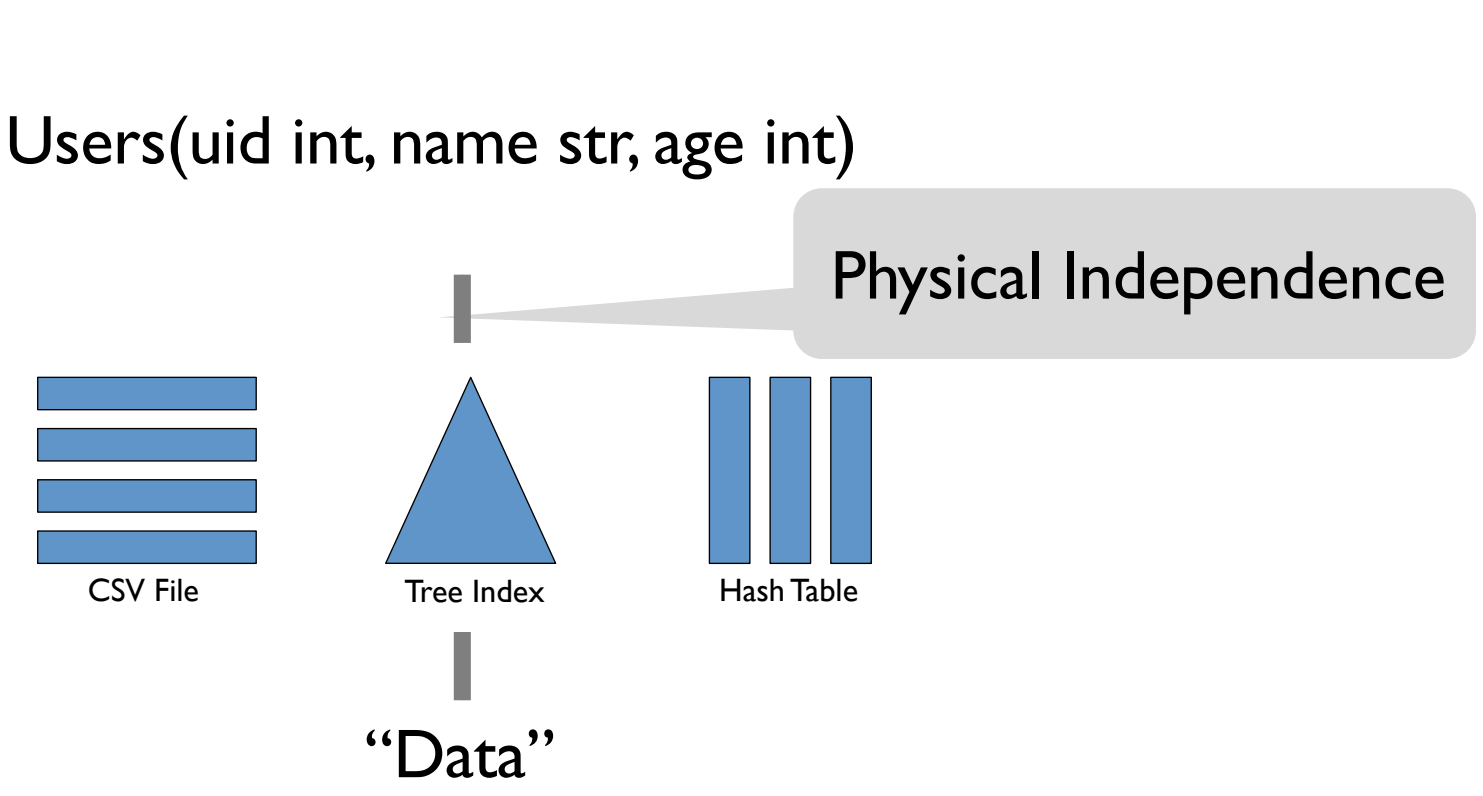
Users(uid int, fname str, lname str, age int)

"Welcome back Mr. Wu"

Data Independence

Conceptual Schema
Describes logical structure

Physical Schema
Describes files and indexes

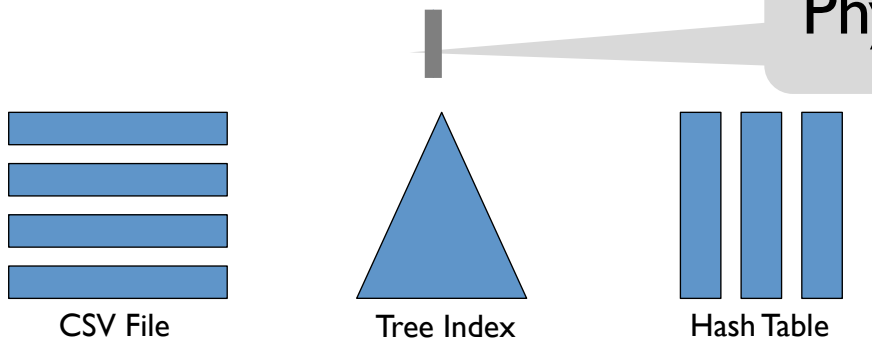


Data Independence

Conceptual Schema
Describes logical structure

Users(uid int, fname str, lname str, age int)

Physical Schema
Describes files and indexes



Data Independence

External Schema
Describe how users see data

Users(uid int, name str, age int)

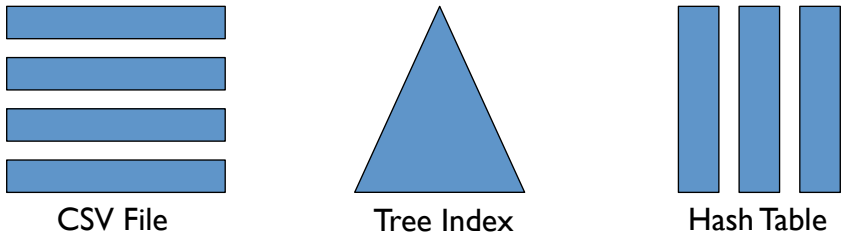
Logical Independence

Conceptual Schema
Describes logical structure

Users(uid int, fname str, lname str, age int)

Physical Independence

Physical Schema
Describes files and indexes



"Data"

Data Independence

Physical Independence

Protection from changes in physical structure of data

Logical Independence

Protection from changes in logical structure of data

One of most important properties of a DBMS

Declarative Interface

Mechanism that enables data independence

Insulates programmer from physical schema

Rather than a list of functions,

the API is a *query language*

Declarative Interface

What you want, **not how to do it.**

"Make me a sandwich"

Buy from pb&j store
Make BLT
½ Tuna
Veggie

"Take two slices of wheat bread out of the 2nd shelf, put them next to each other…"

What if on 1st shelf?
Out of wheat bread?
No counter space?

