CS150A Database

Lu Sun

School of Information Science and Technology

ShanghaiTech University

Feb. 28, 2024

Today:

- Introduction to database systems
- Course logistics

Readings:

- Database Management Systems (DBMS), Chapter 1

Acknowledgement: Joe Hellerstein@UCBerkeley's scourse notes

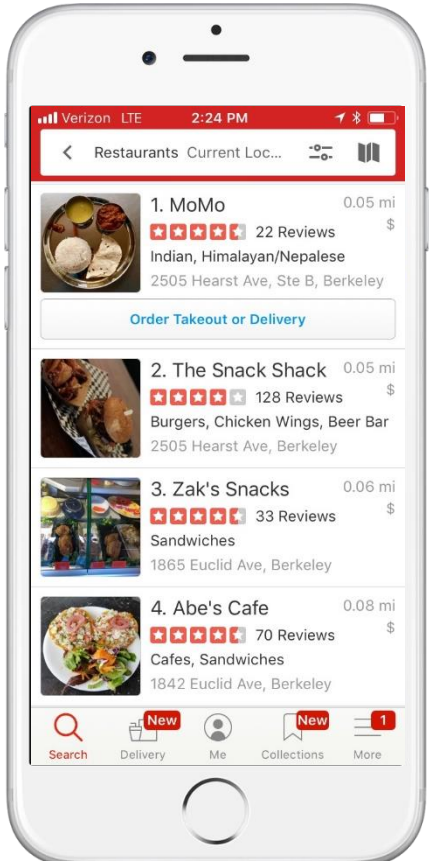
Essential Queries

- **Why** take this class?
- **What** is this class all about?
- **Who** is running this?
- **How** will this class work?

Why? Reason #1: Utility

- This class is very, very useful
  - Data processing backs essentially every app
  - Databases of one form or another back most apps
  - The *principles* taught in this class back nearly everything in computing

Where shall I eat, Database?

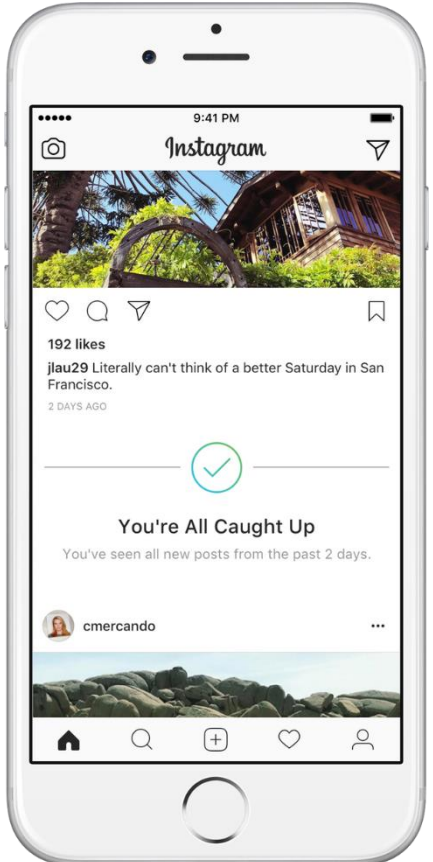


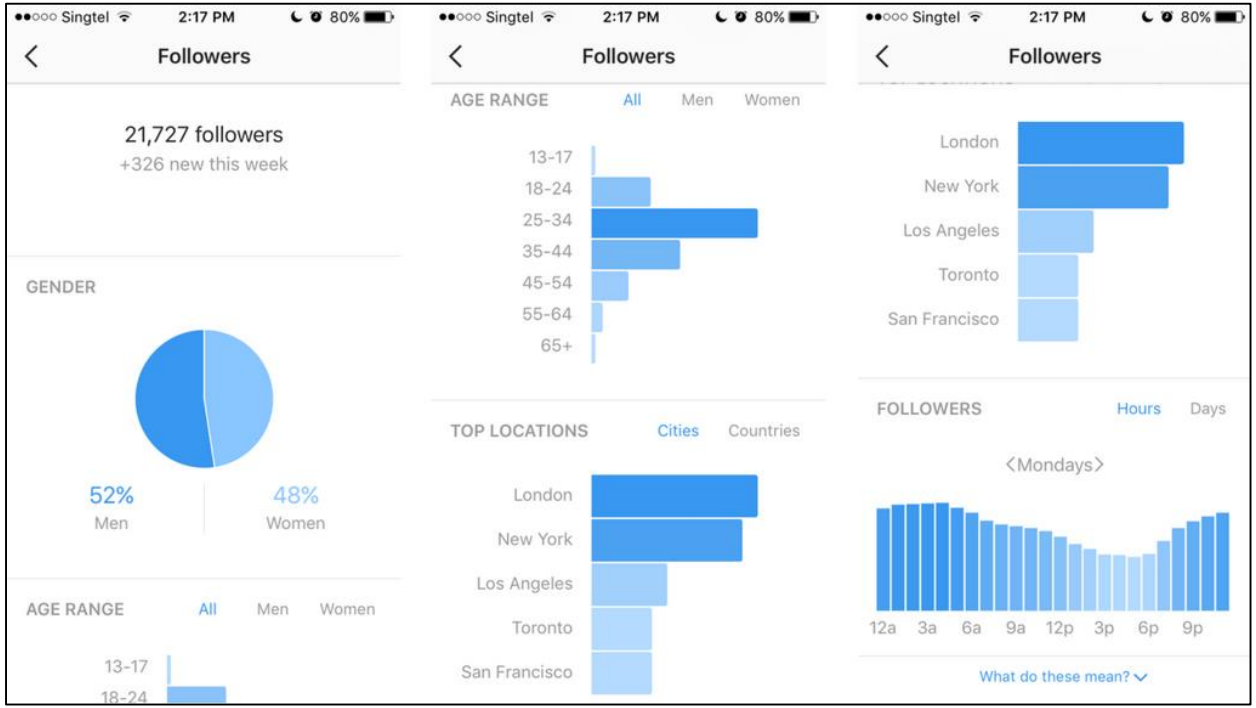
Each ratings star added on a Yelp restaurant review translated to anywhere from a 5% to a 9% effect on revenues.

