

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.

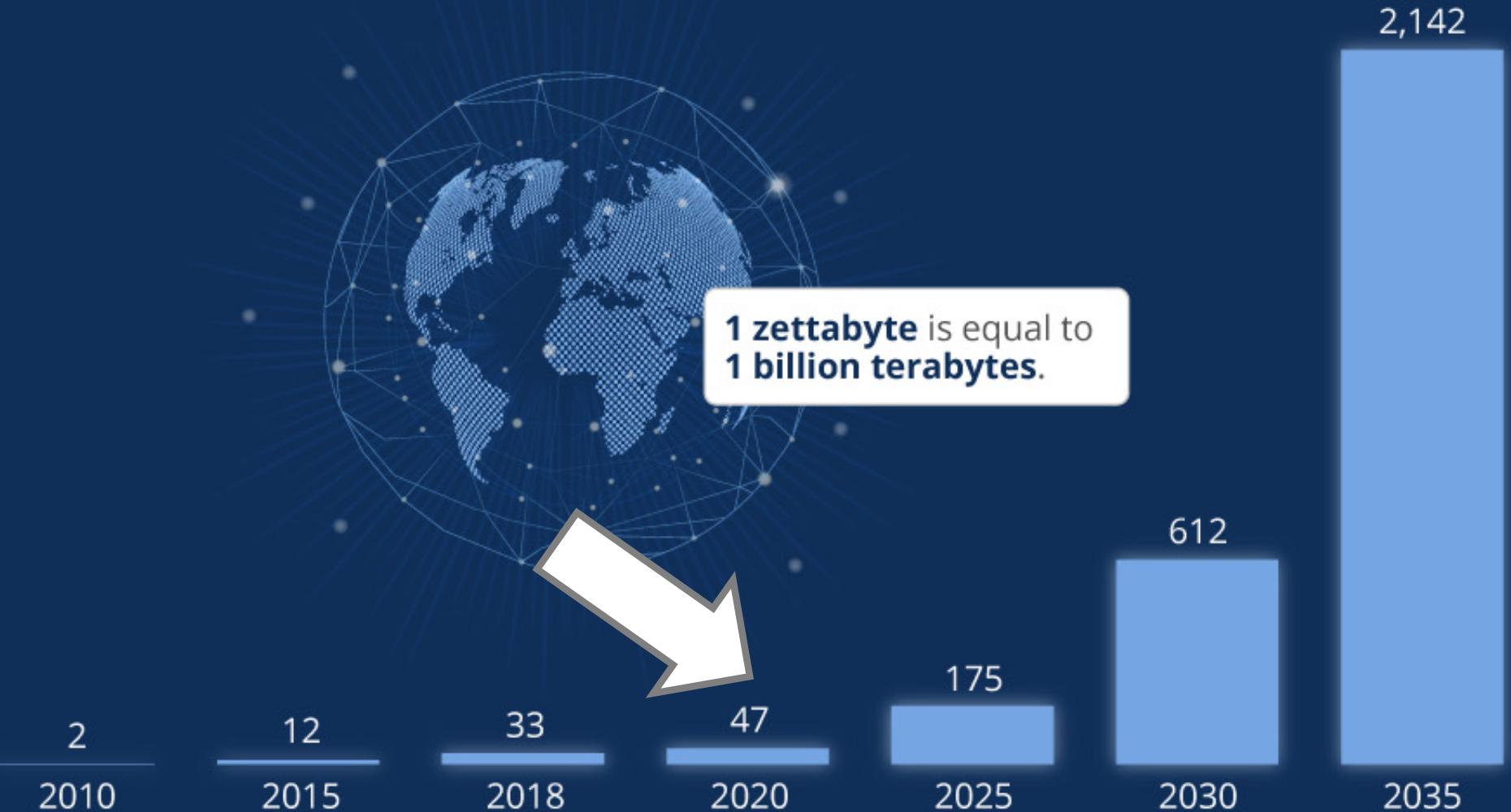
Data

is at the center of *everything*



Global Data Creation is About to Explode

Actual and forecast amount of data created worldwide 2010-2035 (in zettabytes)



@StatistaCharts

Source: Statista Digital Economy Compass 2019

statista

How did we get here?

Data was *Manual*

67

1928

June 11.	Geo. A. Kelly
June 16	Mrs. Chas. Long Jr
June 16	Nellora Wright
June 16	Charity A. Bowes
" "	Mrs. M. A. Carpenter
" "	Mr. & Mrs. Carpenter
July 10	James Stevens trap I
July 10	A. W. Gennings
July 10	Millicent Gennings
	Walt Kulin
July 11.	Mrs. Rawe, & Daughter
	Mrs. Ralph Rogers
" "	Mrs. A. H. Favout
" "	Mrs. J. A. Miller
	Mrs. G. J. Hayes
	Mrs. C. J. Mata
	Mary B. Cosgrave
	Mrs. & Maiden Hoffman
	Mrs. Key D. Young
	Mrs. A. S. Whitney
	Mrs. F. L. St. Gerda
	Mrs. E. H. Patman

Data was *Expensive*



Data is Cheap

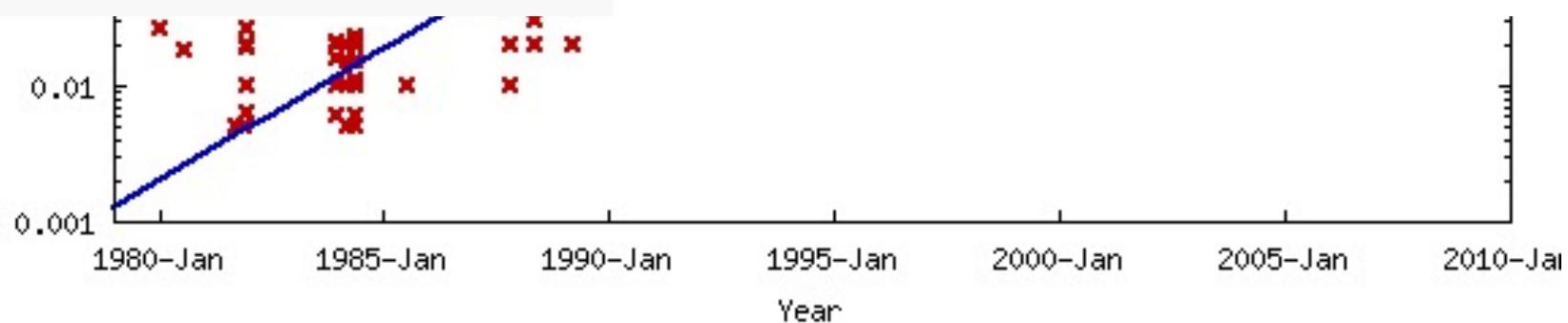


Sponsored ⓘ

2TB USB 3.0 Flash Drive - Read Speeds up to 100MB/s | 2000GB Pen Drive 2TB Swivel Metal Style

\$37⁹⁹

✓prime FREE Delivery Sat, Jan 22



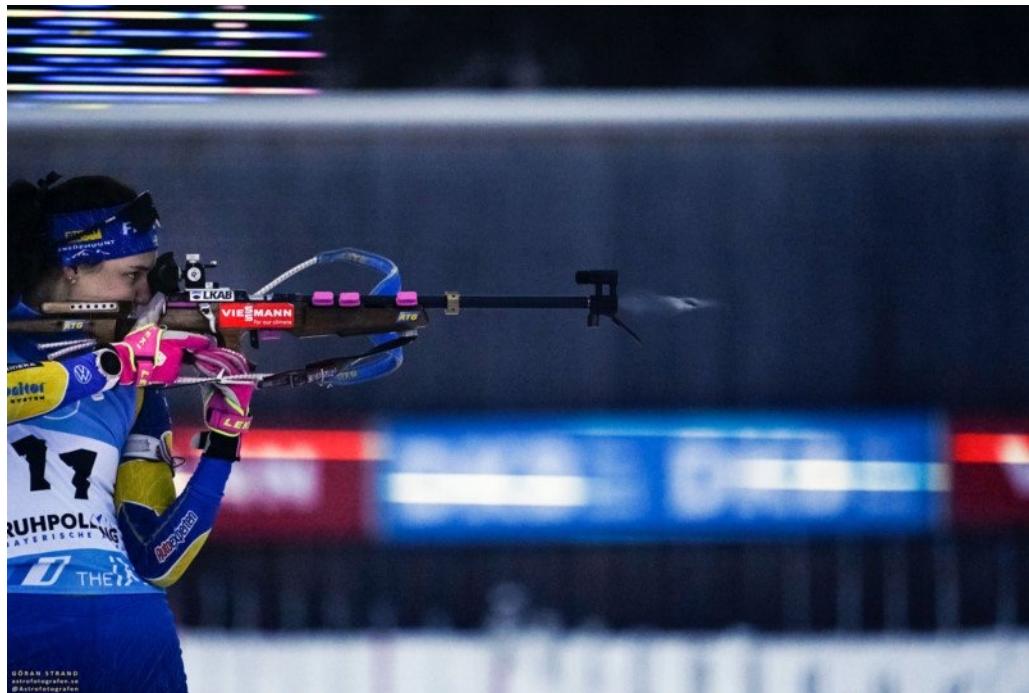
Data is Automated

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

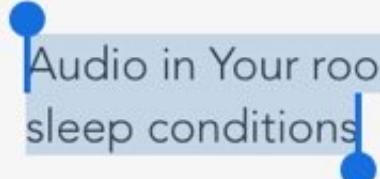
Phones

GPS/Cars



- 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 Services

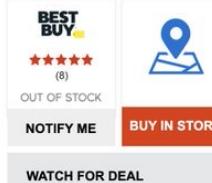


Want playtime ideas & special offers? [SIGN UP](#)[Home](#) / [Toddler & Preschool Toys](#) / Chatter Telephone™ With Bluetooth®

Chatter Telephone™ with Bluetooth®

★★★★★ (0)

Introducing the special edition Fisher-Price® Chatter Telephone™ — a phone smart enough not to come with any apps. Its intuitive bulky face design comes with a 'super-advanced' rotary dial and connects to your mobile device via Bluetooth® wireless technology, so you can... [Read More](#)



Images 1/7

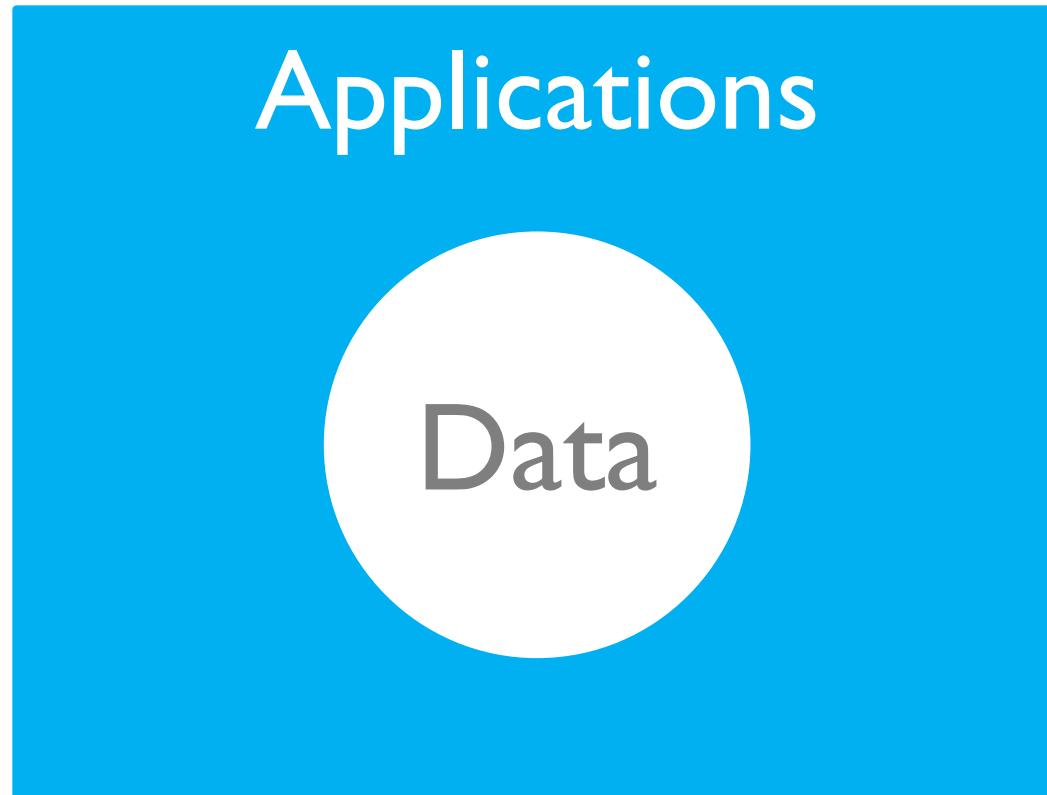
Double click to zoom



**Connect to your smart phone
to make & receive calls**



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



The image shows the homepage of the ProPublica Surgeon Scorecard. At the top left is the ProPublica logo, which consists of the word "PRO" in a white circle and "PUBLICA" in a black circle. To the right of the logo is the text "Patient Safety". Below the logo is the title "Surgeon Scorecard" in large, bold, white serif font. Underneath the title is the text "by Sisi Wei, Olga Pierce and Marshall Allen, ProPublica, Updated July 15, 2015" in a smaller white serif font. At the bottom of the page is a paragraph of white text on a black background, explaining the purpose of the scorecard.

Surgeon Scorecard

by Sisi Wei, Olga Pierce and Marshall Allen, ProPublica, Updated July 15, 2015

Guided by experts, ProPublica calculated death and complication rates for surgeons performing one of eight elective procedures in Medicare, carefully adjusting for differences in patient health, age and hospital quality. Use this database to know more about a surgeon before your operation.