Declarative Interface

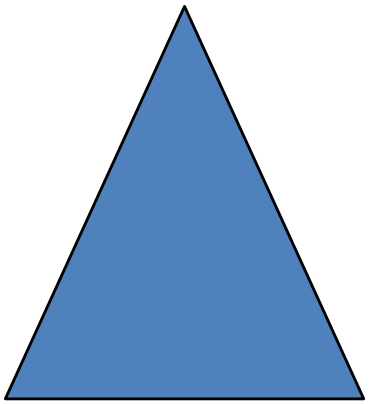
"I want all highly rated fast drivers"

Declarative Interface

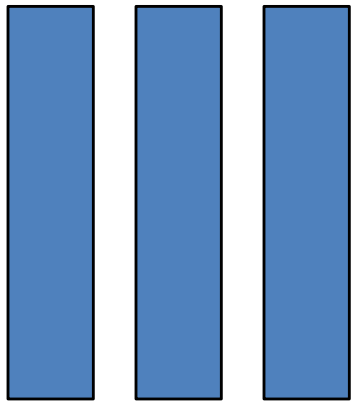
SELECT name FROM users WHERE rating > 8

DBMS

CSV File



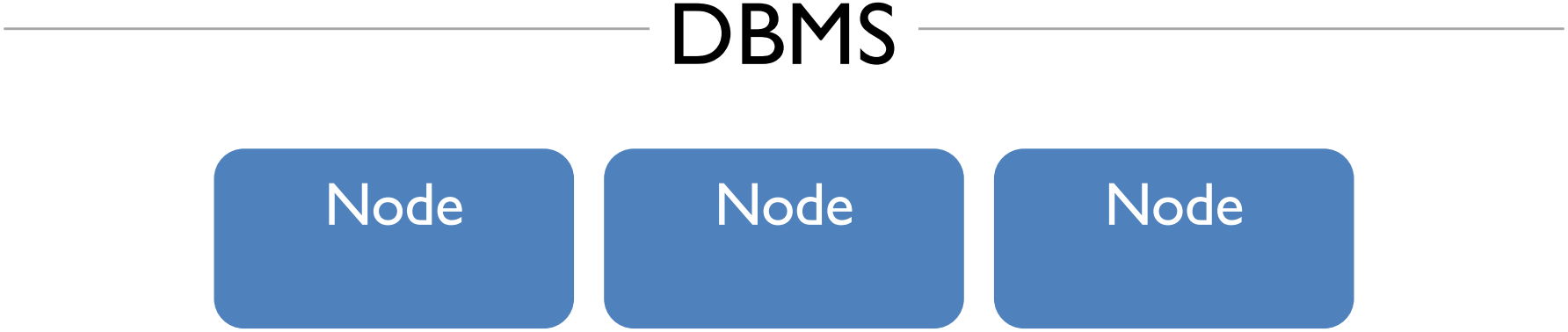
Tree Index



Hash Table

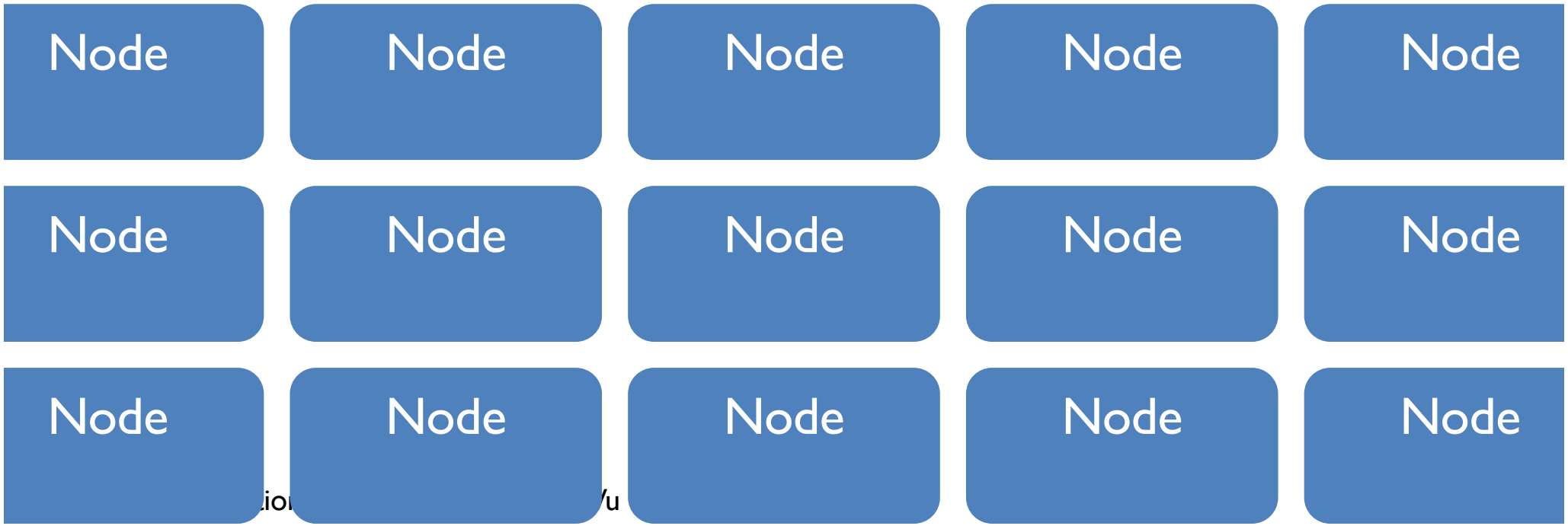
Declarative Interface

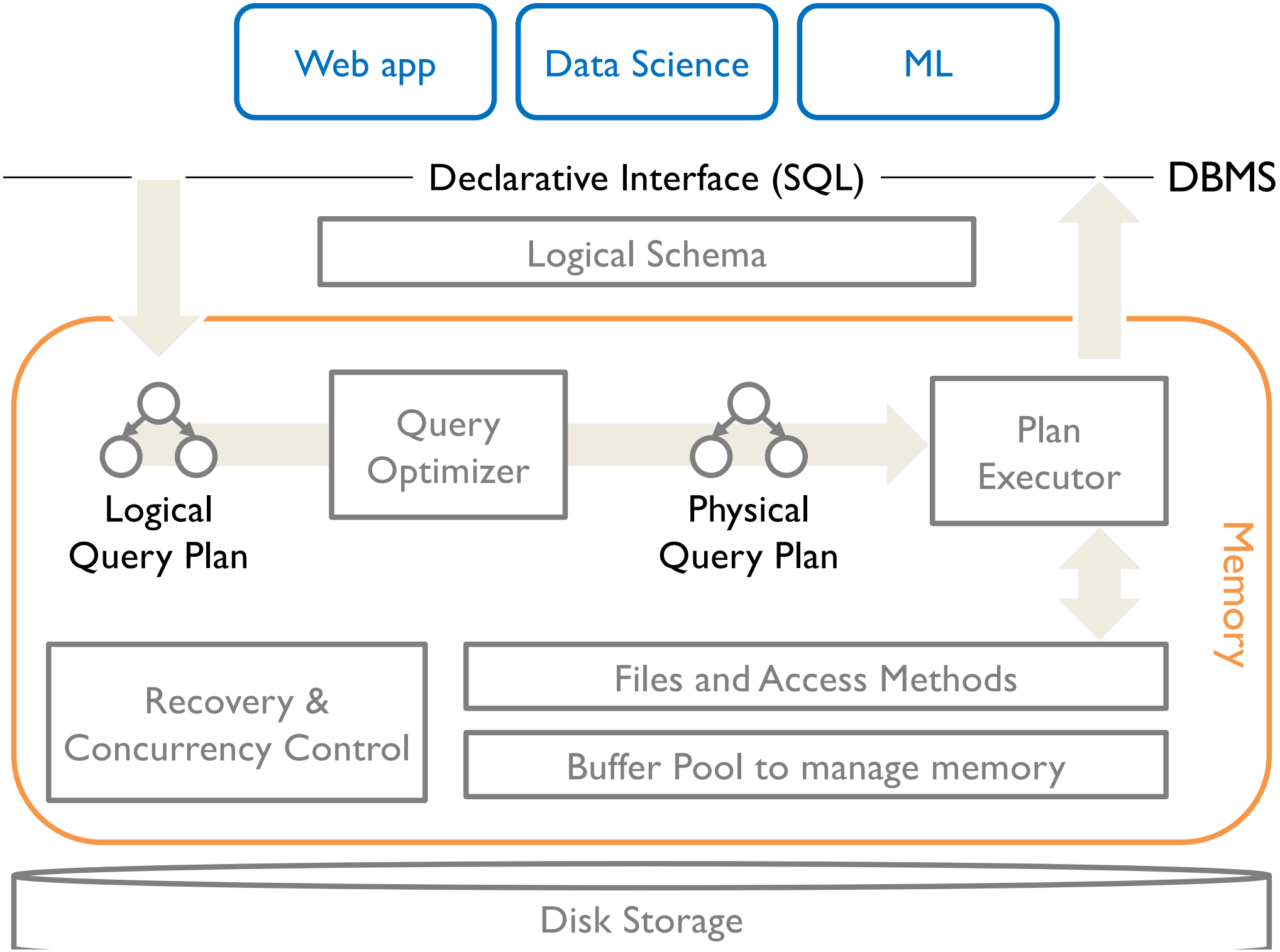
SELECT name FROM users WHERE rating > 8

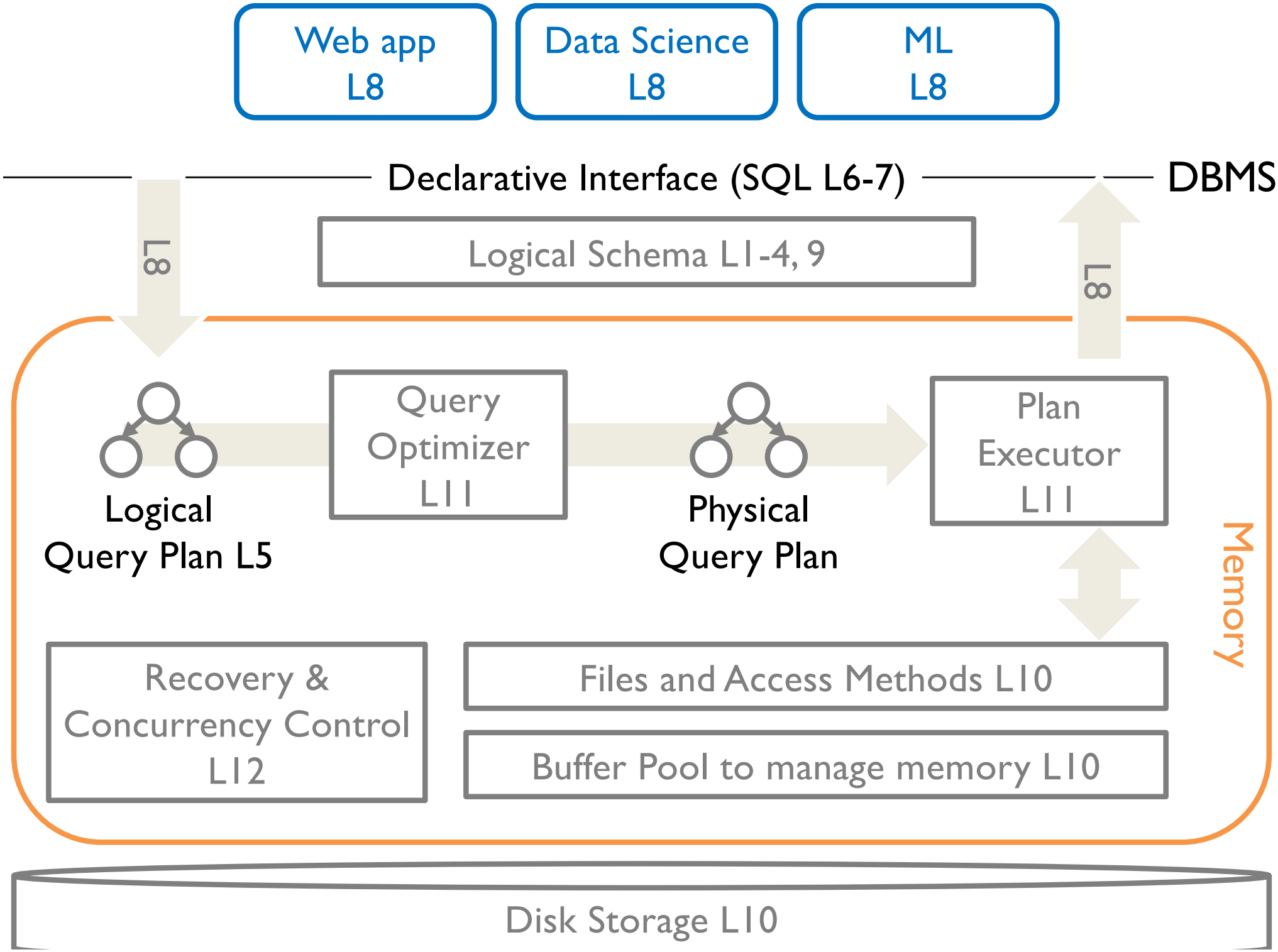


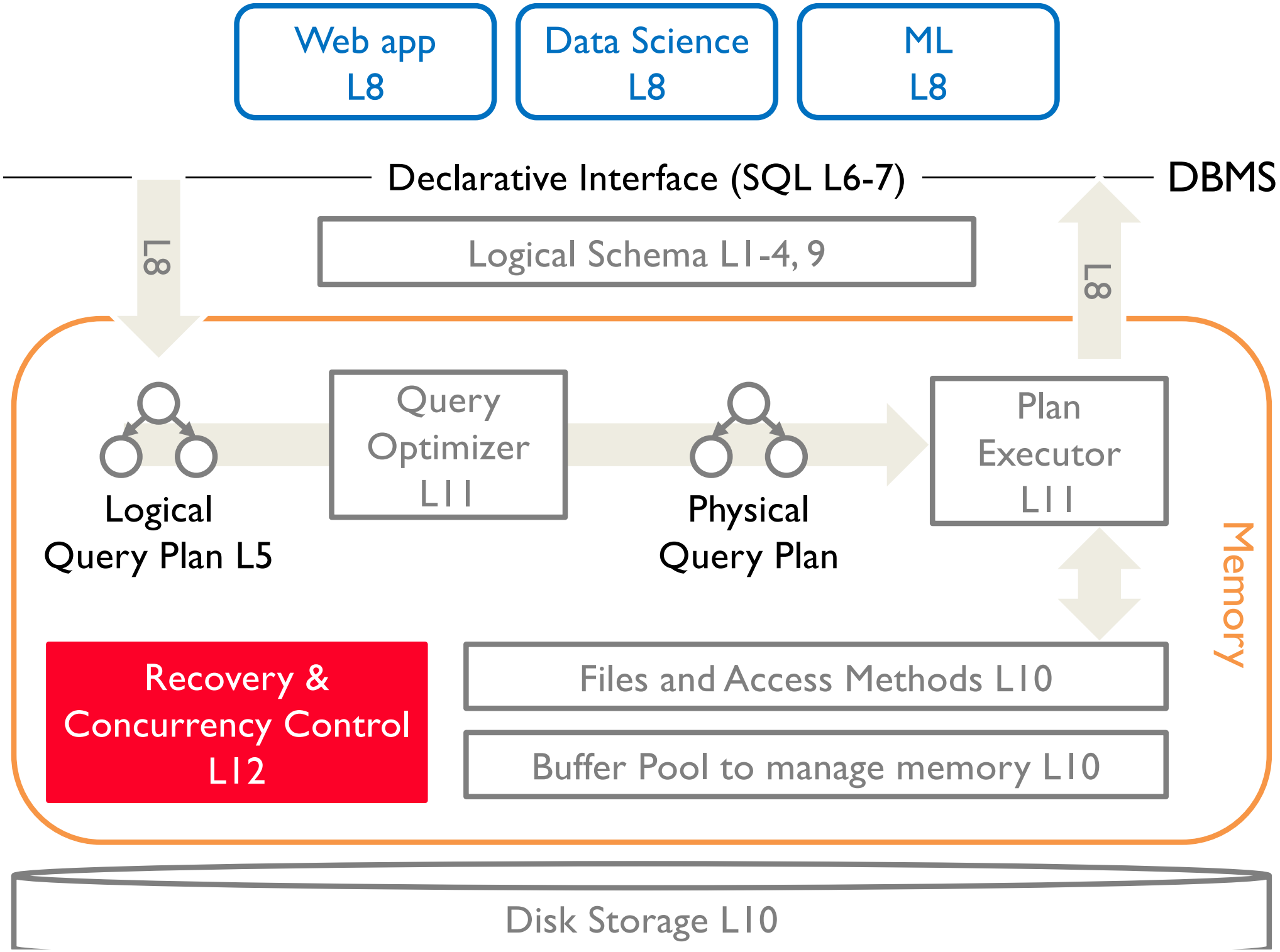
Declarative Interface

SELECT name FROM users WHERE rating > 8




Web app
Data Science
ML
Declarative Interface (SQL)
DBMS
Logical Schema
Query Optimizer
Logical Query Plan
Physical Query Plan
Plan Executor
Memory
Recovery & Concurrency Control
Files and Access Methods
Buffer Pool to manage memory
Disk Storage


Web app
L8
Data Science
L8
ML
L8
Declarative Interface (SQL L6-7)
DBMS
L8
Logical Schema L1-4, 9
L8
Query
Optimizer
L11
Plan
Executor
L11
Logical
Query Plan L5
Physical
Query Plan
Memory
Recovery &
Concurrency Control
L12
Files and Access Methods L10
Buffer Pool to manage memory L10
Disk Storage L10


Web app
L8
Data Science
L8
ML
L8
Declarative Interface (SQL L6-7)
DBMS
L8
Logical Schema L1-4, 9
L8
Query
Optimizer
L11
Plan
Executor
L11
Logical