—Harvard Business School, 2011

http://hbswk.hbs.edu/item/the-yelp-factor-are-consumer-reviews-good-for-

What am I missing, Database?

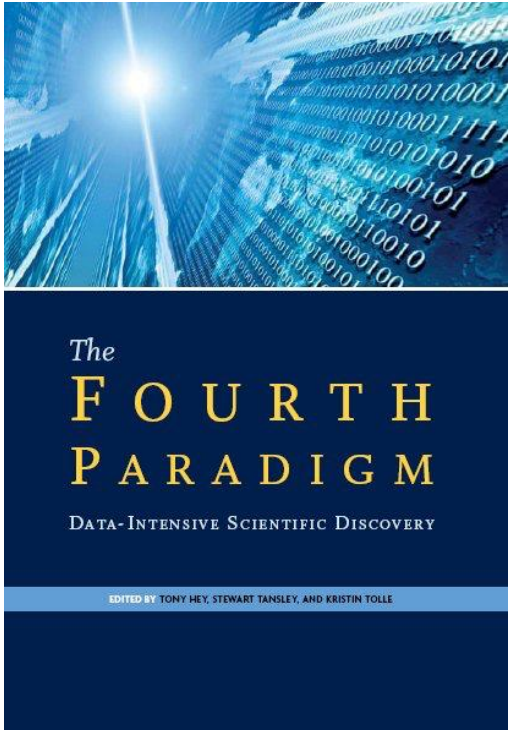




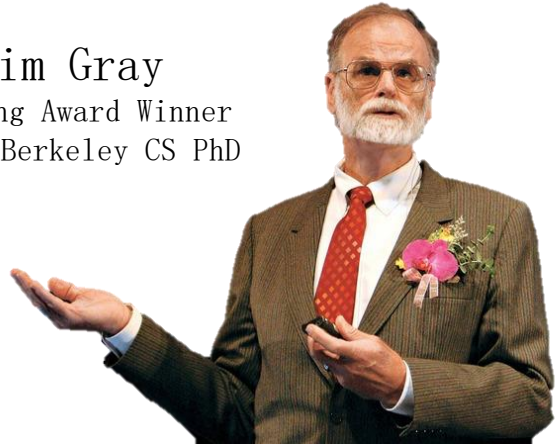
https://blog.bufferapp.com/instagram-analytics

https://instagram-press.com/blog/2018/07/02/introducing-youre-all-caught-up-in-

How does Science work? Database.



Jim Gray
Turing Award Winner
First Berkeley CS PhD



How does Science work? Database. Pt 2

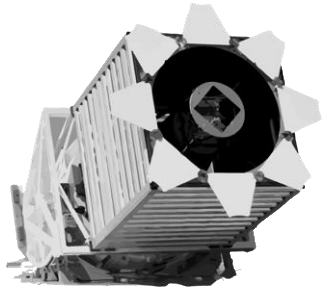
Experimental

Theoretical

Simulation

Data
Intensive

Astronomy in the 4th Paradigm



Sloan Digital
Sky Survey (SDSS)

+

Database
Systems

Sky Server

Science in the 4th Paradigm

http://www.sciserver.org

Why? Reason #1: Utility (again)

- This class is very, very useful
  - Data processing backs essentially every app
  - Databases of one form or another back most apps
  - The *principles* taught in this class back nearly everything in computing
- This material will empower you.

Why? Reason #2: Centrality

- Data is at the center of modern society.
- Unprecedented in its nature and significance
  - *Particular* and *voluminous*
  - Often asymmetric
    - low value in isolation, high value when aggregated
  - Difficult to protect