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



Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

8 MIN READ

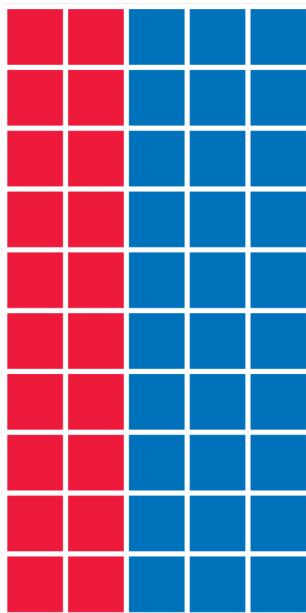


CS 4

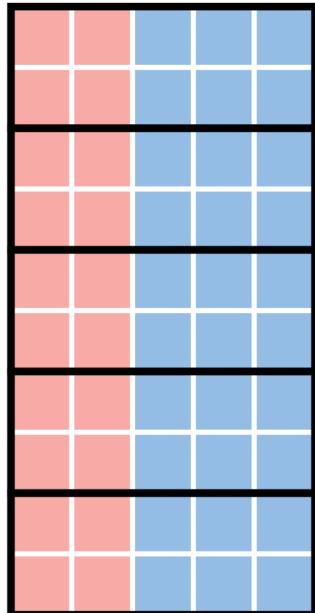
SAN FRANCISCO (Reuters) - Amazon.com Inc's [AMZN.O](#) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

What are we doing with data?

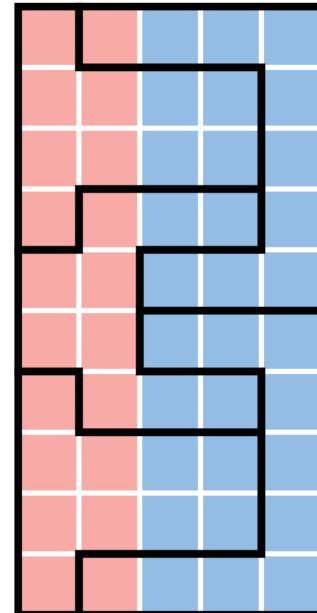
HOW TO STEAL AN ELECTION



50 PRECINCTS
60% BLUE
40% RED



5 DISTRICTS
5 BLUE
0 RED
BLUE WINS



5 DISTRICTS
3 RED
2 BLUE
RED WINS

Congressional District 4



What are we doing with data?

Health

Investigative Journalism

Recommendations

Politics

Make a
contribution

Subscribe

Find a job

Sign in / Register

Search ▾

News

Opinion

Sport

Culture

Lifestyle

More ▾

US edition ▾

The Guardian



The Cambridge Analytica Files

A year-long investigation into
Facebook, data, and influencing
elections in the digital age

Key stories

Hide



TIME

Subscribe

2012 ELECTION

Inside the Secret World of the Data Crunchers Who Helped Obama Win

Data-driven decisionmaking played a huge role in creating a second term for the 44th President and will be one of the more closely studied elements of the 2012 cycle

What are we doing with data?

Health

Investigative Journalism



set station news arts & life music programs

shop

parallels MANY STORIES, ONE WORLD



4:21

+ QUEUE

DOWNLOAD

EMBED

Facial Recognition In China Is Big Business As Local Governments Boost Surveillance

April 3, 2018 · 10:40 AM ET



ROB SCHMITZ



Every day, the NSA intercepts and stores **1.7 billion** emails, phone calls, texts and other electronic

What are we doing with data?

Health

Investigative Journalism

Recommendations

Politics

Surveillance

Identity

**“YOU ARE BASICALLY DENIED ALMOST EVERYTHING
IF YOU CAN’T PROVE WHO YOU ARE.”**



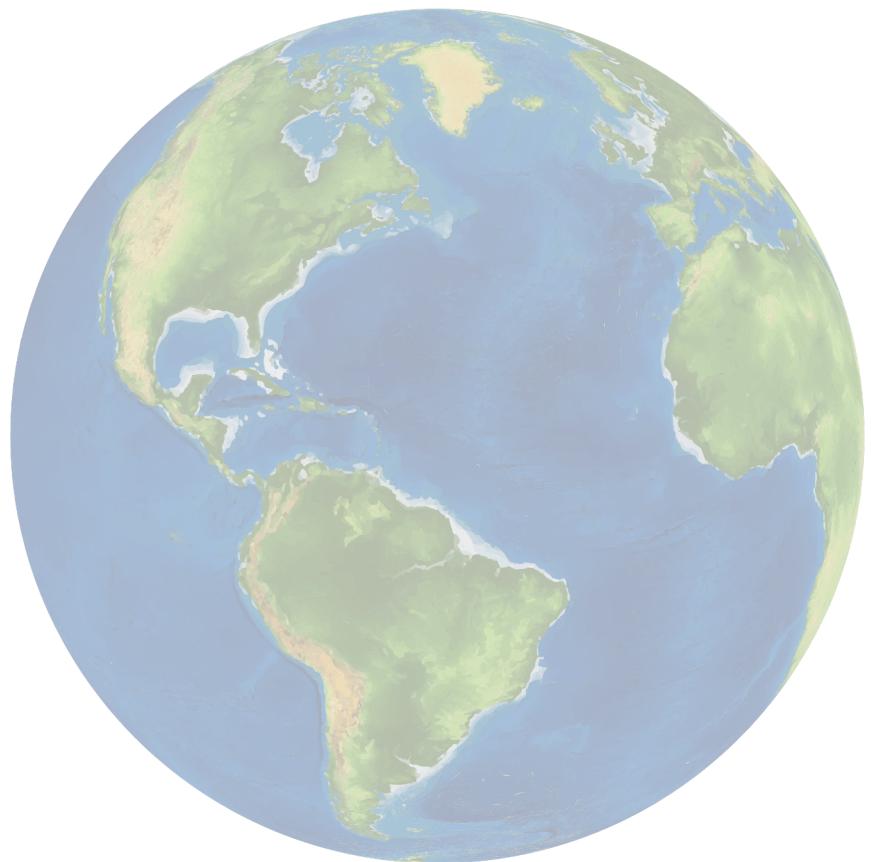
30 APR 2012 RESEARCH & IDEAS

India’s Ambitious National Identification Program

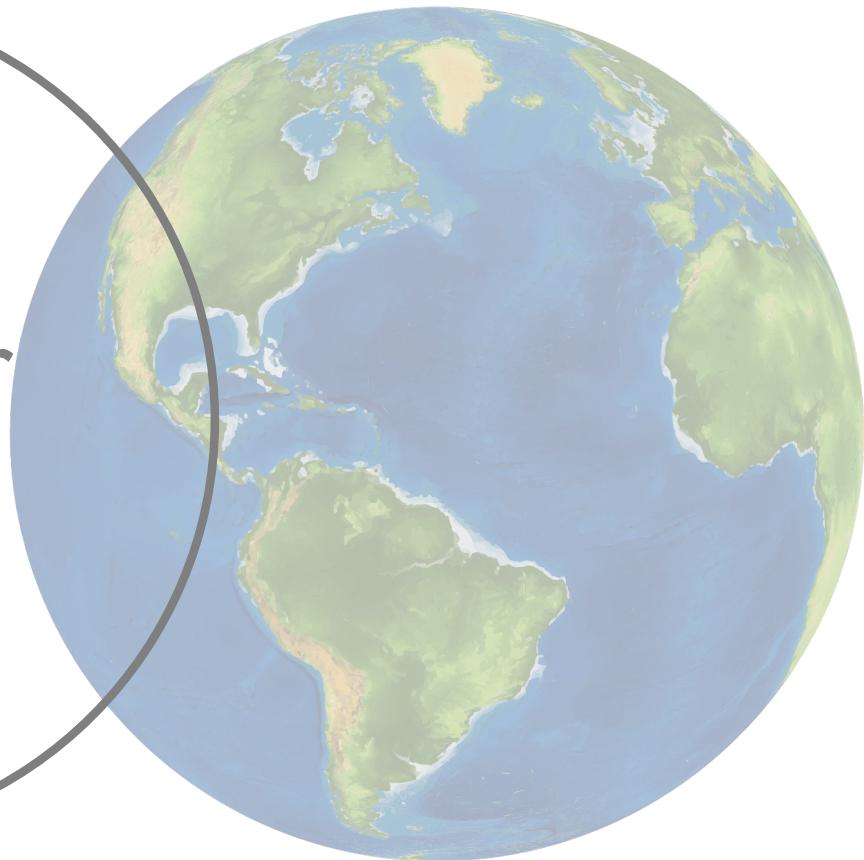
[Comments 30](#) [Email](#) [Print](#) [Download](#) [Share](#) [f Recommend](#) [Share 92](#)

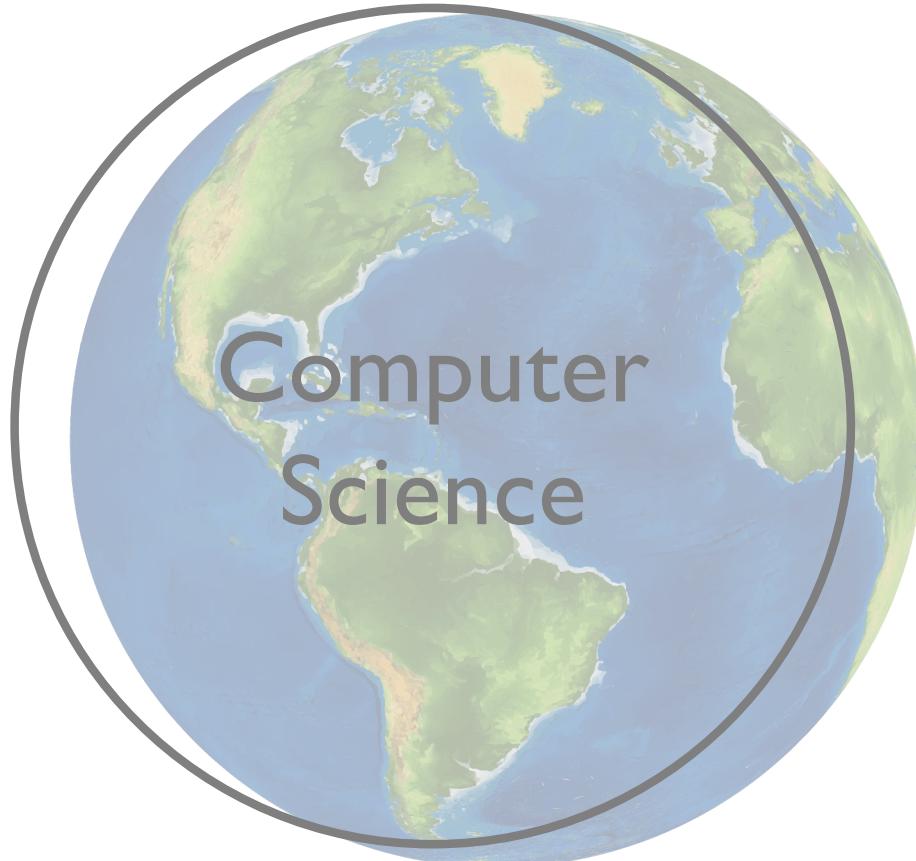
The Unique Identification Authority of India has been charged with implementing a nationwide program to register and assign a unique 12-digit ID to every Indian resident—some 1.2 billion people—by 2020. In a new case, Professor Tarun Khanna and HBS India Research Center Executive Director Anjali Raina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