Query Plan L5
Physical
Query Plan
Memory
Recovery &
Concurrency Control
L12
Files and Access Methods L10
Buffer Pool to manage memory L10
Disk Storage L10

Concurrency Control

Want to let many users use database concurrently.

How to ensure they run correctly?

Concurrency Control

Want to let many users **use** database concurrently.

How to ensure they run correctly?

What does "**use**" mean?

Run transaction, which groups all of the user's DBMS actions together. They either all run, or none run.

```
Begin;
<read beth's account>
<deduct from beth's account>
<increase eugene's account>
Commit;  (or Abort;)
```

Concurrency Control

Want to let many users use database concurrently.

How to ensure they run **correctly**?

What does "**Correctly**" mean?

Transactions run as if only one DBMS user at a time.

DBMS maintains all integrity constraints over the database.

Recovery

If the DBMS crashes, can recover to a **Correct** state.

What does "**Correct**" mean?

All committed transactions are preserved in the database.

Undo all incomplete (uncommitted) transactions.

How?

DBMS keeps a log of all actions each transaction performs.

Track whether an action is committed or uncommitted.

A bit about the class

Next Up

HW0 is out.

Due by 1/26 11:59PM.

No Late Submission Accepted

0 in class if not submitted in time

Class Information: Prerequisites

COMS W3134 - *Data Structures in Java* or
COMS W3137 - *Data Structures and Algorithms*

(equivalent courses taken elsewhere are acceptable as well)

Fluency in Python

Class Information: Lectures

Fridays

10-12:40PM 😱

501 Schermerhorn (when in person resumes)

w4111.github.io

COLUMBIA UNIVERSITY COMS W4111

INTRODUCTION TO DATABASES

Information

- Fri 10-12:40
  501 Schermerhorn
  3 units
- Syllabus
- Ed Discussion
- Provide Feedback
- Course Github

Staff

- Eugene Wu Instructor
  Weds 2-3PM
- Zachary Huang
- Jacob Fisher
- Ashwathy Menon
- Sughosh V Kaushik
- Rachel Halpern

Prereqs

- Required: Students are expected to be comfortable with data structures and Python.
- Required: COMS W3134, COMS W3137, or COMS

Overview

The goal of this class is two-fold. First, to introduce you to core database concepts (e.g., data modeling, logical design, SQL) so that you too can build a billion dollar application. Second, to teach enough about database engine internals (e.g., physical database design, query optimization, transaction processing) so you have a good sense of why queries may be running slowly/incorrectly. We will also discuss their relevance to systems used in industry.