Google Knows

Opinion Facebook

Are you ready? Here is all the data Facebook and Google have on you

Dylan Curran

Google knows where you've been

Google knows everything you've ever searched – and deleted

Google has an advertisement profile of you

Google knows all the apps you use

l of your YouTube history

tores everything from your stickers to your login

access your webcam and microphone

knows which events you attended, and when

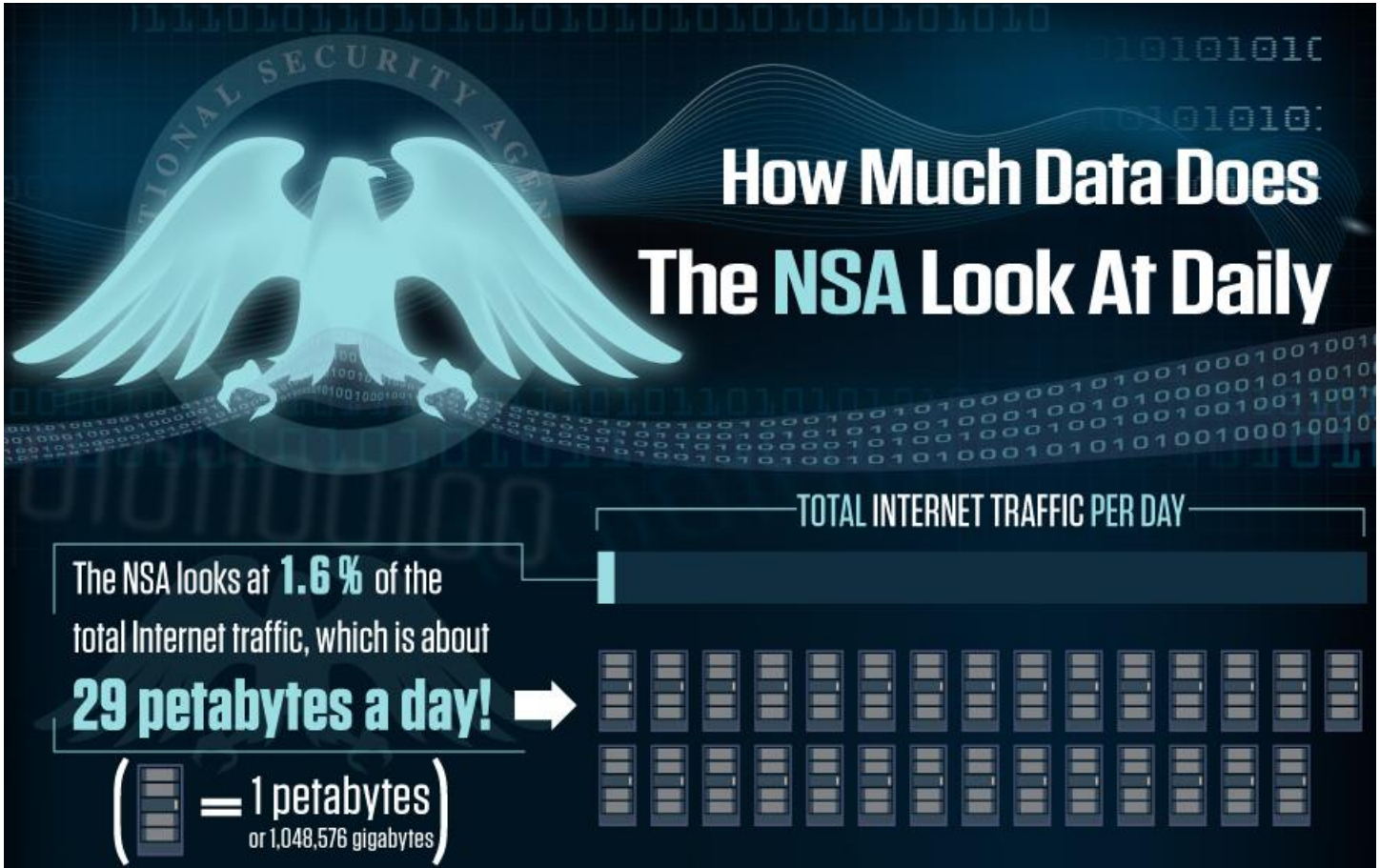
gle can know your workout routine

nd they have years' worth of photos

Google has every email you ever sent

> "Manage to gain access to someone's Google account? Perfect, you have a diary of everything that person has done