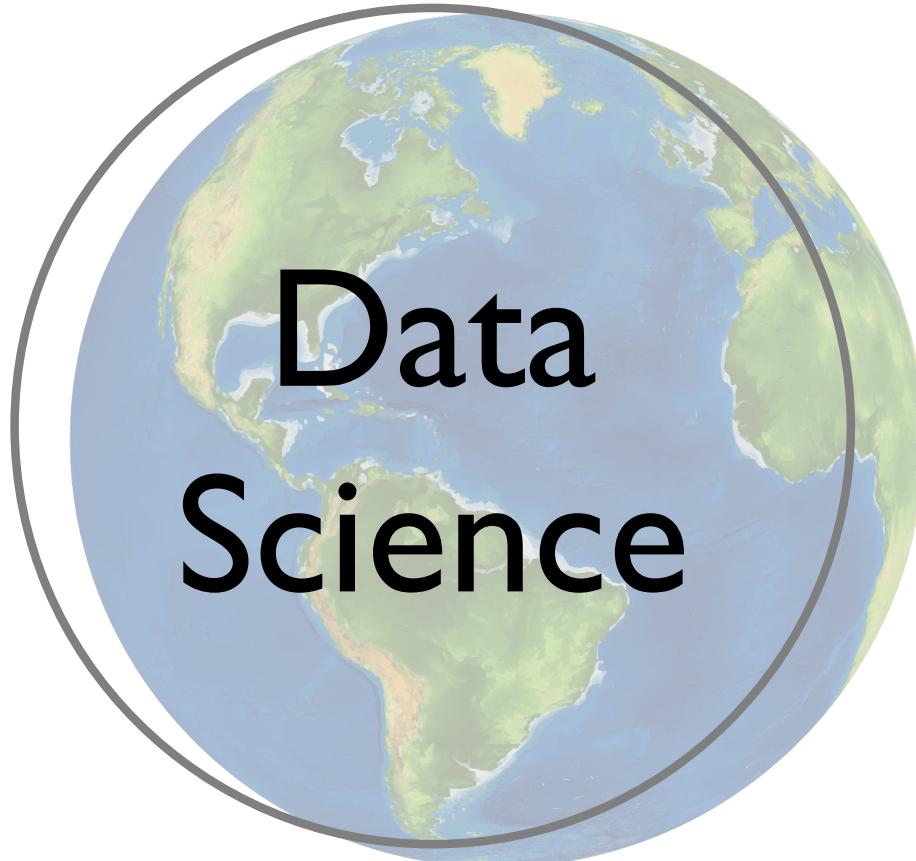
Computer
Science



Computer
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

Lone data scientist uses a static, clean table,

applies statistics or fits an ML model

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

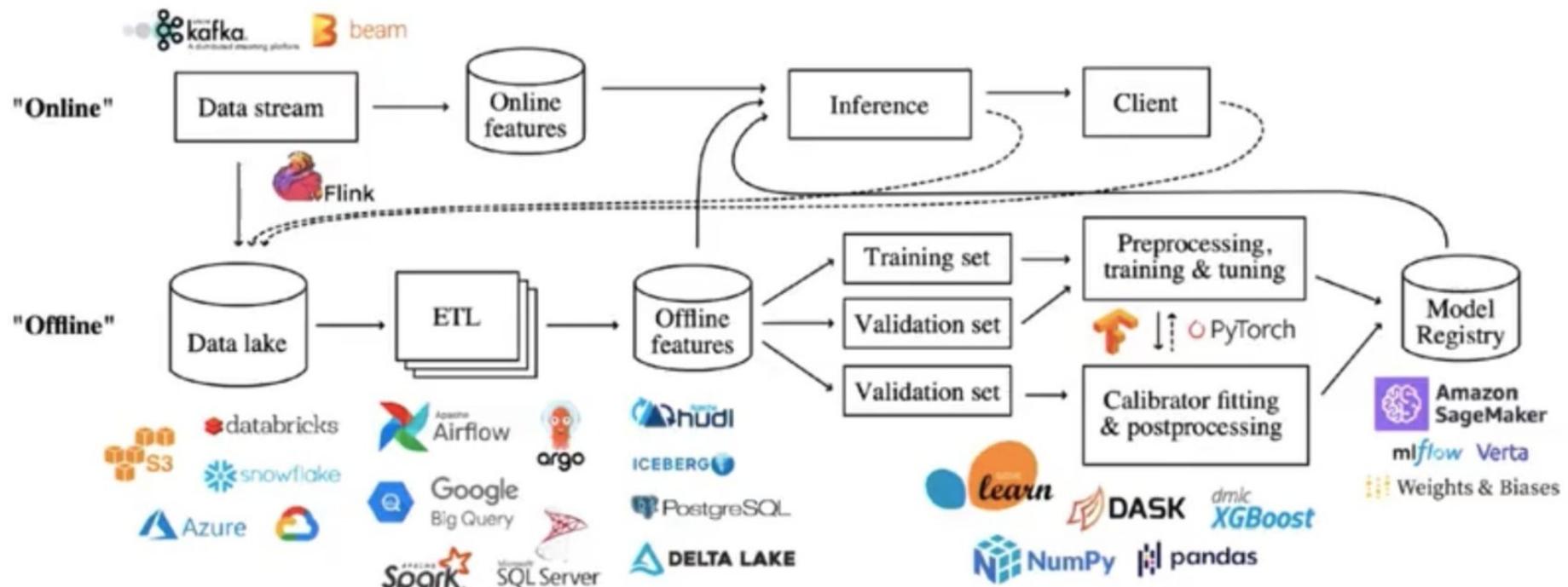


Figure 1: High-level architecture of a generic end-to-end machine learning pipeline. Logos represent a sample of tools used to construct components of the pipeline, illustrating heterogeneity in the tool stack. *Shankar et al. 2021*

<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



Big Data Borat

@BigDataBorat

 Follow

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.

Most time spent on data engineering

Often not viewed as sexy, but critical

Data Eng Dominates Data Science Projects

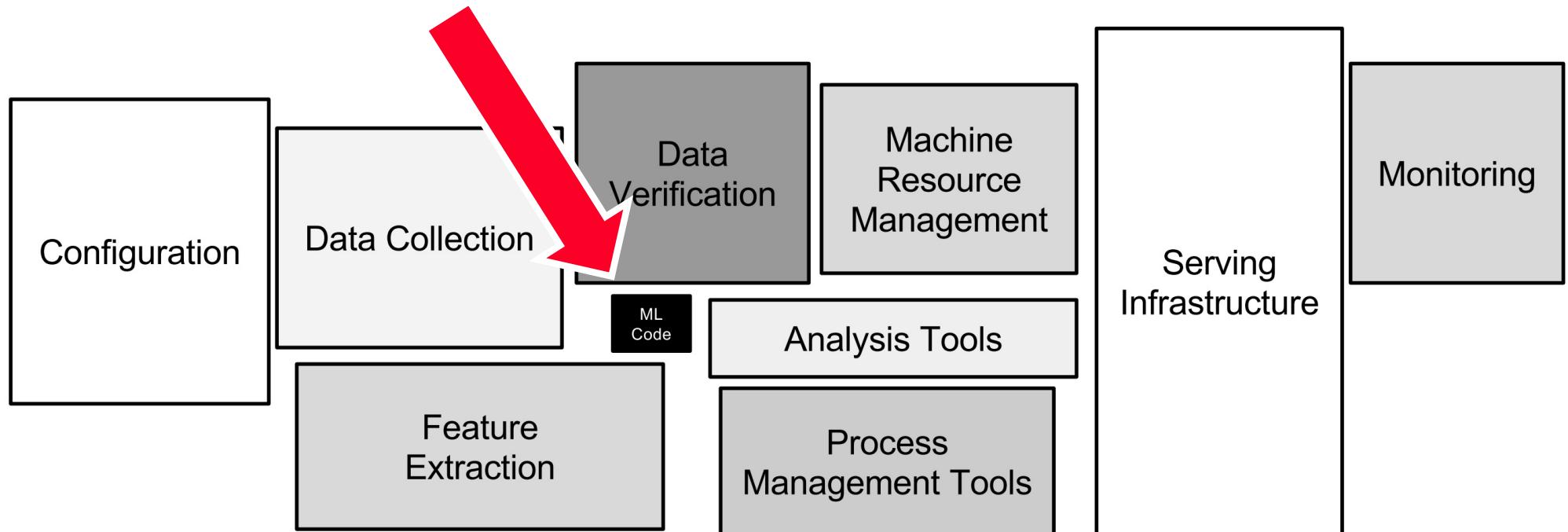
Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley, gholt, dg, 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



Andrej Karpathy

@karpathy

...

But as of approx. last two years, even the neural net architectures across all areas are starting to look identical - a Transformer (definable in ~200 lines of PyTorch [github.com/karpathy/minGPT...](https://github.com/karpathy/minGPT)), with very minor differences. Either as a strong baseline or (often) state of the art.

7:03 PM · Dec 7, 2021 · Twitter Web App

THE DATA SCIENCE HIERARCHY OF NEEDS

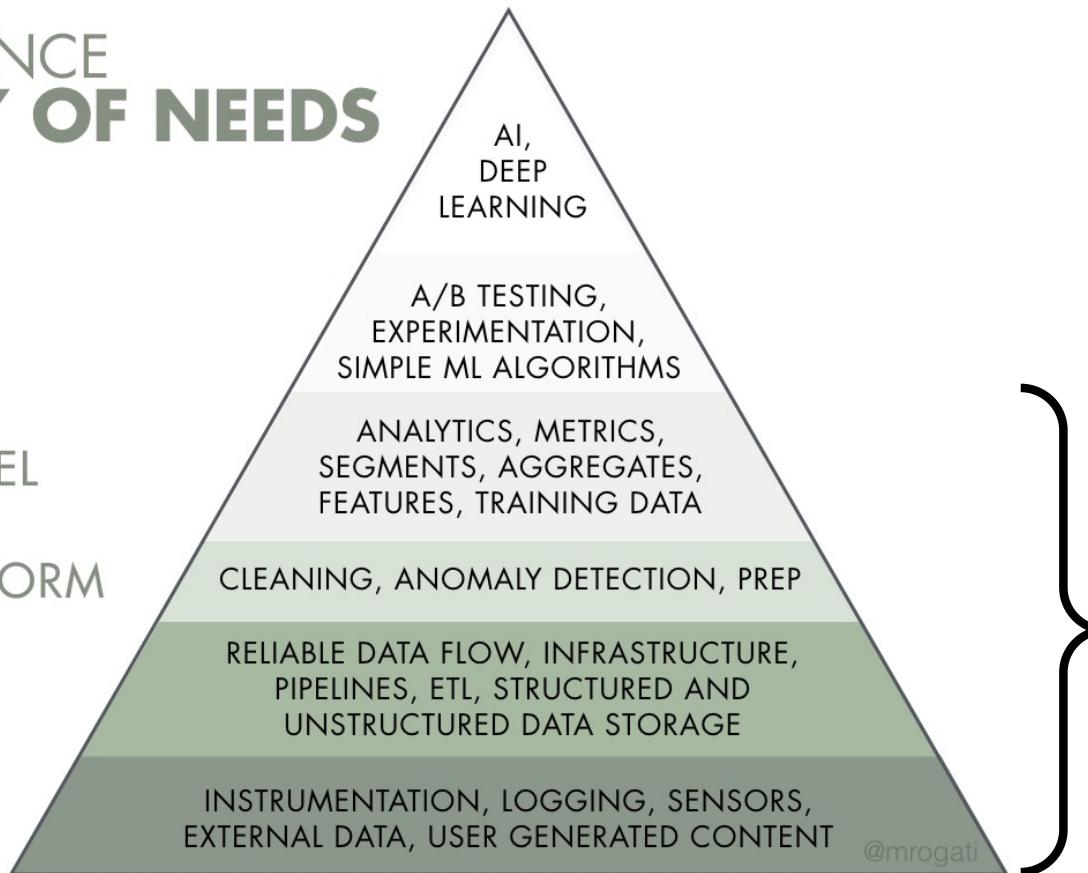
LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

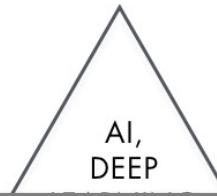
MOVE/STORE

COLLECT



Data eng. work
that must
happen first

THE DATA SCIENCE **HIERARCHY OF NEEDS**



“However, under the strong influence of the current AI hype, people try to plug in data that’s dirty & full of gaps, that spans years while changing in format and meaning, that’s not understood yet, that’s structured in ways that don’t make sense, and expect those tools to magically handle it.”

COLLECT

INSTRUMENTATION, LOGGING, SENSORS,
EXTERNAL DATA, USER GENERATED CONTENT

@mrogati

“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,loraal]@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

Feb 4, 2019, 08:15am EST | 171,732 views

Why There Will Be No Data Science Job Titles By 2029



Noah Gift
Forbes Council
Forbes Technology Council
Innovation

We Don't Need Data Scientists,
We Need Data Engineers

January 2021

“..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

PDF

Python Scripts, Spark/Databricks, etc

Data Preparation

Transform/clean data into useful form

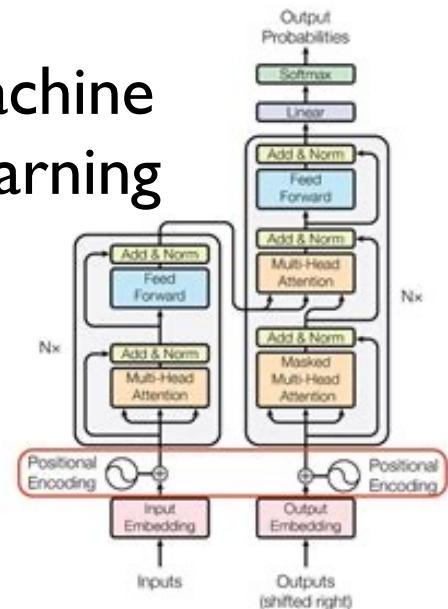


```
client.createAnnotationTask({
  callback_url: 'http://www.example.com/callback',
  instruction: 'Draw a box around each rooftop and pool.',
  attachment: 'http://i.imgur.com/X0JbalC.jpg',
  objects_to_annotate: ['pool', 'rooftop'],
  with_labels: true,
  min_width: 30,
  min_height: 30
}, (err, task) => {
  // do something with task
}).
```