Please do not ask me about the waitlist

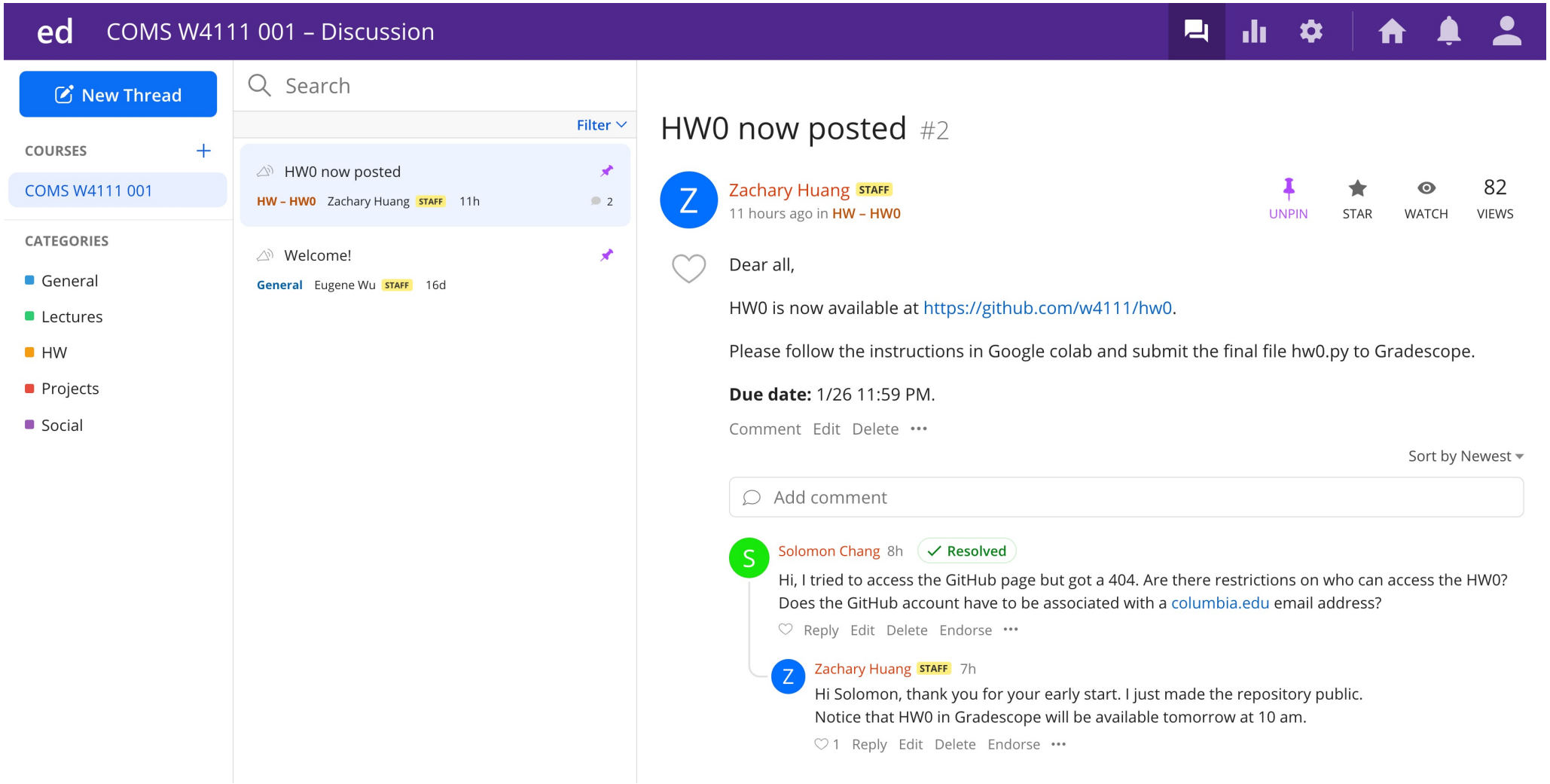
Announcements

- HW0 released.