National Security Data: 2010

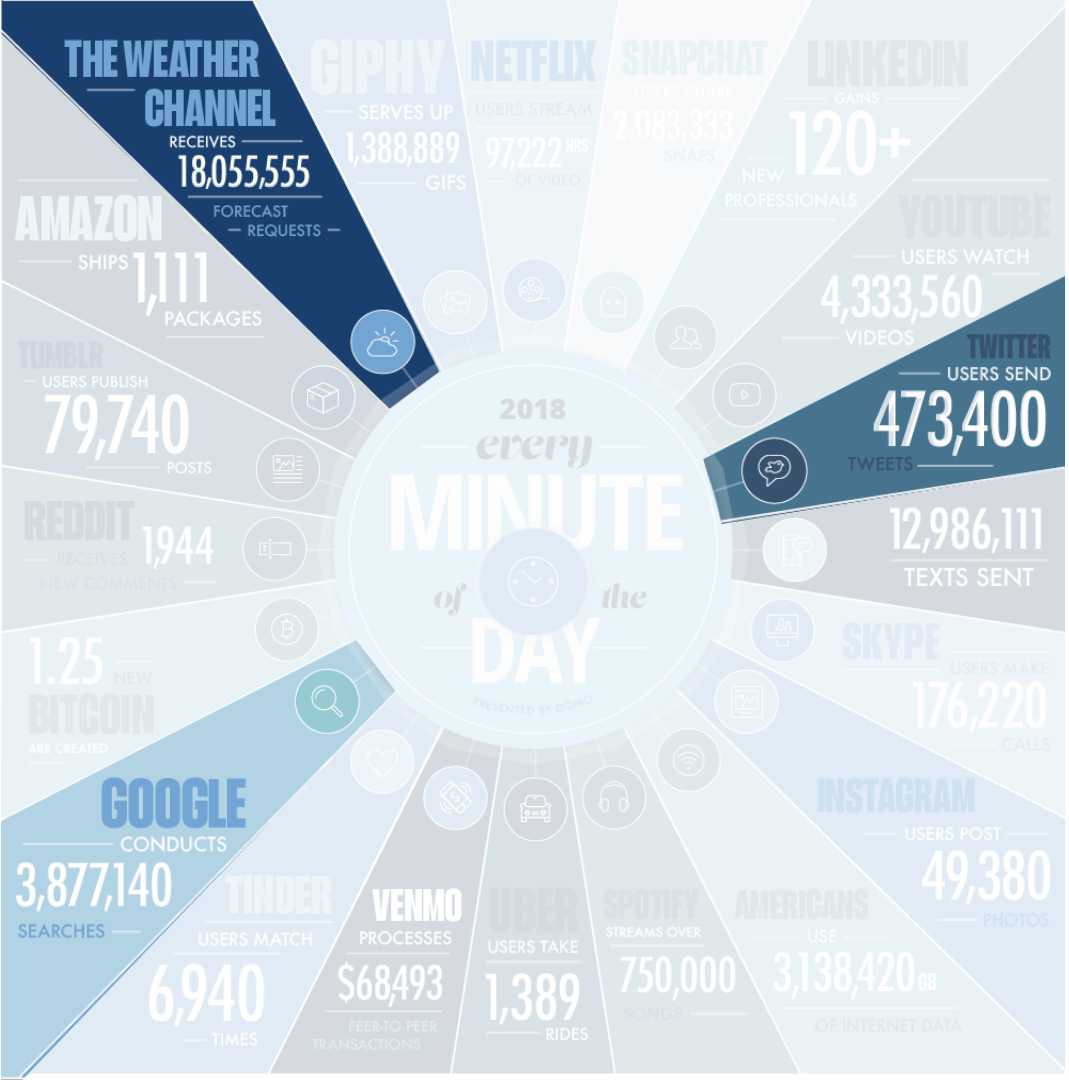


Why? Reason #2: Centrality (again)

- Data is at the center of modern society.
- Unprecedented in its nature and significance
  - *Particular* and *voluminous*
  - Often asymmetric
    - low value in isolation, high value when aggregated
  - Difficult to protect
- The infrastructure determines what's possible

Why #3? The Core of Computing

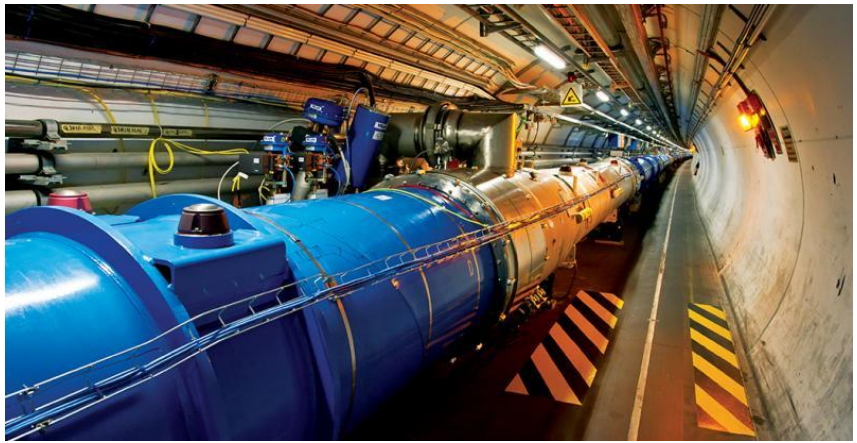
- Data growth will continue to outpace computation
- Systems for Data at Scale: the core of modern computing



Every Minute!

https://www.domo.com/learn/data-never-sleeps-5

Scale of Scientific Data



Metric prefixes in everyday use			
Text	Symbol	Factor	Power
yotta	Y	1 000 000 000 000 000 000 000 000	10^{24}
zetta	Z	1 000 000 000 000 000 000 000	10^{21}
exa	E	1 000 000 000 000 000 000	10^{18}
peta	P	1 000 000 000 000 000	10^{15}
tera	T	1 000 000 000 000	10^{12}
giga	G	1 000 000 000	10^{9}
mega	M	1 000 000	10^{6}
kilo	k	1 000	10^{3}

Large Hadron Collider, CERN

- Raw data: 1MB/event. 600,000,000 events/sec.
  = 1.9×10^{22} bytes/year = **19 ZettaBytes/year**
- Downsampled: 25GB/sec = 7.88×10^{17} bytes/year = **788 PetaBytes/year**
- Downsampled further: 1050MB/sec = $3.3*10^{16}$/year = **33 PetaBytes/year**

https://home.cern/about/computing/processing-what-record

Forces Driving Data Growth

- Ubiquitous sensors and reporting:
  - Cameras, mobile computing, blogging, ...
- Large collaborative science projects
- Philosophy: *More Data* → *More Value*?

Enabling Technology

- **Cheap, Scalable** Data Management Systems

http://hyperboleandahalf.blogspot.com

Why #3? The Core of Computing (again)

- Data growth will continue to outpace computation
- Systems for Data at Scale: the core of modern computing
- Techniques you learn in this class underlie many topics in computing

Essential Queries, Pt 2

- **Why** take this class?
- **What** is this class all about?
- **Who** is running this?
- **How** will this class work?

What is this class all about?

- Databases?
  - What is a database?
- Database Management Systems?
- Implementation?

Universal Symbol for a Database

Why the Symbol?