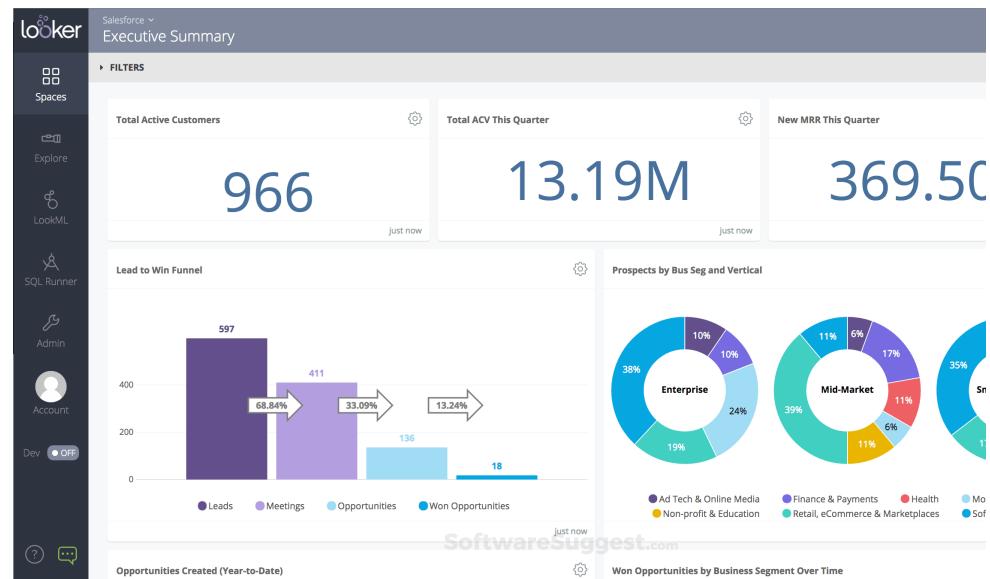
JSON

Machine Learning



Logs

Visualization



Sources



+ Add

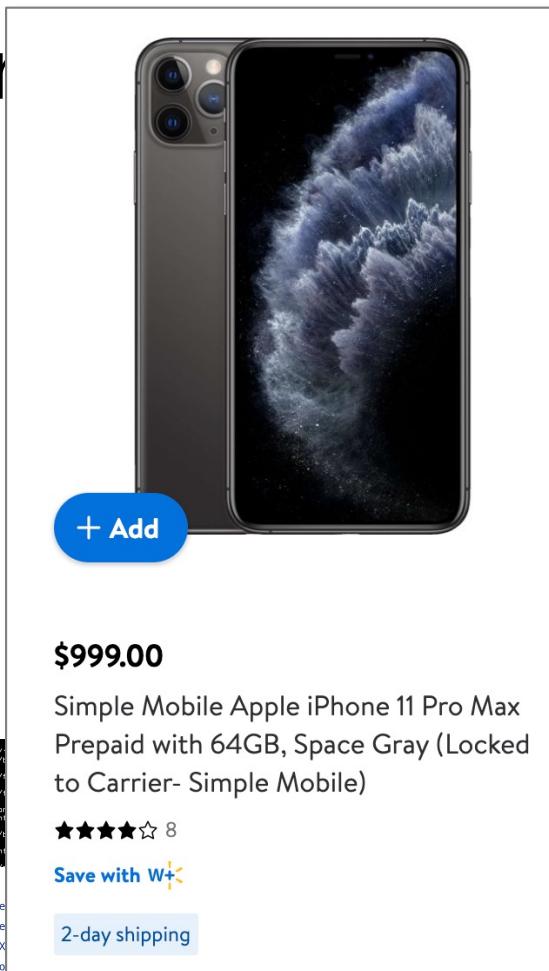
\$999.00

Simple Mobile Apple iPhone 11 Pro Max
Prepaid with 64GB, Space Gray (Locked
to Carrier- Simple Mobile)

8

Save with W+

2-day shipping





+ Add

\$999.00

Total Wireless Apple iPhone 11 Pro Max,
64GB, Space Gray- Prepaid Smartphone

★★★★★ 11

Save with W+ 

2-day shipping

Use Cases

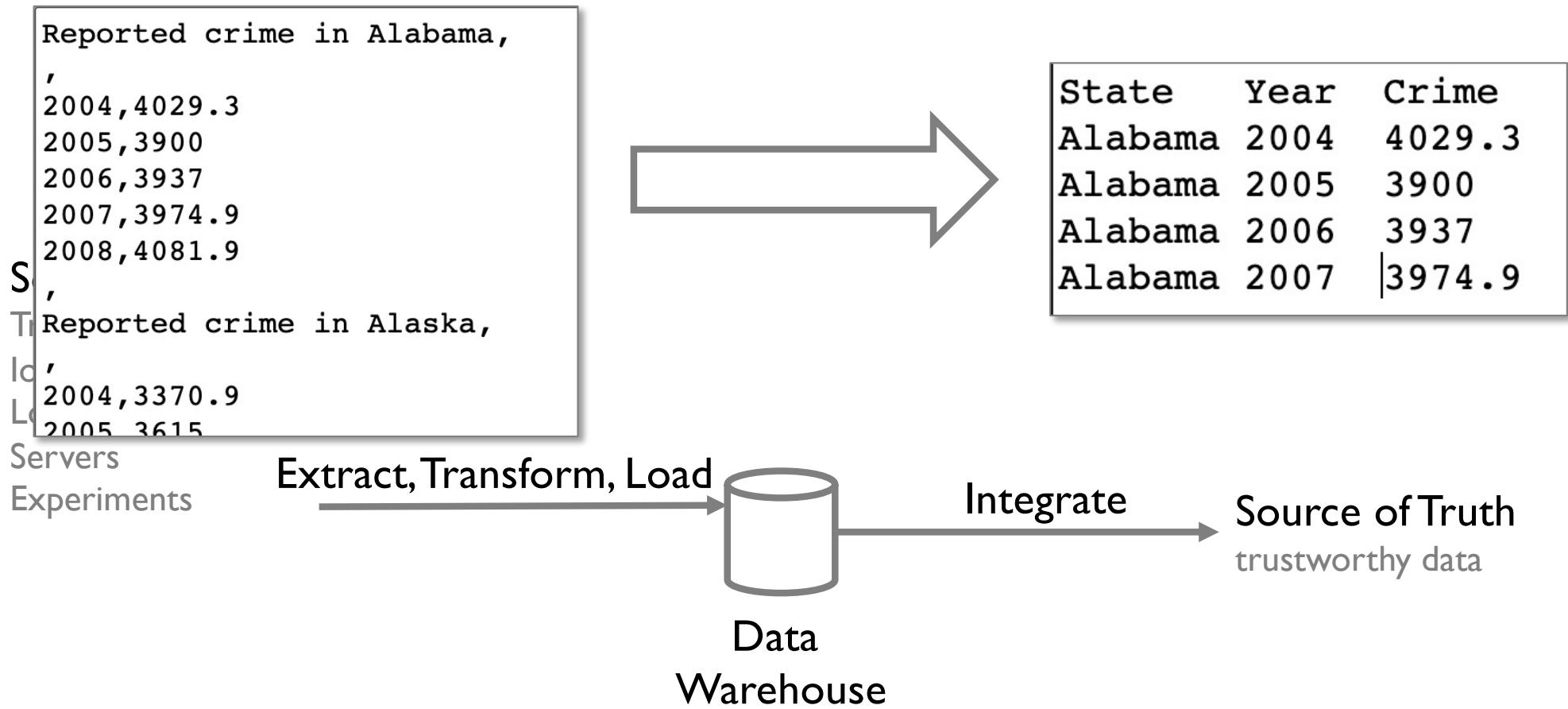
AI, ML, Data science, Apps, Webservices

Source of Truth

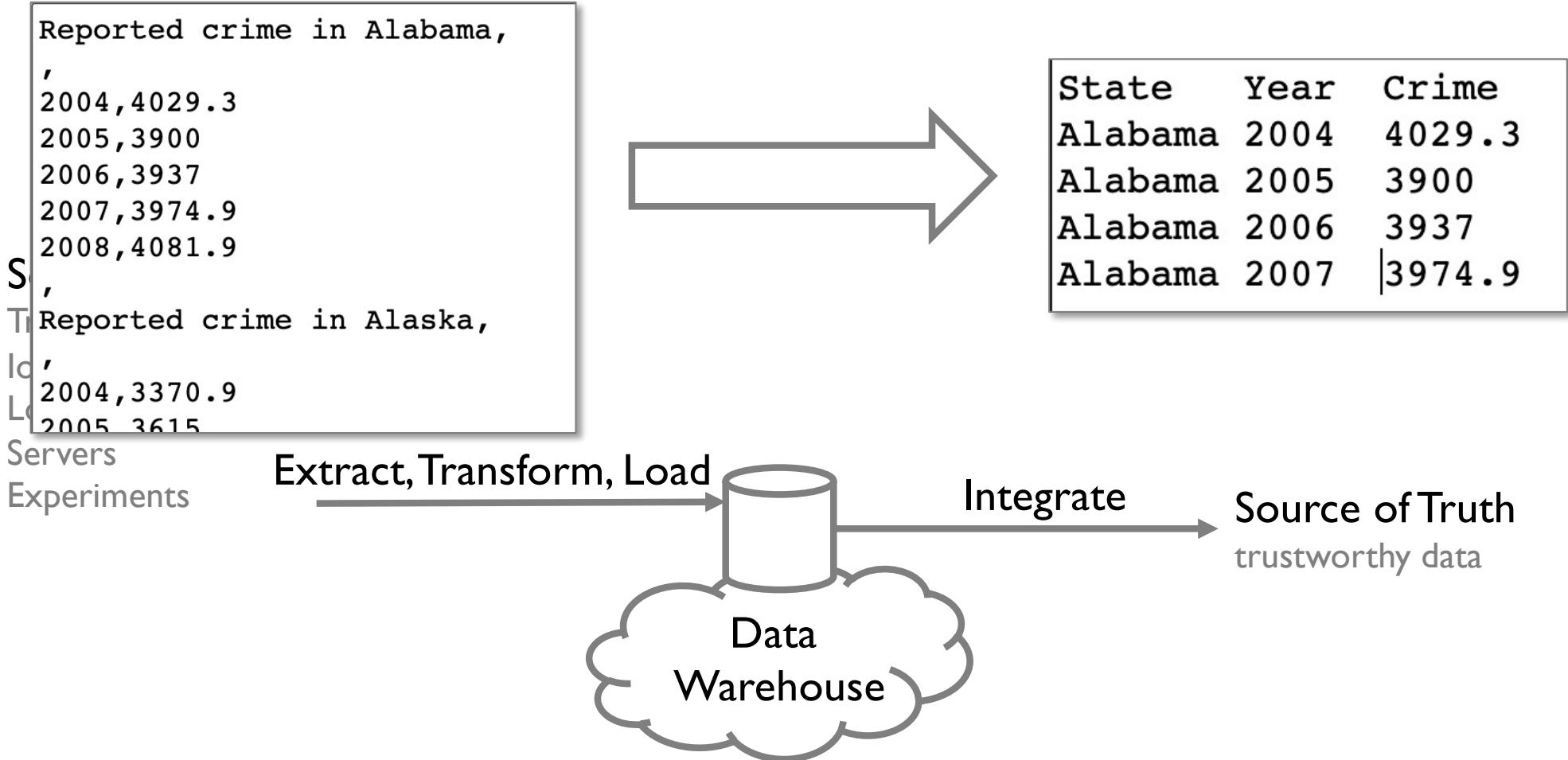
trustworthy data

	#	2004	#	2005
4829.3		3988		
3378.9		3615		
2 Arizona	5873.3	4827		
3 Arkansas	4833.1	4068		
4 California	3423.9	3321		
5 Colorado	3918.5	4041		
6 Connecticut	2684.9	2579		
7 Delaware	3280.6	3118		
8 District of Columbia	4852.8	4490		
9 Florida	4182.5	4013		

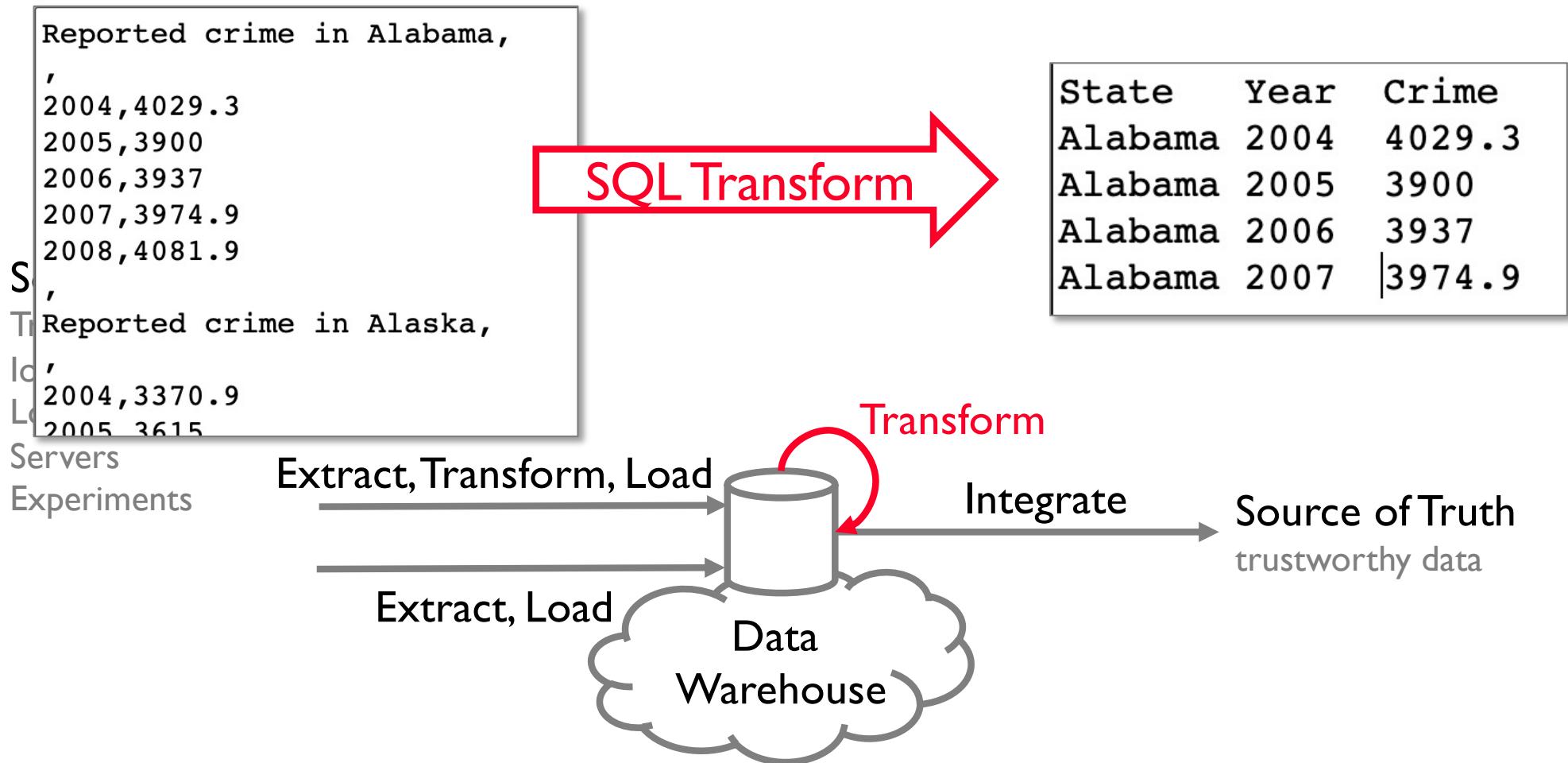
ETL: Extract Transform Load (1980s)