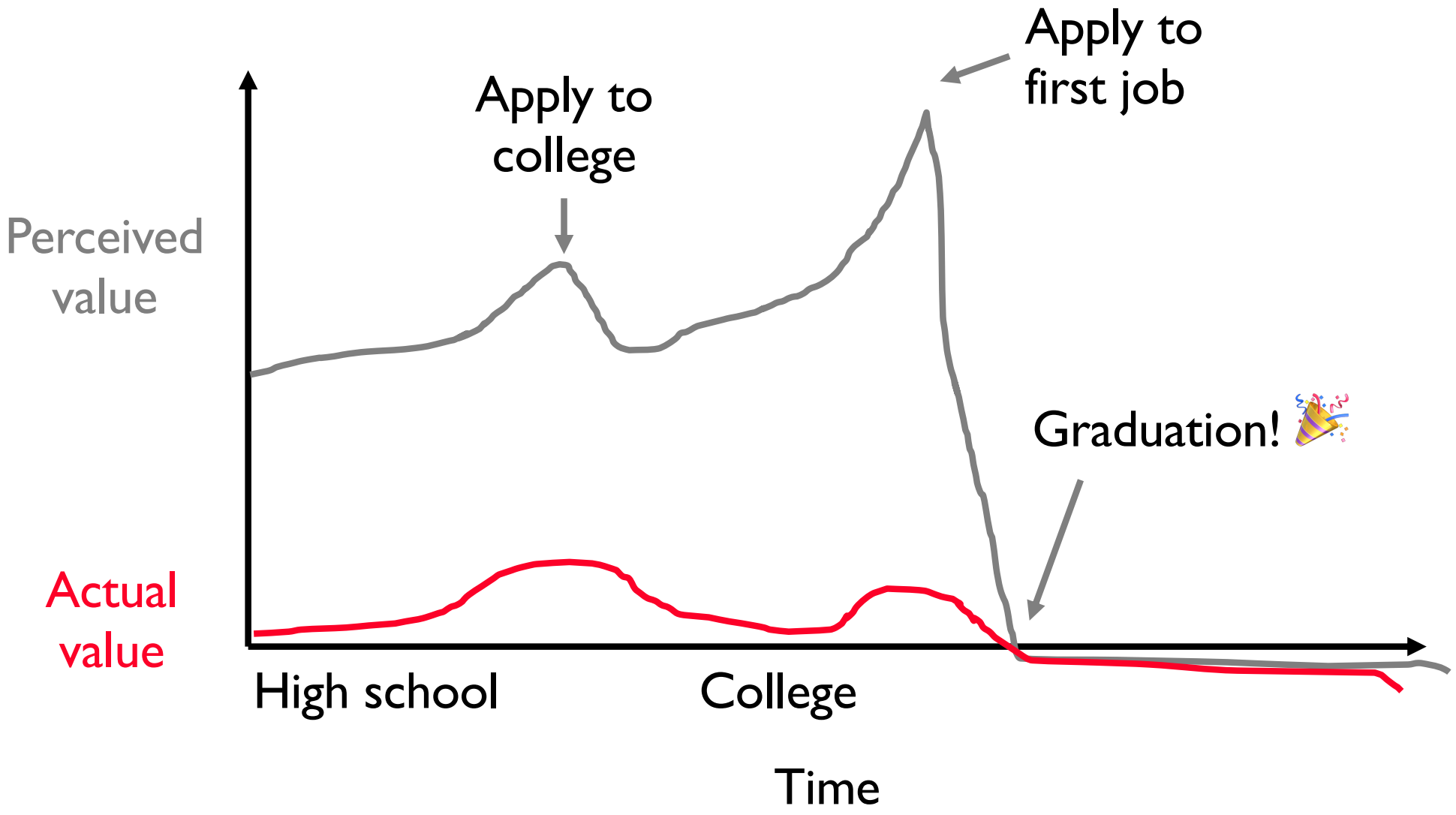
Schedule

Date	Topic	Assigned	Due
21-Jan	**Intro + ER Models** optional: Textbook Ch 1, 2	HW 0	
28-Jan	**ER Models** optional: Textbook Ch 2	HW 1 Project 1 Part 1. LOOK FOR TEAMMATE	HW 0
4-Feb	**Data Models** optional: What goes around comes around optional: NoSQL data modeling techniques optional: Textbook Ch 3		HW1 Part1
11-Feb	**Data Models + ER->Relational** optional: Original Relational Model paper optional: Textbook Ch 3		Project 1 Part 1 approval phase

Discussion Board



Grades. How do they work?



Grading Information

Midterm 1	25%
Midterm 2	40%
HW	15% (4 HWs equally weighed)
Project 1	15%
Project 2	5%
Extra credit	variable

Median grade: B or slightly higher.

Exam Dates

Midterm 1 3/11, format tbd

Midterm 2 5/13, format tbd

Makeup exams are not scheduled

Homework

Homework usually due at 10AM of due date.

Assignment will specify submission instructions.

No extensions or exceptions.

5 grace days for hws throughout the semester.

Can be applied to any assignment *unless otherwise specified*

After using all grace days, 25% grade deduction per day.

Don't need to tell us, staff will assign grace days in your favor

Check full details on web site under syllabus.

Projects (more details soon)

Two projects.