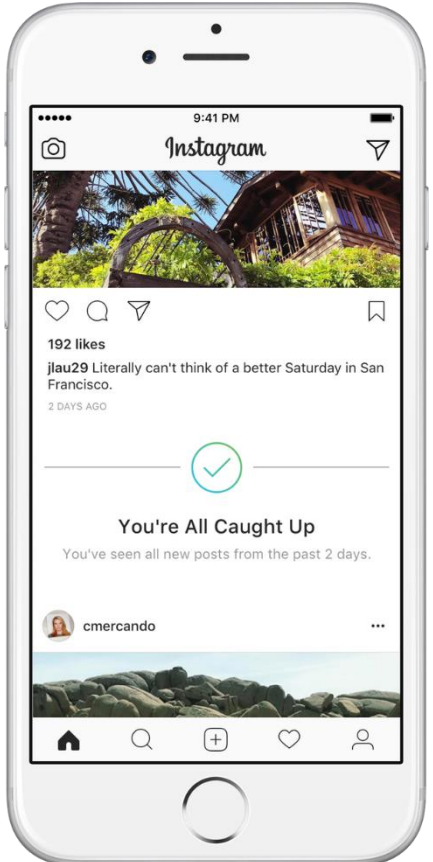
Looks Like?

Platters on a Disk Drive

Is This a Database?

- Rolodex
- Alphbetically ordered cards
- Indexed access by first letter

Is This a Database?, cont



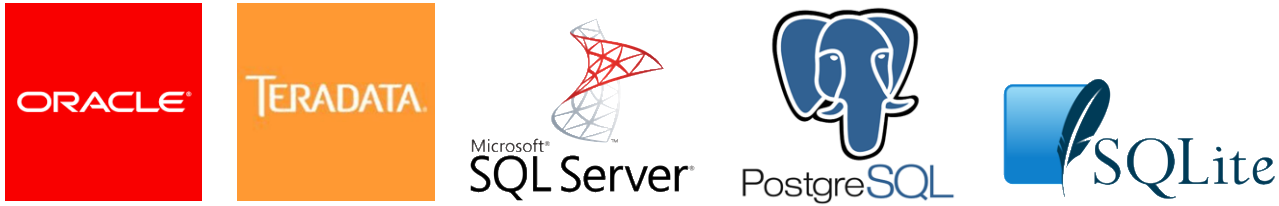
- A database + “business logic” + user interface?

What is a Database?

- Let’s not split hairs.
  - *A database is a large, organized collection of data.*
- Sometimes confused with a Database Management System (DBMS)
  - *A DBMS is software that **stores, manages,** and facilitates **access** to data.*