ELT: Extract Load Transform (2010s)

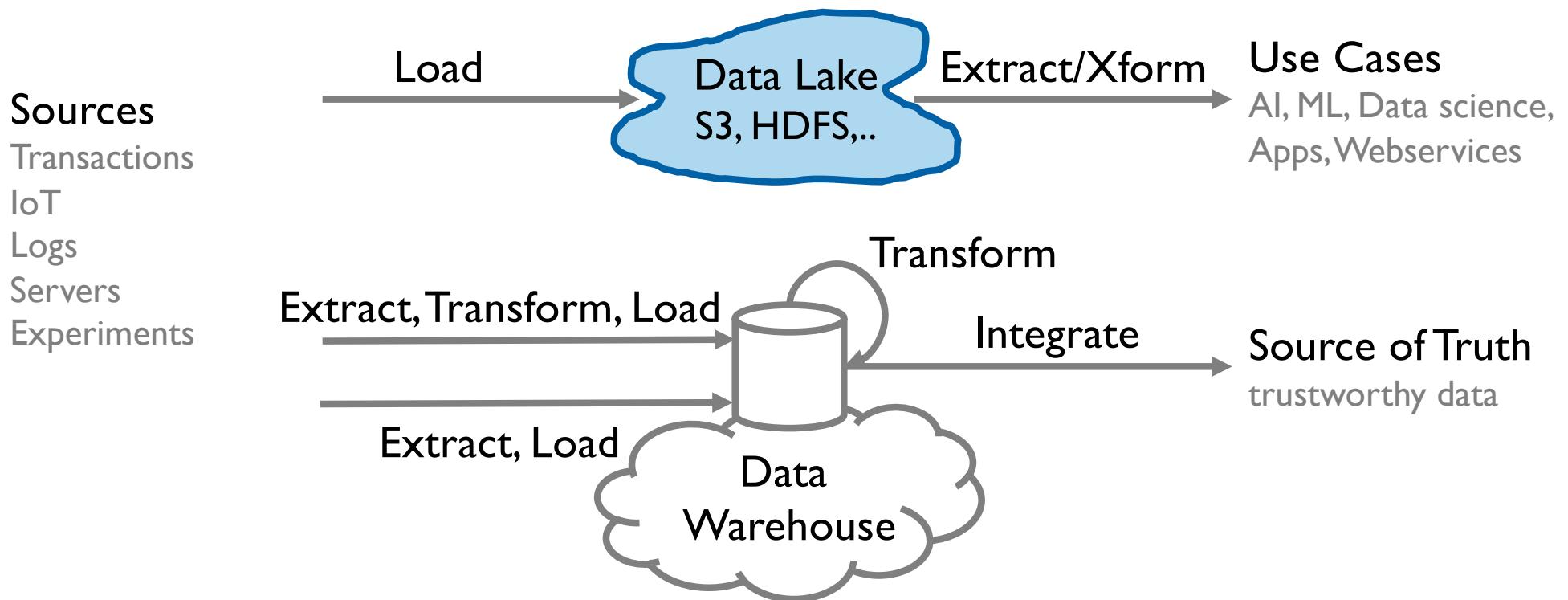


ELT: Extract Load Transform (2010s)

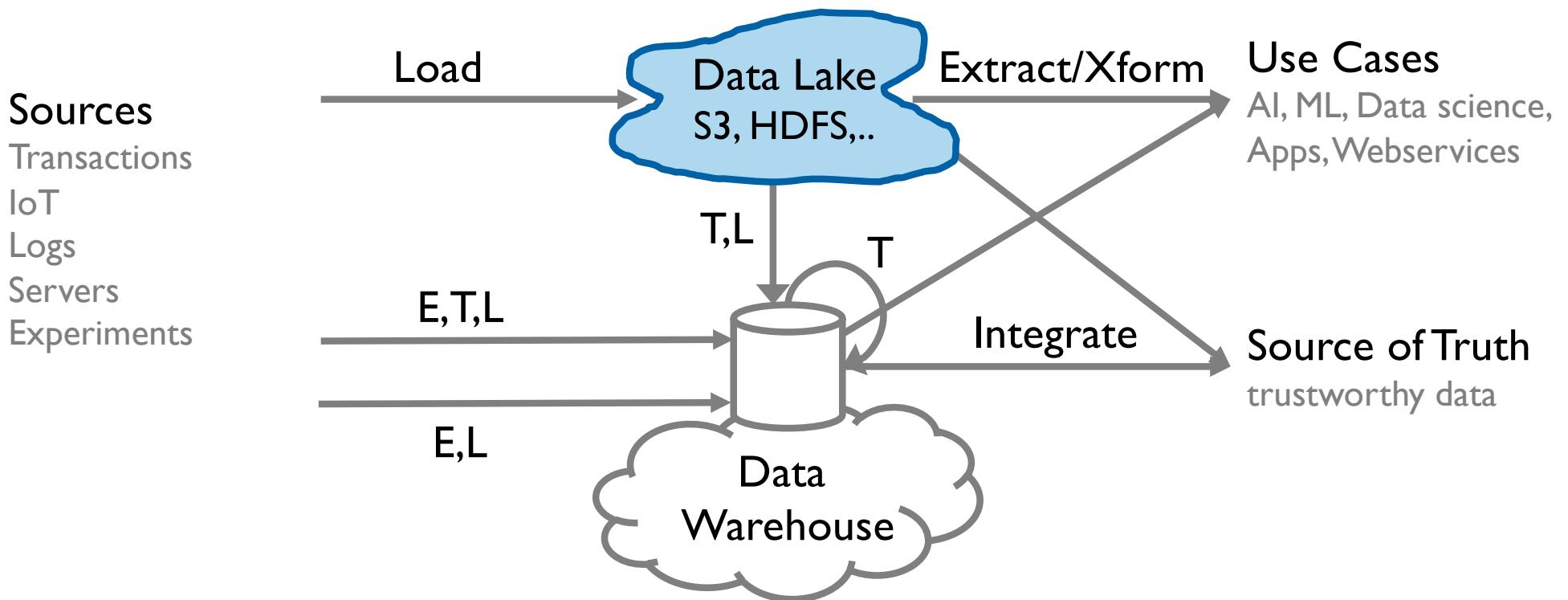


Data Lakes (2000s)

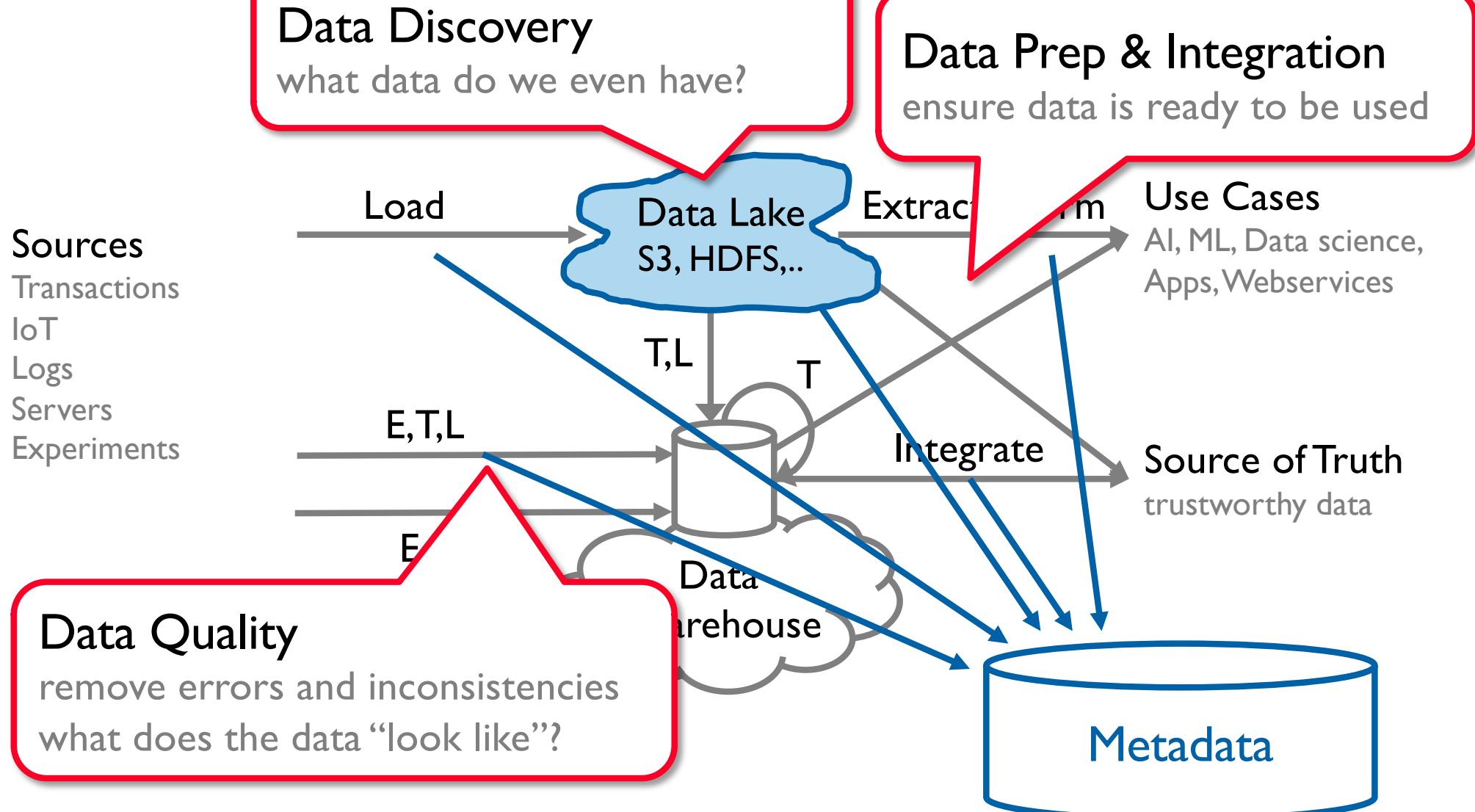
Store as files, transform when needed

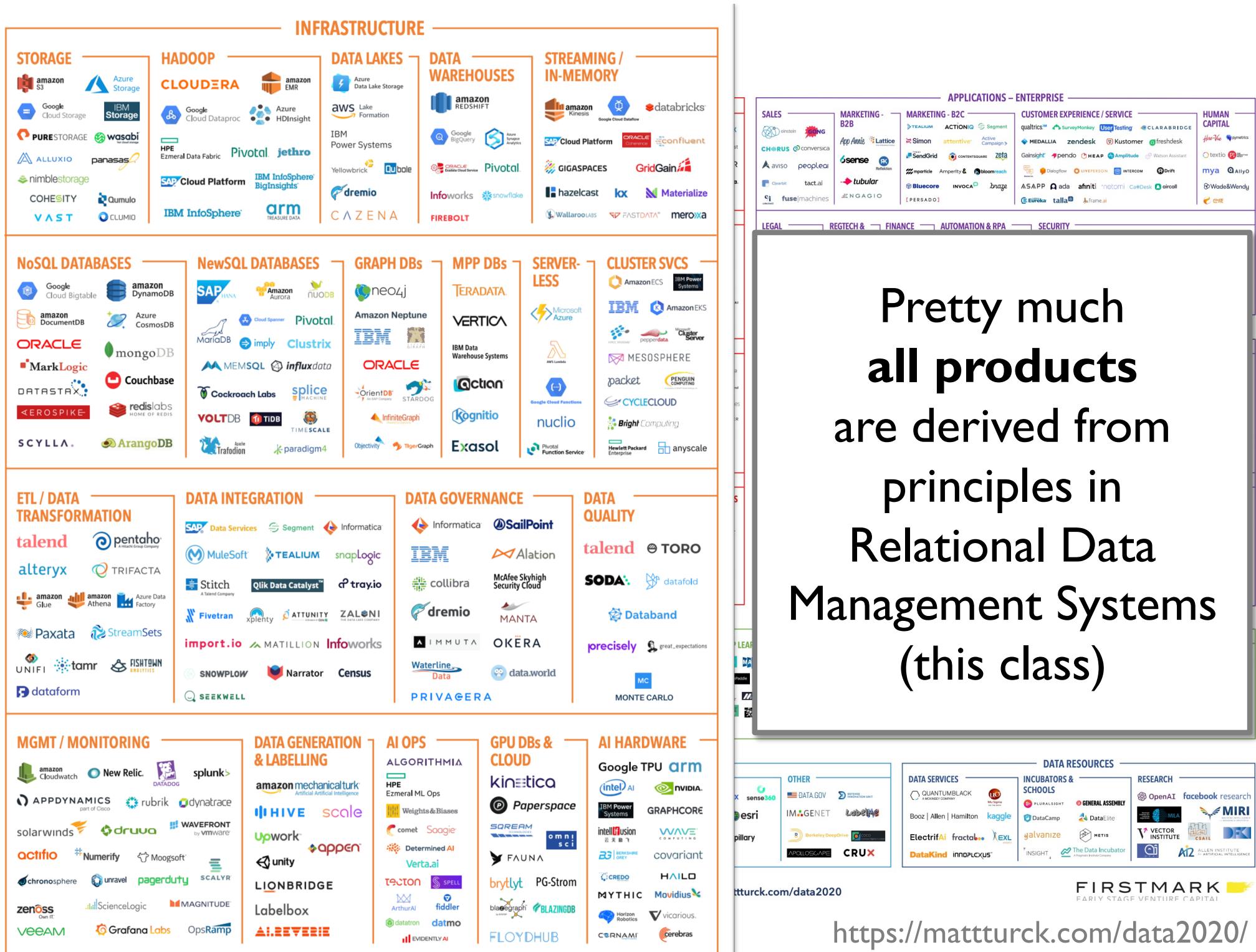


Everything All At Once (2010s)



Understanding the Data





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://github.com/w4111/hw0	Ch 1, 2		< a href="https://github.com/w4111/hw0" > HW 0 		
3	#e7f8ff	28-Jan	ER Models				Ch 2		< a href="https://github.com/w4111/hw1-s22" > HW 1 	0	HW 0
4	#e7f8ff	4-Feb	Data Models				Ch 3	optional: HW 1 		0	HW1 Part1
5	#e7f8ff	11-Feb	Data Models + ER->Relational				Ch 3	optional: HW 1 		0	Project 1 Part 1 approval phase
6	#f2f9ed	18-Feb	Relational Algebra				Ch 4		< a href="https://github.com/w4111/project1" > Project 1 	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/blob/main/midterm.pdf" > Midterm 	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" > HW 4 	0	HW3
13	#ddf9ff	8-Apr	Physical Design				Ch 8		< a href="https://github.com/w4111/project2_s22" > Project 2 	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 (Cumulative)	one 8x11 page cheat sheet both sides						0	Project 2
18											
19											
20											

What is a Database?