- Teams of two
- Run on cloud infrastructure
- Python & SQL

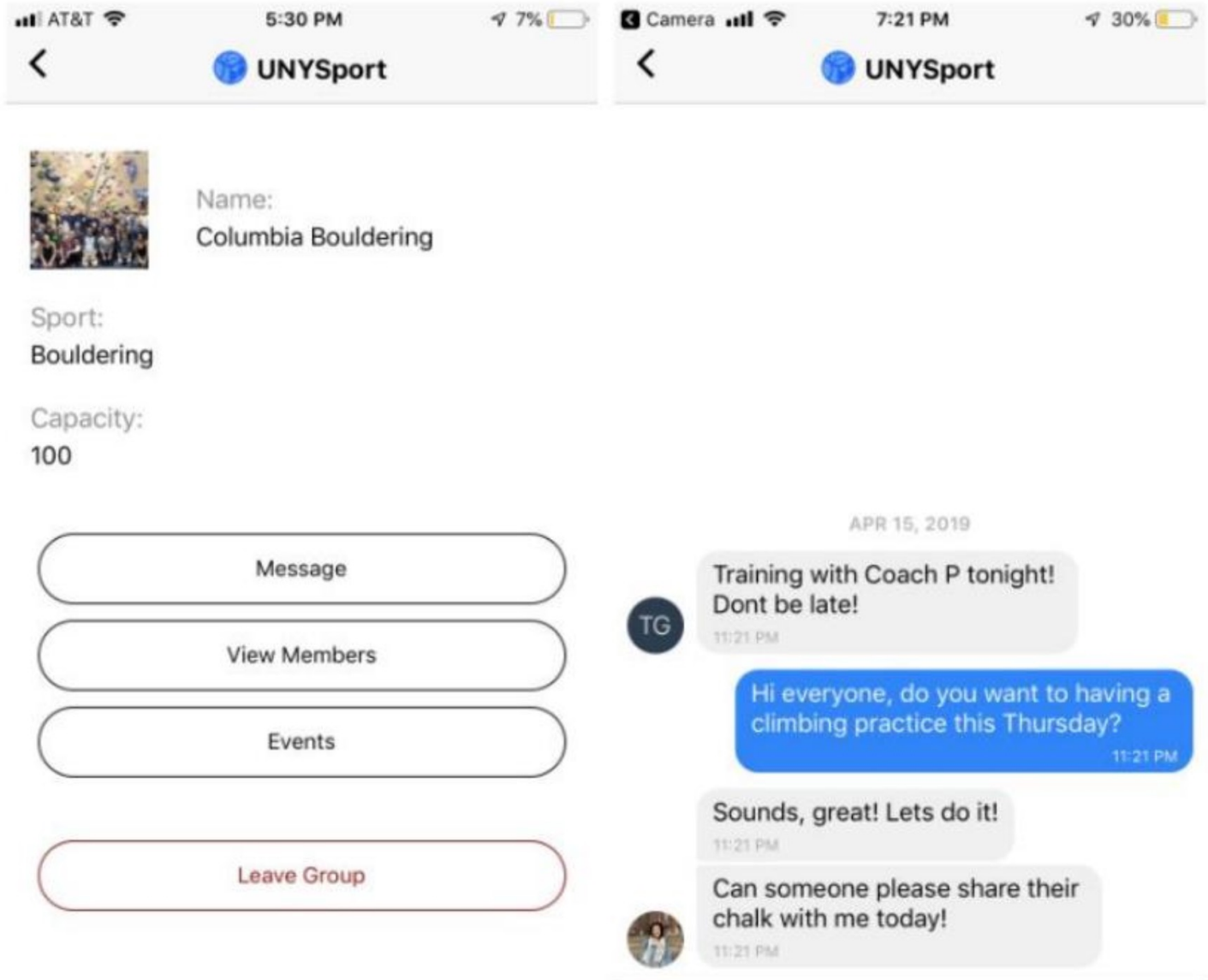
Project 1

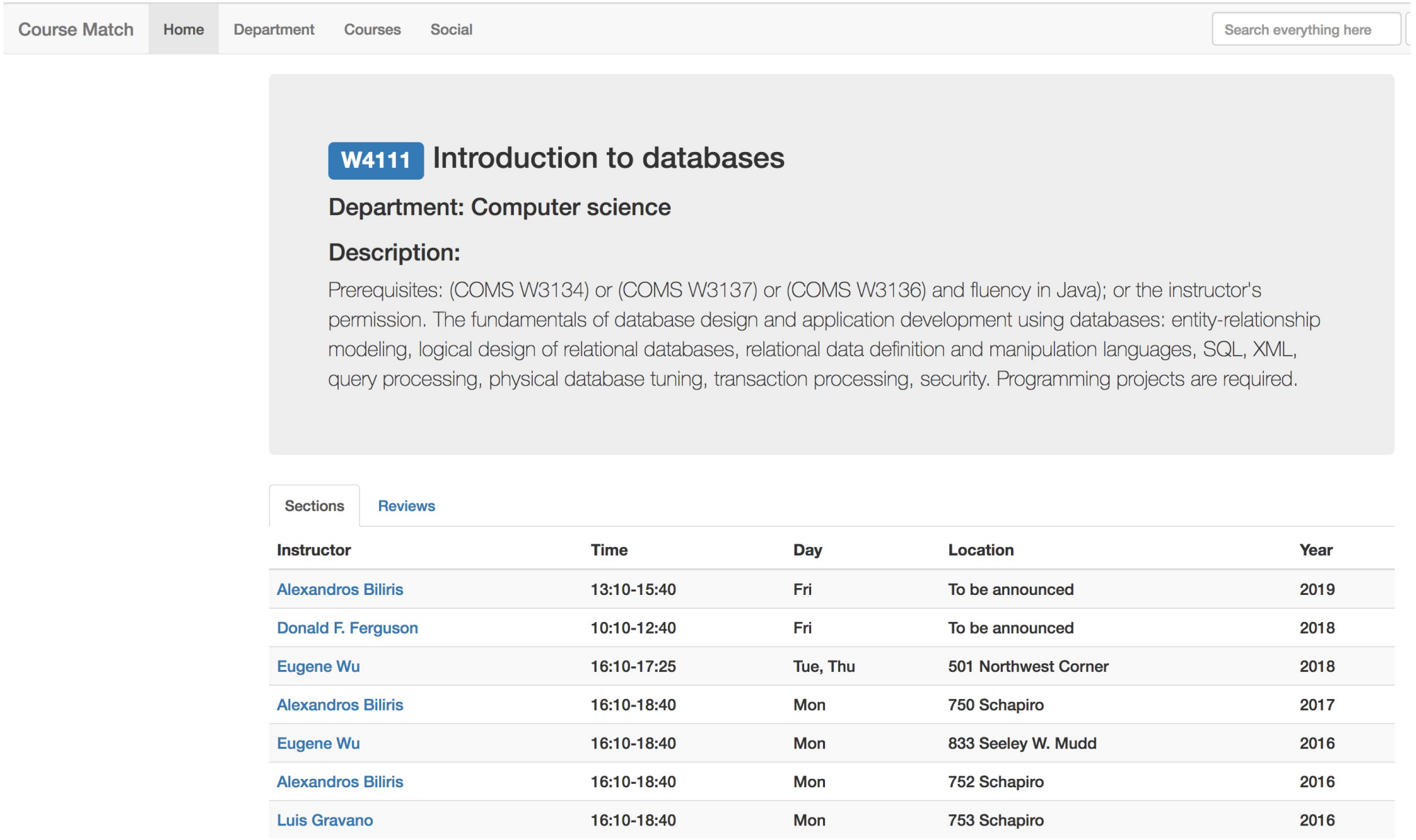
- Model and build your own database web application
- Explore "traditional" relational database features.
- Non-programming option

Project 2

- Do cool things with DBMSes (TBD)

Sports Community Mobile App



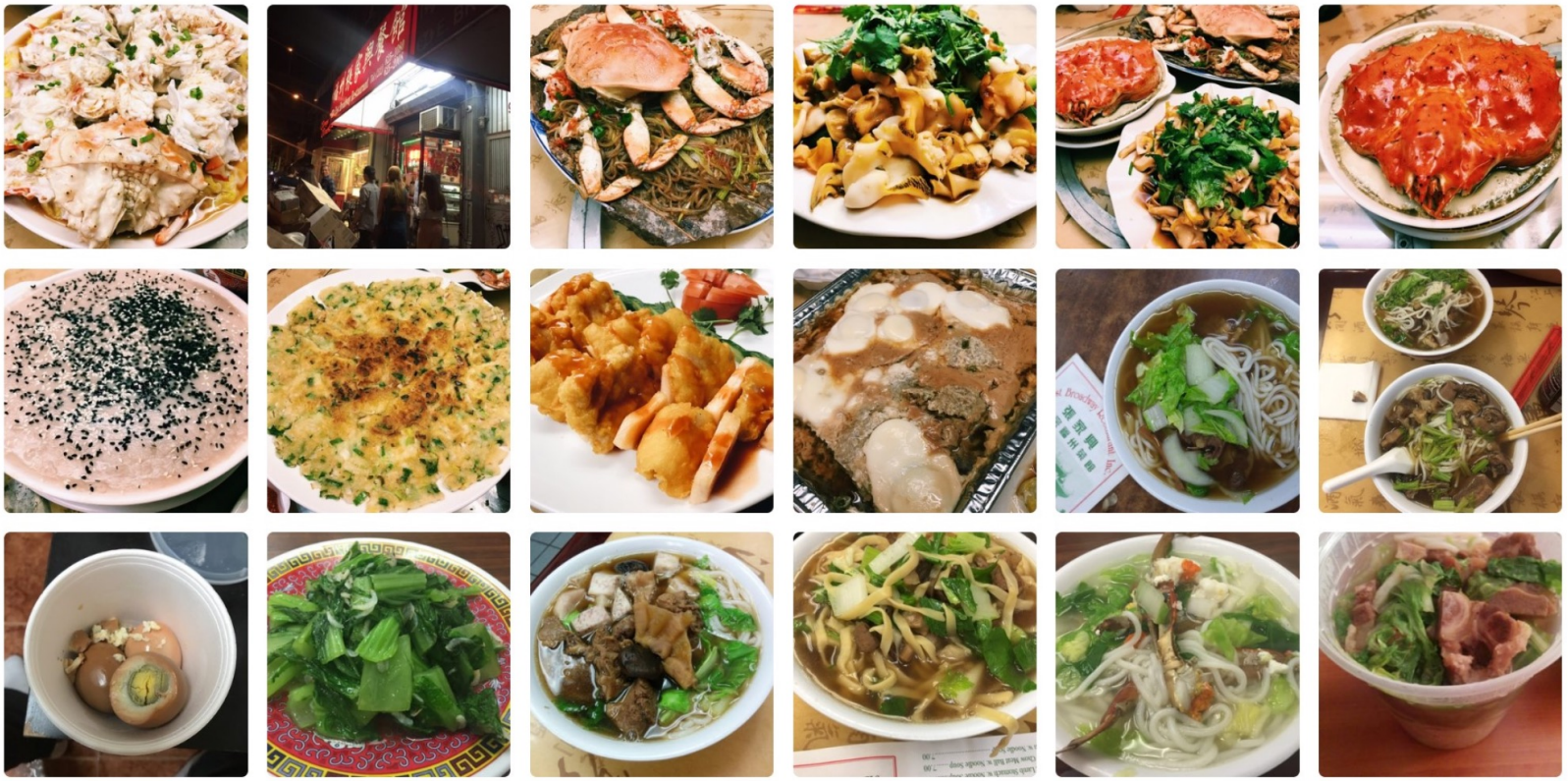

Course Match
Home
Department
Courses
Social
Search everything here
W4111 Introduction to databases
Department: Computer science
Description:
Prerequisites: (COMS W3134) or (COMS W3137) or (COMS W3136) and fluency in Java); or the instructor's permission. The fundamentals of database design and application development using databases: entity-relationship modeling, logical design of relational databases, relational data definition and manipulation languages, SQL, XML, query processing, physical database tuning, transaction processing, security. Programming projects are required.
Sections
Reviews
Instructor
Time
Day
Location
Year
Alexandros Biliris
13:10-15:40
Fri
To be announced
2019
Donald F. Ferguson
10:10-12:40
Fri
To be announced
2018
Eugene Wu
16:10-17:25
Tue, Thu
501 Northwest Corner
2018
Alexandros Biliris
16:10-18:40
Mon
750 Schapiro
2017
Eugene Wu
16:10-18:40
Mon
833 Seeley W. Mudd
2016
Alexandros Biliris
16:10-18:40
Mon
752 Schapiro
2016
Luis Gravano
16:10-18:40
Mon
753 Schapiro
2016