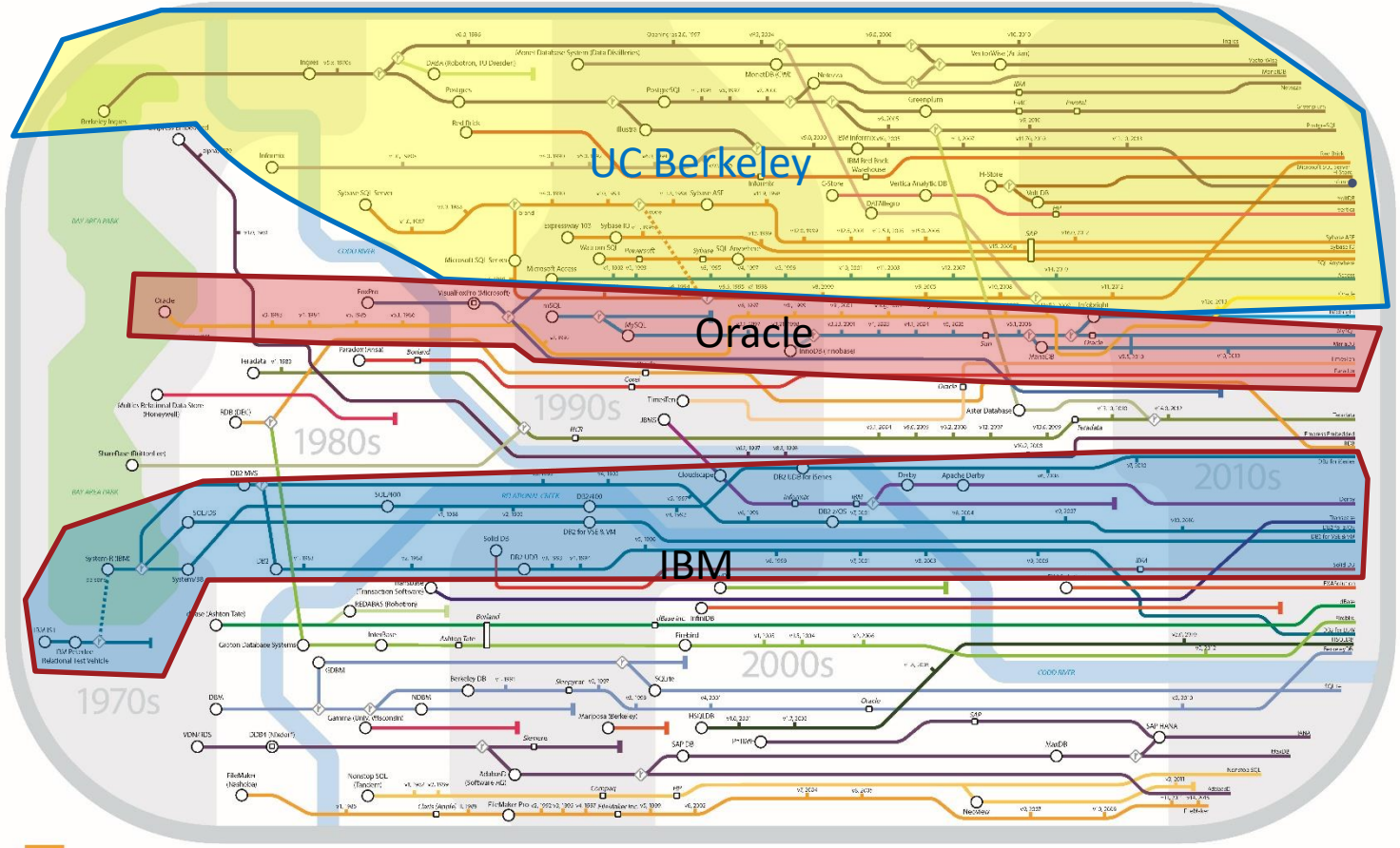
Relational DBMSs

- Traditionally DBMS referred to relational databases



- **RDBMS** is a more appropriate term
- **SQL** data description and manipulation language
- **ACID** transaction consistency
- **Durable** writes (prevent data loss)
- **Mature** technologies …

Genealogy of Relational Database Management Systems



Berkeley Roots!

- Ingres / Postgres
- Sybase
- Informix

Ranking of DBMS Technologies 2024

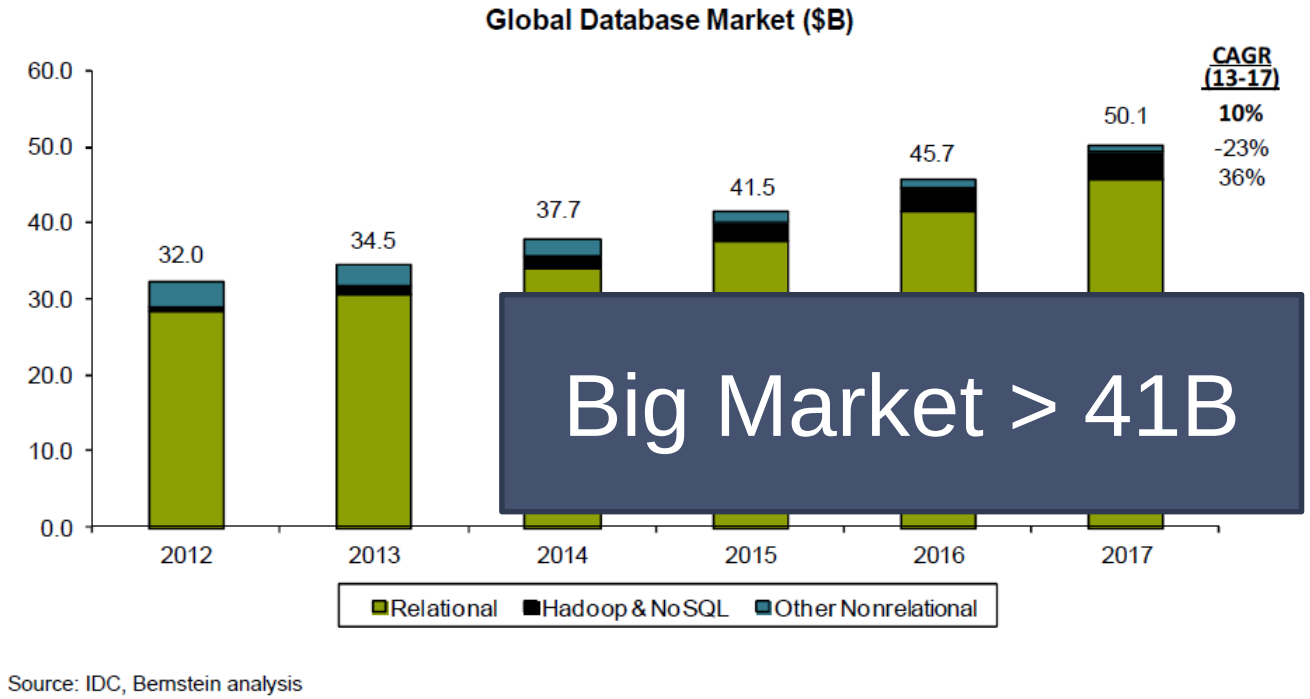
417 systems in ranking, February 2024

Rank			DBMS	Database Model	Score		
Feb 2024	Jan 2024	Feb 2023			Feb 2024	Jan 2024	Feb 2023
1.	1.	1.	Oracle	Relational, Multi-model	1241.45	-6.05	-6.08
2.	2.	2.	MySQL	Relational, Multi-model	1106.67	-16.79	-88.78
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model	853.57	-23.03	-75.52
4.	4.	4.	PostgreSQL	Relational, Multi-model	629.41	-19.55	+12.90
5.	5.	5.	MongoDB	Document, Multi-model	420.36	+2.88	-32.41
6.	6.	6.	Redis	Key-value, Multi-model	160.71	+1.33	-13.12
7.	7.	↑ 8.	Elasticsearch	Search engine, Multi-model	135.74	-0.33	-2.86
8.	8.	↓ 7.	IBM Db2	Relational, Multi-model	132.23	-0.18	-10.74
9.	9.	↑ 12.	Snowflake	Relational	127.45	+1.53	+11.80
10.	↑ 11.	↓ 9.	SQLite	Relational	117.28	+2.08	-15.38