••••• AT&T ⌂ 3:00 PM 1 3G

Contacts +

Search

A

Apple Inc.

C

Call Recorder

F

Julia Fillory

Mike Fillory me

G

Justin Gilmore

Thomas Gilmore

Willa Good

H

Barry T. Hubbard

M

Favorites Recents Contacts Keypad Voicemail

A screenshot of a smartphone contacts application. The screen shows a list of contacts sorted by initial. At the top right is a red '+' button. Below it is a search bar with the placeholder 'Search'. The contact list includes: Apple Inc., Call Recorder, Julia Fillory, Mike Fillory (with a 'me' tag), Justin Gilmore, Thomas Gilmore, Willa Good, Barry T. Hubbard, and several entries starting with 'H' and 'M'. At the bottom are navigation icons for Favorites, Recents, Contacts (which is blue and bolded), Keypad, and Voicemail.

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,  
cliID: 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. $\neg \backslash (\exists) \neg$

Is the file system a DBMS?

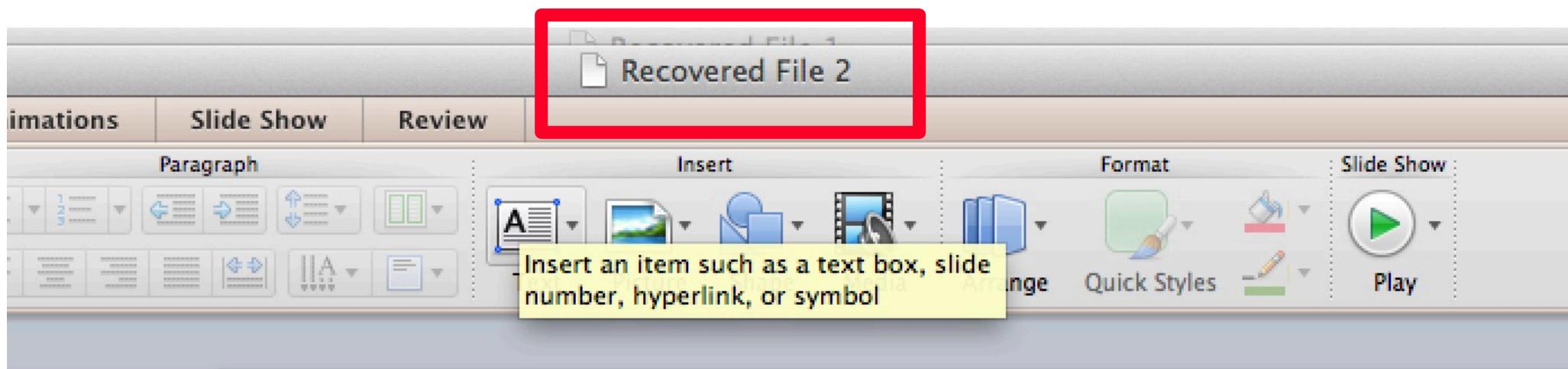
You edit a text file

Computer crashes

What happens?

1. All changes survive
2. No changes survive
3. Changes from last save survive
4. $\neg \backslash (\exists) \backslash$

Is the file system a DBMS?



COMS W4111
Introduction to Databases

Who... would ever do this?

Real \$IB+ Companies...

Store extracted data in a file

Every change → rewrite the file

EUGENE WU

BIO

Eugene Wu is broadly interested in technologies that help users play well at technical levels to effectively and quickly make sense of their information that ultimately improve the user interface between people and data, and uses data from various systems and sources, such as the Web, sensors, and from MIT, B.S. from Cal, and was a postdoc in the AMPLab. A profile, an Eugene Wu has received the VLDB 2018 10-year test of time award, best VLDB, the SIGMOD 2016 best demo award, the NSF CAREER, and the G The WuLab Website & Blog

We are recruiting PhDs + Postdocs, and Interns + UGrad + Mast

Overview of My Research and Teaching

SELECTED PUBLICATIONS (SHOW ALL)

Private Federated Exploration of Inference Queries
Young Wu, Yiqia Lu, Lampros Flekas, Jianan Wang, Eugene Wu
VLDB 2022

Explaining SQL-M: Queries with Bayesian Optimization
Brandon Lockhart, Jianan Wang, Eugene Wu

From Debugging Before ML to Cleaning For ML
Felix Nezura, Binger Chen, Ziaessah Abdine, Eugene Wu
Invited, IEEE Data Engineering Bulletin 2021

Continuous Preference for Interactive Data Applications
Haneen Mohammed, Ziyun Wei, Ravi Netravali, Eugene Wu
VLDB 2020 Talk Video Bioghost

Complaint-driven Training Data Debugging for Query 2.0
Young Wu, Lampros Flekas, Jianan Wang, Eugene Wu
SIGMOD 2020 Talk Video Bioghost

Monte Carlo Tree Search for Generating Interactive Data Analysis Interfaces
Yiqia Chen, Eugene Wu

NEWS

Jun-2022: Looking forward to giving one of the keynotes at SEA DATA at VLDB 2022 this summer!

NotEnough is a blog post to introduce our Khamenei paper. Hansen also recorded a short YouTube video about it.

Jul-2020: FLAME EBS VCT-DIST Khamenei, our rethink of client-server communication for interactive applications will be presented at VLDB 2020! With Haneen Mohammed, Tracy Wei, and Ravi Netravali. This is a very mortal combinatorial system and has failed to submit.

Jun-2020 - Haneen participated in, and won, first place at the 2020 SIGMOD student research competition for her work on Khamenei!

Mar-2020: FATALITY! A new mortal kombat-themed system has been beaten into submission. Our full paper about it is now available on the TrainDev site.

Browser

New Tab

EUGENE WU



BIO

Eugene Wu is broadly interested in technologies that help users play with their data. His goal is for all technical levels to effectively and quickly make sense of their information. He is interested in solutions that ultimately improve the interface between users and data, and uses techniques borrowed from fields such as data management, systems, crowd sourcing, visualization, and HCI. Eugene Wu received his B.S. from MIT, and was a postdoc in the AMPLab. A profile, an obit.

Eugene Wu has received the VLDB 2018 10-year test of time award, best-of-conference citations at VLDB, the SIGMOD 2016 best demo award, the NSF CAREER, and the Google and Amazon faculty awards.

The WuLab Website & Blog
We are recruiting PhDs + Postdocs, and Interns + UGrad + Masters!

[Overview of My Research and Teaching](#)

NEWS

Jun-2021: Looking forward to giving one of the keynotes at SEA DATA at VLDB 2021 this summer!

Aug-2020: For Highly Interactive Apps, Prediction is Not Enough is a blog post to introduce our Kameleon paper. Haneen also recorded a short YouTube video summarizing our work.

Jul-2020: FLAWLESS VICTORY! Kameleon, our

SELECTED PUBLICATIONS (SHOW ALL)

Private Federated Explanation of Inference Queries
Young Wu, Yiqia Lu, Lampros Flekas, Jianan Wang, Eugene Wu
VLDB 2022

Explaining SQL-ML Queries with Bayesian Optimization
Brandon Lockhard, Jianan Wang, Eugene Wu

From Debugging Before ML to Cleaning For ML
Felix Nezura, Binger Chen, Ziaessab Abdess, Eugene Wu
Invited, IEEE Data Engineering Bulletin 2021

Continuous Preference for Interactive Data Applications
Haneen Mohammad, Ziyun Wei, Rav Nettivalli, Eugene Wu
VLDB 2020 Talk Video Biogpost

Complaint-driven Training Data Debugging for Query 2.0
Young Wu, Lampros Flekas, Jianan Wang, Eugene Wu
SIGMOD 2020 Talk Video Biogpost

Monte Carlo Tree Search for Generating Interactive Data Analysis Interfaces
Yiqia Chen, Eugene Wu

EUGENE WU



BIO

Eugene Wu is broadly interested in technologies that help users play with their data. His goal is for all technical levels to effectively and quickly make sense of their information. He is interested in solutions that ultimately improve the interface between users and data, and uses techniques borrowed from fields such as data management, systems, crowd sourcing, visualization, and HCI. Eugene Wu received his B.S. from MIT, and was a postdoc in the AMPLab. A profile, an obit.

Eugene Wu has received the VLDB 2018 10-year test of time award, best-of-conference citations at VLDB, the SIGMOD 2016 best demo award, the NSF CAREER, and the Google and Amazon faculty awards.

The WuLab Website & Blog
We are recruiting PhDs + Postdocs, and Interns + UGrad + Masters!

[Overview of My Research and Teaching](#)

NEWS

Jun-2021: Looking forward to giving one of the keynotes at SEA DATA at VLDB 2021 this summer!

Aug-2020: For Highly Interactive Apps, Prediction is Not Enough is a blog post to introduce our Kameleon paper. Haneen also recorded a short YouTube video summarizing our work.

Jul-2020: FLAWLESS VICTORY! Kameleon, our

SELECTED PUBLICATIONS (SHOW ALL)

Private Federated Explanation of Inference Queries
Young Wu, Yiqia Lu, Lampros Flekas, Jianan Wang, Eugene Wu
VLDB 2022

Explaining SQL-ML Queries with Bayesian Optimization
Brandon Lockhard, Jianan Wang, Eugene Wu

From Debugging Before ML to Cleaning For ML
Felix Nezura, Binger Chen, Ziaessab Abdess, Eugene Wu
Invited, IEEE Data Engineering Bulletin 2021

Continuous Preference for Interactive Data Applications
Haneen Mohammad, Ziyun Wei, Rav Nettivalli, Eugene Wu
VLDB 2020 Talk Video Biogpost

Complaint-driven Training Data Debugging for Query 2.0
Young Wu, Lampros Flekas, Jianan Wang, Eugene Wu
SIGMOD 2020 Talk Video Biogpost

Monte Carlo Tree Search for Generating Interactive Data Analysis Interfaces
Yiqia Chen, Eugene Wu

EUGENE WU



BIO

Eugene Wu is broadly interested in technologies that help users play with their data. His goal is for all technical levels to effectively and quickly make sense of their information. He is interested in solutions that ultimately improve the interface between users and data, and uses techniques borrowed from fields such as data management, systems, crowd sourcing, visualization, and HCI. Eugene Wu received his B.S. from MIT, and was a postdoc in the AMPLab. A profile, an obit.

Eugene Wu has received the VLDB 2018 10-year test of time award, best-of-conference citations at VLDB, the SIGMOD 2016 best demo award, the NSF CAREER, and the Google and Amazon faculty awards.

The WuLab Website & Blog
We are recruiting PhDs + Postdocs, and Interns + UGrad + Masters!

[Overview of My Research and Teaching](#)

NEWS

Jun-2021: Looking forward to giving one of the keynotes at SEA DATA at VLDB 2021 this summer!

Aug-2020: For Highly Interactive Apps, Prediction is Not Enough is a blog post to introduce our Kameleon paper. Haneen also recorded a short YouTube video summarizing our work.