C-Food

Reviews Complaints Register Log In Log Out

C-Food: Your guide to clean NYC Restaurants

East Broadway Restaurant



Borough: manhattan

Address: 94 East Broadway

Health Investigation Score: 41/50

Average User Rating: 3.20

Domino's

Projects (cont.)

3 grace days total for project parts 1 and 2.

No extensions or exceptions for project part 3 submission.

After using all grace days, 25% grade deduction per late day.

Check full details on web site.

Extra Credit

Added *after* the curve

Does NOT affect those that don't do extra credit

Collaboration Policy

Read Syllabus on course site for allowed conduct

CS Dept academic honesty policies
http://www.cs.columbia.edu/education/honesty

We will not tolerate *any* cheating

Collaboration Policy

Discussing lectures and course material strongly encouraged

Homework and exams are *individual*. No exceptions
Any libraries or code however minor must be disclosed.

Projects are done in *teams*; no collaboration between teams.

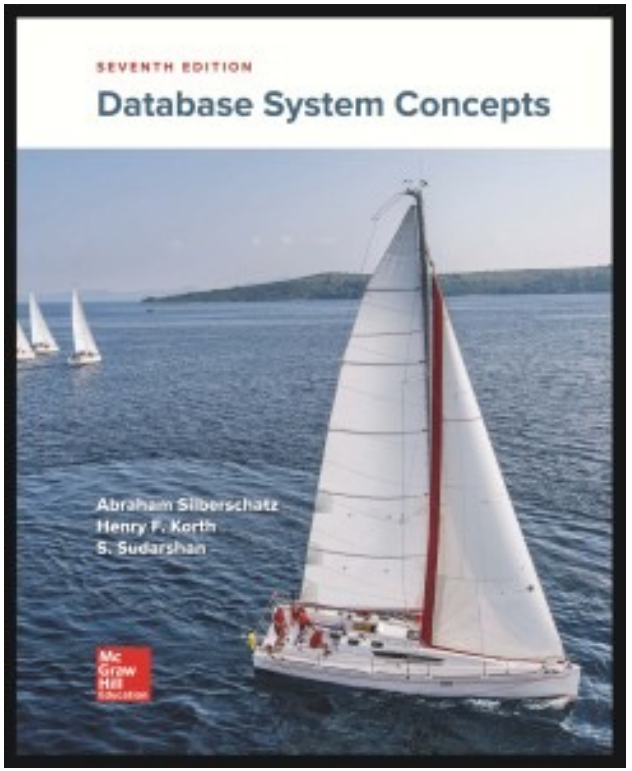
Contact the Professor Wu
right away if you have any questions or are falling behind.

Optional Textbook

Silberschatz et al.

Database System Concepts

7th ed



On-going Feedback

COLUMBIA UNIVERSITY COMS W4111

INTRODUCTION TO DATABASES

Information

- Fri 10-12:40
  501 Schermerhorn
  3 units
- Syllabus
- Ed Discussion
- Provide Feedback
- Course Github

Staff

- Eugene Wu Instructor
  Weds 2-3PM
- Zachary Huang
- Jacob Fisher

Overview

The goal of this class is two-fold. First, to introduce you to core database concepts (e.g., data mod
you too can build a billion dollar application. Second, to teach enough about database engine inte
design, query optimization, transaction processing) so you have a good sense of why queries may
will also discuss their relevance to systems used in industry.

Please do not ask me about the waitlist

Announcements

- HW0 released.

Schedule

Date	Topic	Assigned
21-Jan	**Intro + ER Models**	HW 0

Database Courses at Columbia

COMS W4111 - Intro to Databases

Prerequisites: CS3137 or CS3134; fluency in Python

Intro to DBMSes

Data Models

Relational Algebra

SQL

Applications + SQL

Normalization

Peek at DBMS internals:

- Storage and indexing
- Query optimization
- Transaction Processing

COMS W4112-Database Sys. Impl.

Prerequisites: CS3137 or CS3134; fluency in Python

Components of a Database System in Detail

Storage Methods and Indexing

Query Processing and Optimization

Materialized Views

Transaction Processing and Recovery

Parallel & Distributed DBMSes

Performance Considerations Beyond Disk I/Os

COMS E6111-Advanced Databases

Prerequisites: CS4111; fluency in Java or Python

Information Retrieval
Information Extraction
Web Search
Data Mining
Data Warehousing, OLAP, Decision Support

COMS E6xxx-DB Research Seminars

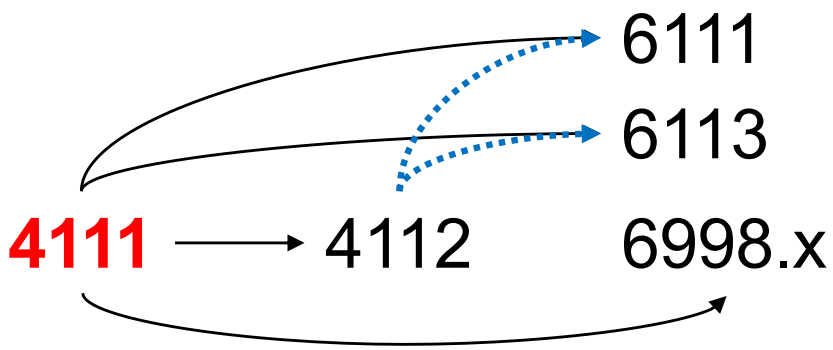
Prerequisites: CS4111; fluency in Java or Python

6113 Database Research Topics

w6113.github.io

6998.002 Systems for
Human Data Interaction

columbiaviz.github.io


6111
6113
4111
4112
6998.x

Data Management at Columbia


Luis Gravano


Kenneth Ross


Eugene Wu


Mihalis Yannakakis

`http://cudbg.github.io/`

Borrowed material from
Prof. Gravano
Prof. Hellerstein (Cal)
Prof. Madden & Stonebraker (MIT)

w4111.github.io

DO HOMEWORK 0!