Based on #mentions (e.g., stack overflow), google trends, job postings, profile data on LinkedIn, tweets …

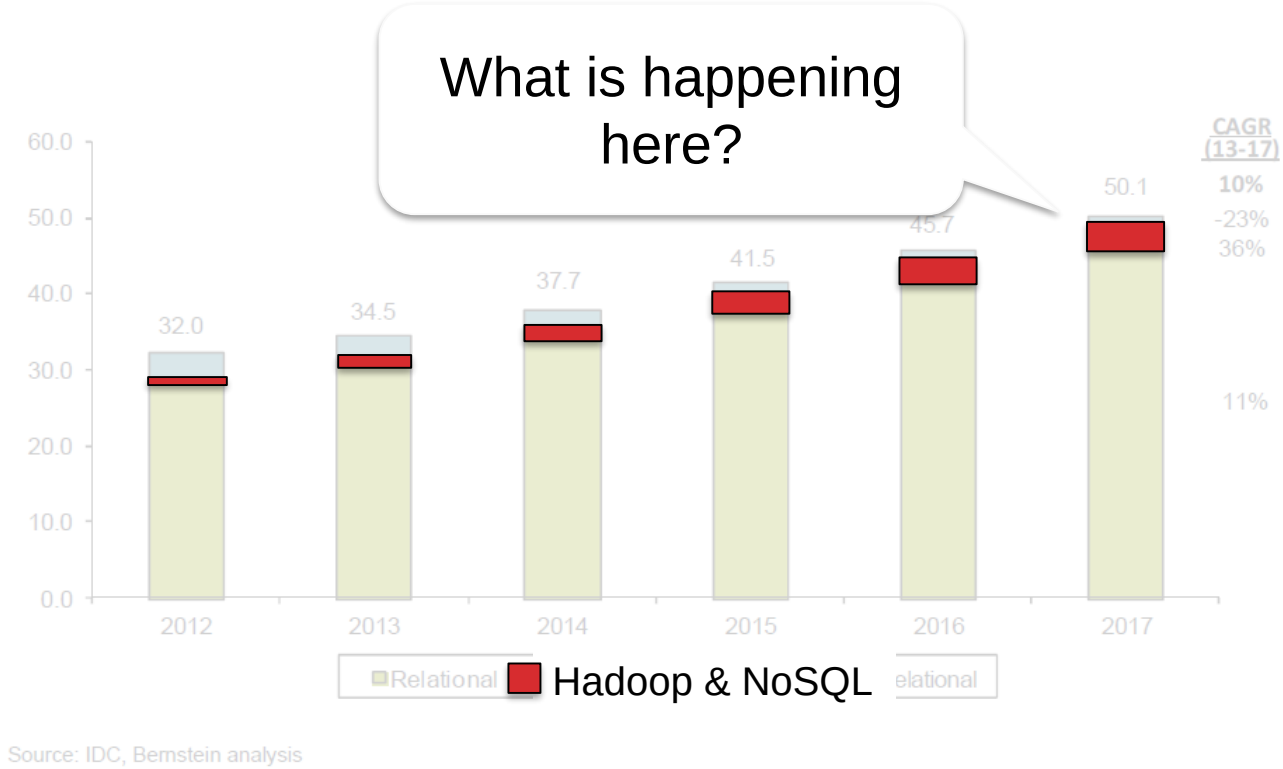
http://db-engines.com/en/ranking

Relational Database Market



http://www.infoworld.com/article/2916057/open-source-software/open-source-threatens-to-eat-the-database-market.html

Relational Database Market, cont



http://www.infoworld.com/article/2916057/open-source-software/open-source-threatens-to-eat-the-database-market.html

Market Trends

- Cloud DBMS disrupting on-premises vendors
  - Cloud is less relational-centric
  - But fastest-growing services at AWS are RDBMSs

- “One size doesn’t fit all”
  - Main-memory DBMS
  - Graph DBMS
  - TimeSeries DBMS
  - Key-Value Stores (NoSQL)
  - Analytics Platforms (Spark, Hadoop)

- Tools for working with data
  - Business Intelligence (charting tools)
  - Data Science platforms
  - Data preparation and next-generation data integration (ETL)

Reasons for Change

- **Hardware** trends: *RAM*, *SSDs*, *NVRAM*, *GPUs*, ...
- **Platform** trends: cloud and elastic computing
- Need to **scale**: *storage* and *transactions*
- New **data-types**: *text*, *json*, *image*, *video*…
- New **workloads**: *machine learning* & advanced *analytics*

Change = Opportunity!

- The DBMS world is rapidly changing
  - Our textbook is rather out of date (2003!)
- Opportunity!
  - You can shape the future of DBMSs
- We will not follow the textbook slavishly.