Jul-2020: FLAWLESS VICTORY! Kameleon, our

SELECTED PUBLICATIONS (SHOW ALL)

Private Federated Explanation of Inference Queries
Young Wu, Yiqia Lu, Lampros Flekas, Jianan Wang, Eugene Wu
VLDB 2022

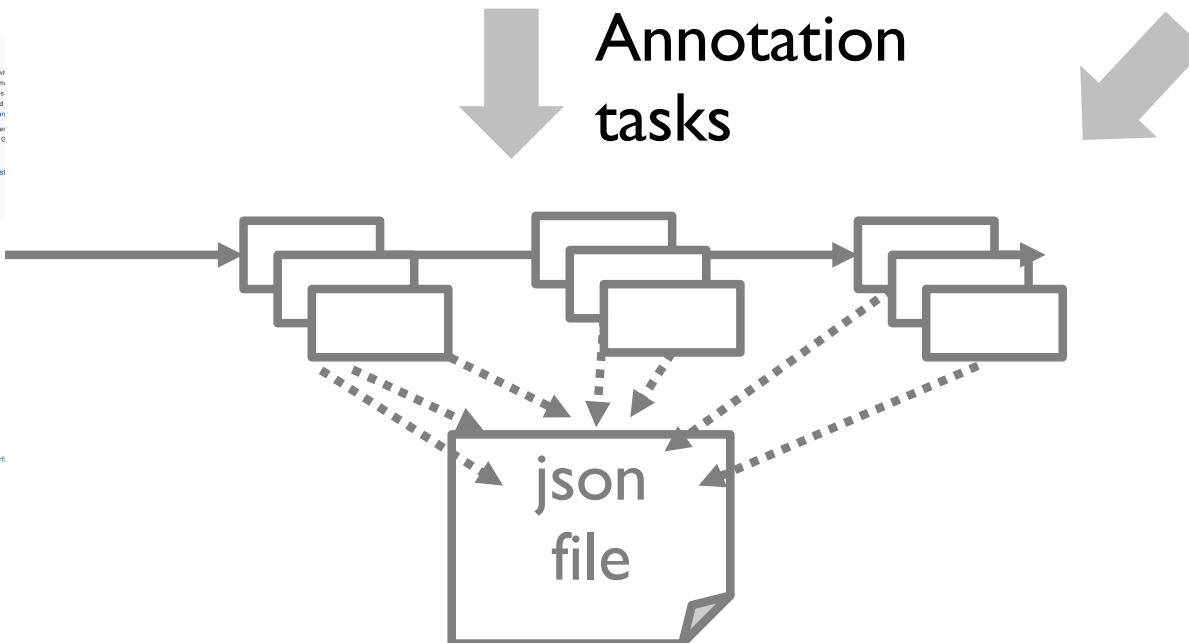
Explaining SQL-ML Queries with Bayesian Optimization
Brandon Lockhard, Jianan Wang, Eugene Wu

From Debugging Before ML to Cleaning For ML
Felix Nezura, Binger Chen, Ziaessab Abdess, Eugene Wu
Invited, IEEE Data Engineering Bulletin 2021

Continuous Preference for Interactive Data Applications
Haneen Mohammad, Ziyun Wei, Rav Nettivalli, Eugene Wu
VLDB 2020 Talk Video Biogpost

Complaint-driven Training Data Debugging for Query 2.0
Young Wu, Lampros Flekas, Jianan Wang, Eugene Wu
SIGMOD 2020 Talk Video Biogpost

Monte Carlo Tree Search for Generating Interactive Data Analysis Interfaces
Yiqia Chen, Eugene Wu



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	<i>Fast: 100k+ queries / sec</i>

Encompasses most of CS

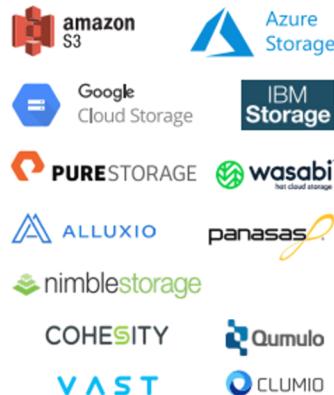
OS	DBMS directly manages hardware
Languages	SQL is a domain specific language
Theory	Algorithms, models, NP-complete
AI/ML	Knowledge Discovery, KDD
Logic	Relational Algebra = 1 st order logic

Scalable Computer Science

Golden Era of Data Systems!

INFRASTRUCTURE

STORAGE



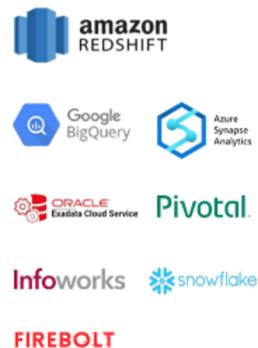
HADOOP



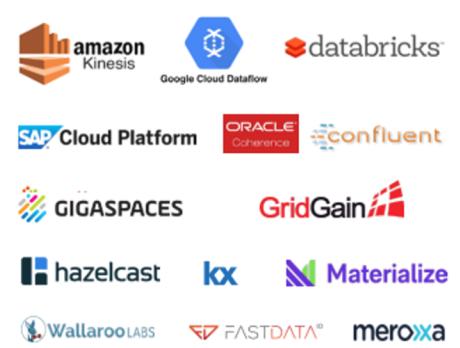
DATA LAKES



DATA WAREHOUSES



STREAMING / IN-MEMORY



NoSQL DATABASES



NewSQL DATABASES



GRAPH DBs



MPP DBs



SERVER-LESS



CLUSTER SVCS



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

External Schema

**Conceptual
Schema**

Describes logical
structure

Conceptual Schema

Physical Schema

Describes files,
formats, indexes

Physical Schema

“Data”

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

Conceptual Schema

Describes logical structure

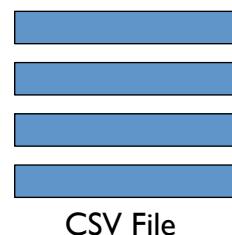
Physical Schema

Describes files and indexes

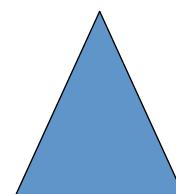
Conceptual Schema is the API!

Users(uid int, name str, age int)

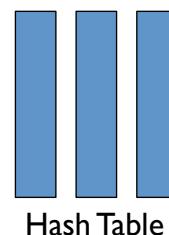
Physical Independence



CSV File



Tree Index



Hash Table

“Data”

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

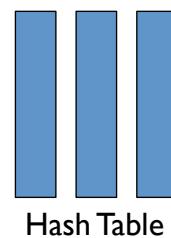
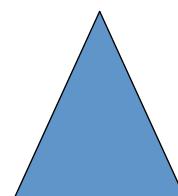
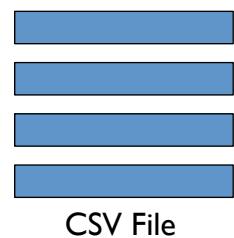
Describes logical structure

Physical Schema

Describes files and indexes

`Users(uid int, name str, age int)`

Physical Independence



“Data”

Data Independence

Conceptual Schema

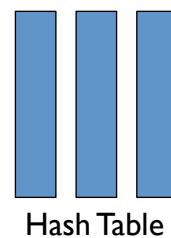
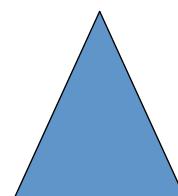
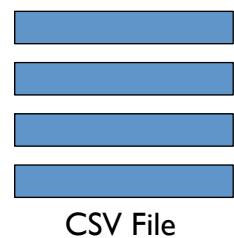
Describes logical structure

Physical Schema

Describes files and indexes

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

Physical Independence



“Data”

Data Independence

External Schema

Describe how users see data

Conceptual Schema

Describes logical structure

Physical Schema

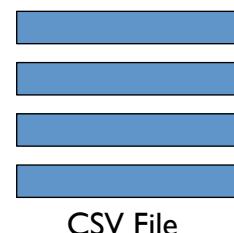
Describes files and indexes

Users(uid int, **name str**, age int)

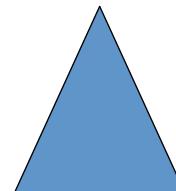
Logical Independence

Users(uid int, **fname str**, **Iname str**, age int)

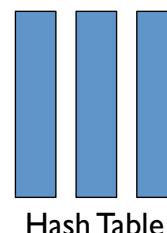
Physical Independence



CSV File



Tree Index



Hash Table

“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”

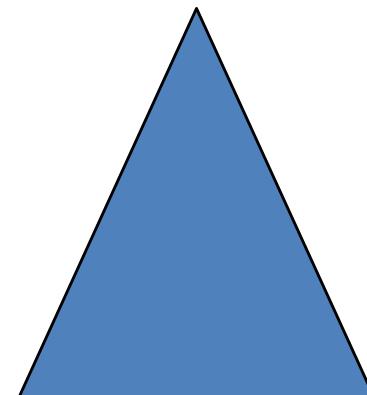
Declarative Interface

`SELECT name FROM users WHERE rating > 8`

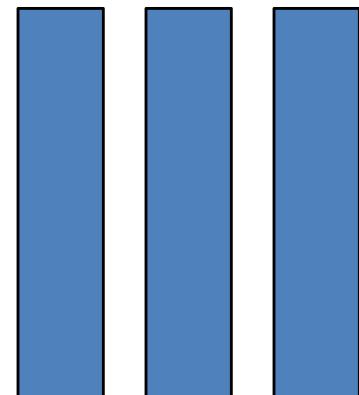
DBMS



CSV File



Tree Index



Hash Table

Declarative Interface

SELECT name FROM users WHERE rating > 8

DBMS

Node

Node

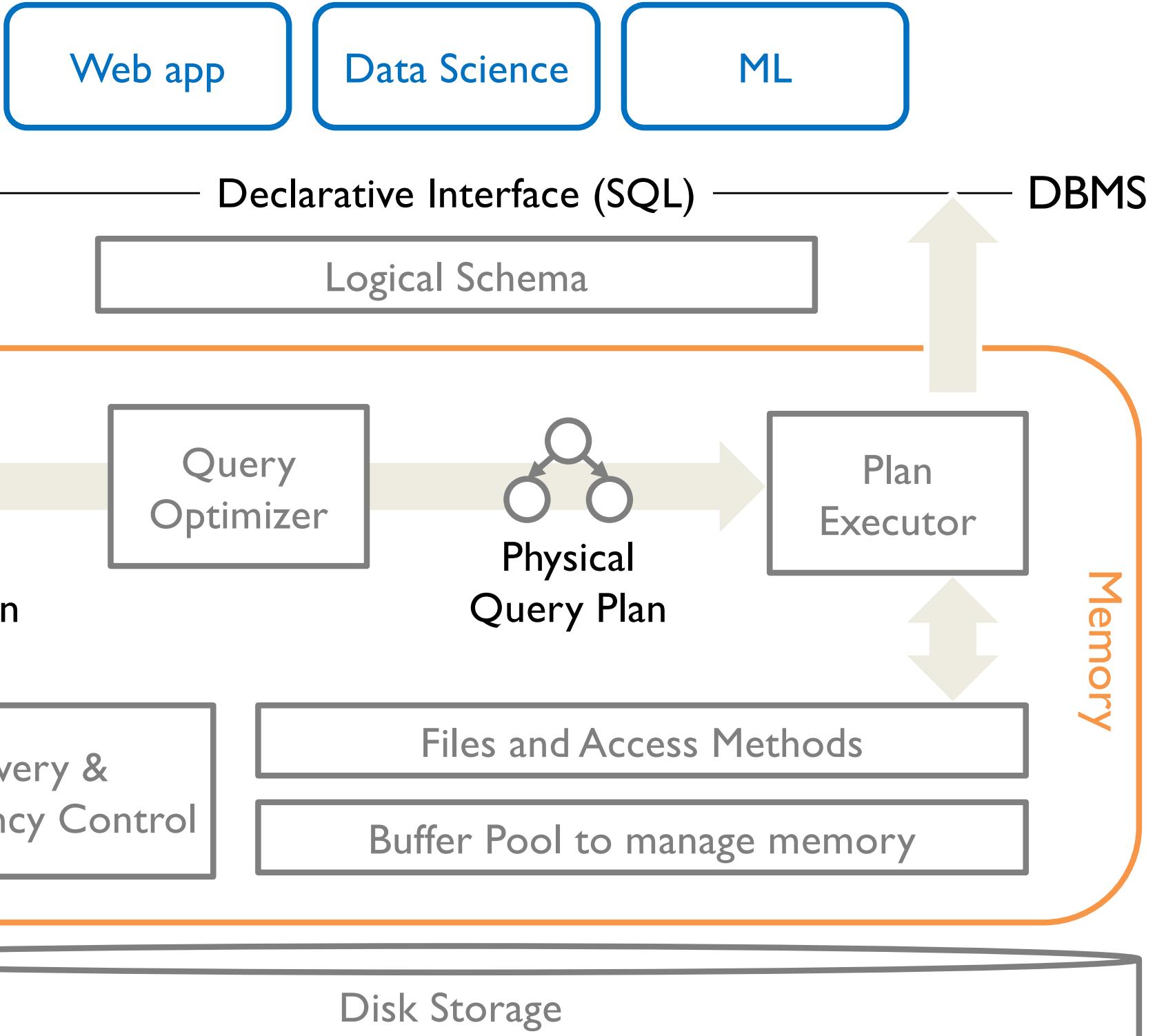
Node

Declarative Interface

`SELECT name FROM users WHERE rating > 8`

DBMS

Node



Web app
L8

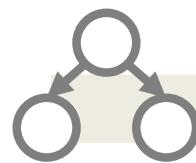
Data Science
L8

ML
L8

Declarative Interface (SQL L6-7)

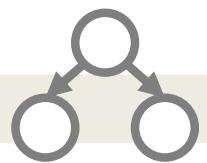
DBMS

Logical Schema L1-4, 9



Logical
Query Plan L5

Query
Optimizer
L11



Physical
Query Plan

Plan
Executor
L11

Memory

Recovery &
Concurrency Control
L12

Files and Access Methods L10

Buffer Pool to manage memory L10

Disk Storage L10

Web app
L8

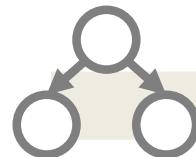
Data Science
L8

ML
L8

Declarative Interface (SQL L6-7)

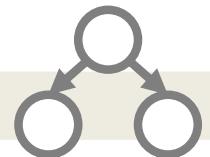
DBMS

Logical Schema L1-4, 9



Logical
Query Plan L5

Query
Optimizer
L11



Physical
Query Plan

Plan
Executor
L11

Recovery &
Concurrency Control
L12

Files and Access Methods L10

Buffer Pool to manage memory L10

Memory

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)

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

ed COMS W4111 001 – Discussion

New Thread

COURSES +

COMS W4111 001

CATEGORIES

- General
- Lectures
- HW
- Projects
- Social

Search Filter ▾

HW0 now posted

Zachary Huang STAFF 11h

11 hours ago in HW - HW0

UNPIN STAR WATCH 82 VIEWS

HW0 now posted #2

Zachary Huang STAFF 11h ago in HW - HW0

Dear all,

HW0 is now available at <https://github.com/w4111/hw0>.