Instead…

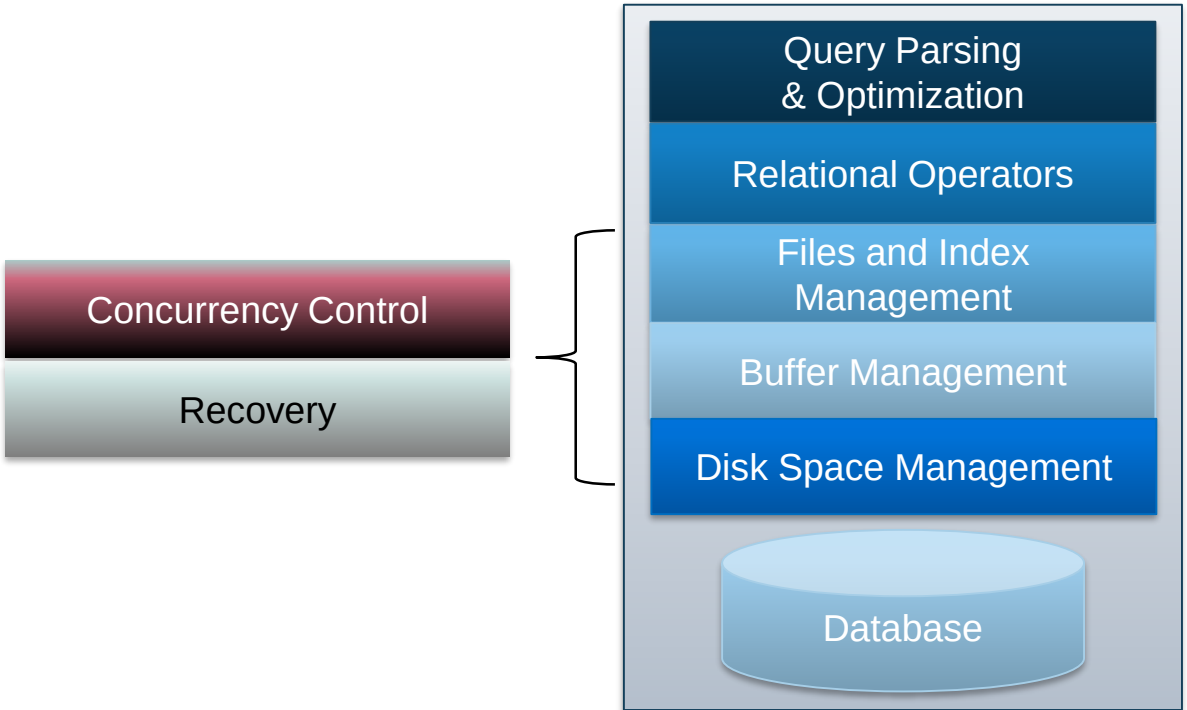
- Focus: **Foundational System Principles**
  - Reusable ideas and components
  - Compositional approach
- Goal:
  - You will be able to **use** existing & **build new** DBMS technologies!

You will learn...

- Data Oriented Programming with SQL
- Foundations of Database System Design
  - Storage, indexing
  - Query processing and optimization
- Transactions
  - Concurrency, Consistency, Recovery
- Data Modeling
  - Application-level representations of data

Systems

We will examine various levels of a DBMS



Essential Queries, Pt 3

- **Why** take this class?
- **What** is this class all about?
- **Who** is running this?
- **How** will this class work?

Instructors

- Prof. Wenjie Wang（王 雯婕）
  - Assistant Professor in SIST
  - Joined in Feb., 2023
  - PhD @ Emory University
  - Email: wangwj1@shanghaitech.edu.cn

- About Me: Lu Sun (孙 露)
  - Assistant Professor in SIST
  - Joined in Nov., 2019
  - PhD @ Hokkaido University
  - Email: sunlu1@shanghaitech.edu.cn

TAs

- Yang Liu (刘 洋)
  - M2 student in CS
  - liuyang12022@shanghaitech.edu.cn
- Chengyan Fu (傅 铖彦)
  - M1 student in CS
  - fuchy2023@shanghaitech.edu.cn
- Lingzheng Dong (董 凌铮)
  - B4 student in CS
  - A+ in CS150A 2022 Fall
  - donglzh@shanghaitech.edu.cn

Essential Queries, Part 4

- **Why** take this class?
- **What** is this class all about?
- **Who** is running this?
- **How** will this class work?

How will this class work?

General information

- Time: Web. & Fri., 10:15-11:55
- Online: Blackboard, Piazza & Gradescope
- 16 weeks (64 credit hours)
- RDBMS in weeks 1-13; Data mining in week 14; NoSQL&Hadoop in weeks 15-16

All class communication via Piazza

- https://piazza.com/shanghaitech.edu.cn/spring2024/cs150a
- announcements and discussion
- read it regularly
- post all questions/comments there
- direct email is not a good idea

How will this class work?

Grading

- Homework: 30%
- Midterm: 30%
- Final exam: 40%

Highlights

- Please write your HW and exam in English
- Submitted to GradeScope:
  - https://www.gradescope.com/courses/743504 (Entry Code: 8E7X2V)
- For late HW, the score will be exponentially decreased
- Once any plagiarism or cheating is confirmed, relevant assignments or exams will receive 0 points

How will this class work?

Recommended textbook

- **Database Management Systems, 3rd Edition**

Johannes Gehrke and Raghu Ramakrishnan

Some useful online resources

- UC Berkeley, CS186 Introduction to Database Systems course

https://cs186berkeley.net/

- Course videos

https://www.youtube.com/playlist?list=PLYp4IGUhNFmw8USiYMJvCUjZe79fvyYge

Our Topics

1. Intro. and SQL
2. Disk, Buffers and Files
3. Index and B+ Trees
4. Buffer Manager
5. Relational Algebra
6. Sorting and Hashing
7. Iterations and Joins
8. Query Optimization
9. Transactions and Concurrency
10. Recovery
11. ER Modeling
12. Parallel Querying
13. Distributed Transaction
14. Data Mining and ML
15. NoSQL
16. Hadoop and Spark