Please follow the instructions in Google colab and submit the final file hw0.py to Gradescope.

Due date: 1/26 11:59 PM.

Comment Edit Delete ...

Sort by Newest ▾

Add comment

S Solomon Chang 8h ✓ Resolved

Hi, I tried to access the GitHub page but got a 404. Are there restrictions on who can access the HW0? Does the GitHub account have to be associated with a [columbia.edu](#) email address?

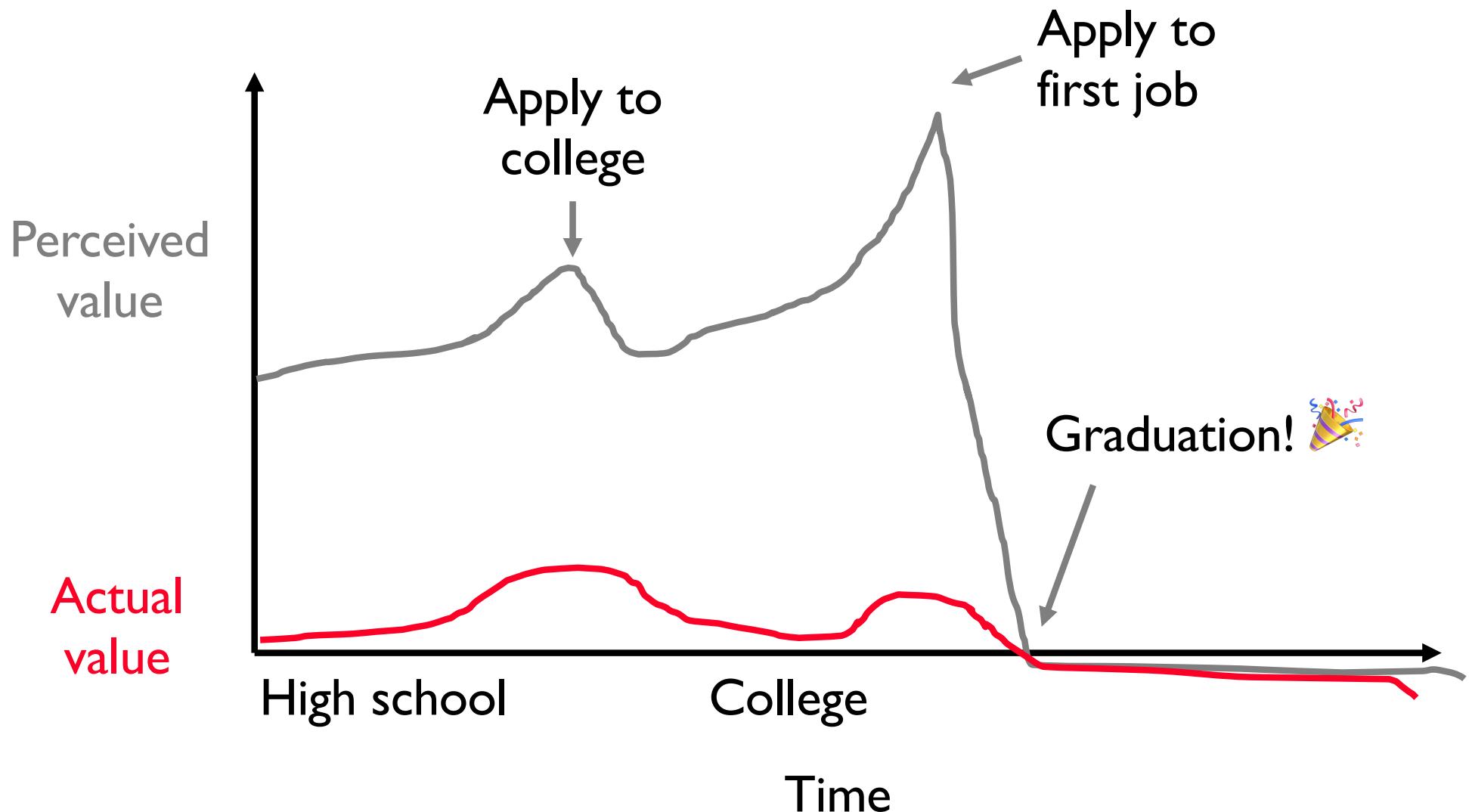
Reply Edit Delete Endorse ...

Z Zachary Huang STAFF 7h

Hi Solomon, thank you for your early start. I just made the repository public. Notice that HW0 in Gradescope will be available tomorrow at 10 am.

1 Reply Edit Delete Endorse ...

Grades. How do they work?



Grading Information

Midterm I 25%

Midterm 2 40%

HW 15% (4 HWs equally weighed)

Project I 15%

Project 2 5%

Extra credit variable

Median grade: B or slightly higher.

Exam Dates

Midterm I 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

The image displays two screenshots of a mobile application interface for a sports community group named "UNYSport".

Screenshot 1 (Top): Shows the group profile. At the top, there are two status bars: the left one shows AT&T signal, 5:30 PM, and 7% battery; the right one shows Camera mode, 7:21 PM, and 30% battery. Below the status bars, the group name "UNYSport" is displayed with a blue circular icon containing a person symbol. The group details are listed as follows:

- Name: Columbia Bouldering
- Sport: Bouldering
- Capacity: 100

Screenshot 2 (Bottom): Shows a messaging screen. At the top, it says APR 15, 2019. A message from a user with the initials TG at 11:21 PM reads: "Training with Coach P tonight! Dont be late!". A reply from a user with a blue profile picture at 11:21 PM reads: "Hi everyone, do you want to having a climbing practice this Thursday?". Another reply from the same user at 11:21 PM reads: "Sounds, great! Lets do it!". A final message from the user with the initials TG at 11:21 PM reads: "Can someone please share their chalk with me today!". On the left side of the messaging screen, there is a vertical list of buttons:

- Message
- View Members
- Events
- Leave Group

W4111 Introduction to databases**Department:** Computer science**Description:**

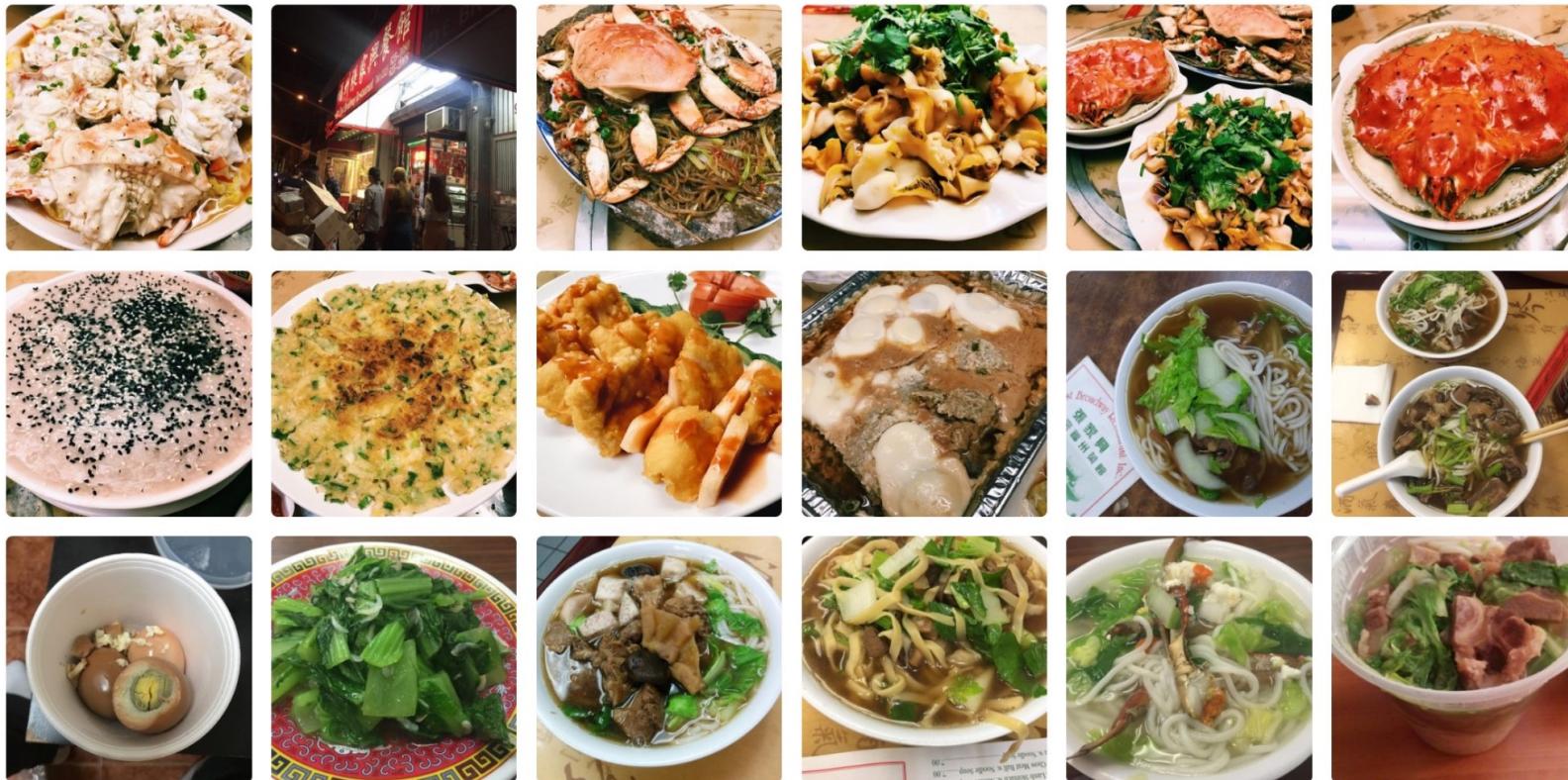
Prerequisites: (COMS W3134) or (COMS W3137) or (COMS W3136) and fluency in Java); or the instructor's permission. The fundamentals of database design and application development using databases: entity-relationship modeling, logical design of relational databases, relational data definition and manipulation 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: 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.

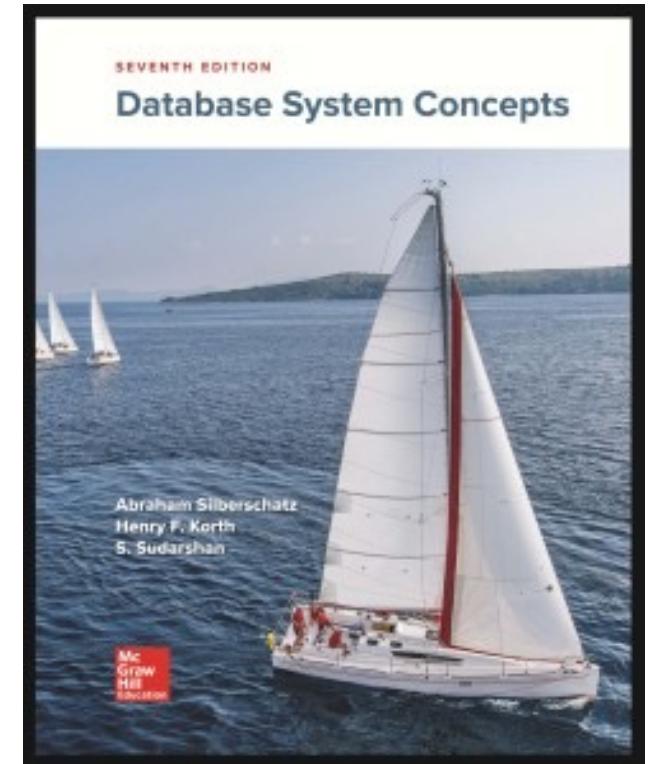
Contact the Professor Wu
right away if you have any questions or are falling behind.

Optional Textbook

Silberschatz et al.

Database System Concepts

7th ed



On-going Feedback

C O L U M B I A U N I V E R S I T Y C O M S W 4 1 1 1

INTRODUCTION TO DATABASES

Information

- Fri 10-12:40
[501 Schermerhorn](#)
- 3 units
- [Syllabus](#)
- [Ed Discussion](#)
- [Provide Feedback](#)
- [Course Github](#)

Overview

The goal of this class is two-fold. First, to introduce you to core database concepts (e.g., data modeling, normalization) so that you too can build a billion dollar application. Second, to teach enough about database engine internals (e.g., storage, query optimization, transaction processing) so you have a good sense of why queries may fail or take a long time. We will also discuss their relevance to systems used in industry.

Please do not ask me about the waitlist



Announcements

- [HW0](#) released.

Staff

- [Eugene Wu](#) Instructor
Weds 2-3PM
- [Zachary Huang](#)
- [Jacob Fisher](#)

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

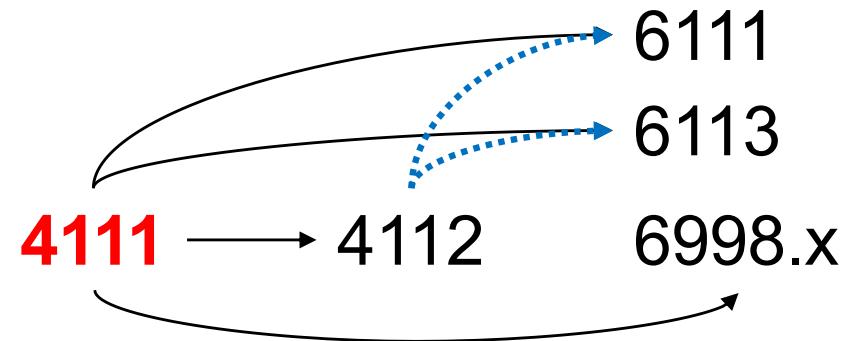
Data Warehousing, OLAP, Decision Support

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`



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!