CS150A Database

Lu Sun

School of Information Science and Technology

ShanghaiTech University

Sept. 6, 2022

Today:

- Introduction to database systems
- Course logistics

Readings:

- Database Management Systems (DBMS), Chapter 1

Acknowledgement: Joe Hellerstein@UCBerkeley's scourse notes

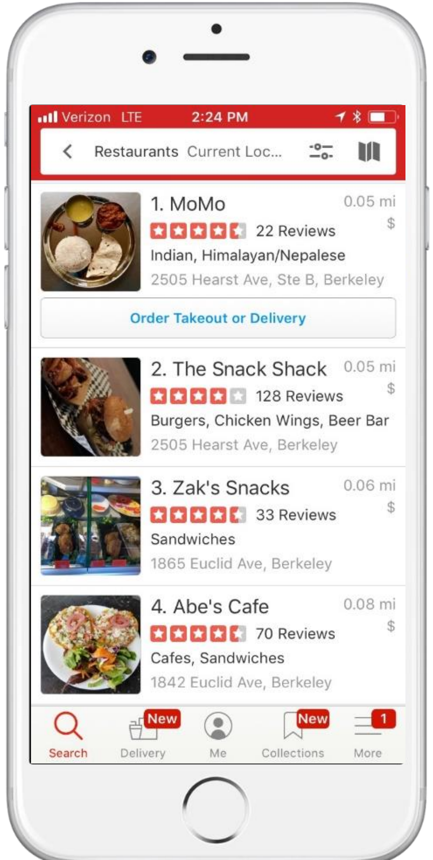
Essential Queries

- **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?



1. MoMo 0.05 mi $
22 Reviews
Indian, Himalayan/Nepalese
2505 Hearst Ave, Ste B, Berkeley

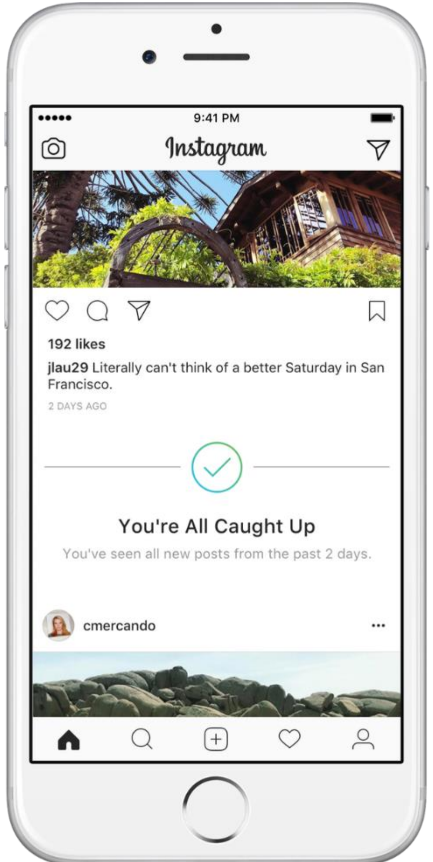
2. The Snack Shack 0.05 mi $
128 Reviews
Burgers, Chicken Wings, Beer Bar
2505 Hearst Ave, Berkeley

3. Zak's Snacks 0.06 mi $
33 Reviews
Sandwiches
1865 Euclid Ave, Berkeley

4. Abe's Cafe 0.08 mi $
70 Reviews
Cafes, Sandwiches
1842 Euclid Ave, Berkeley

> *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?



192 likes

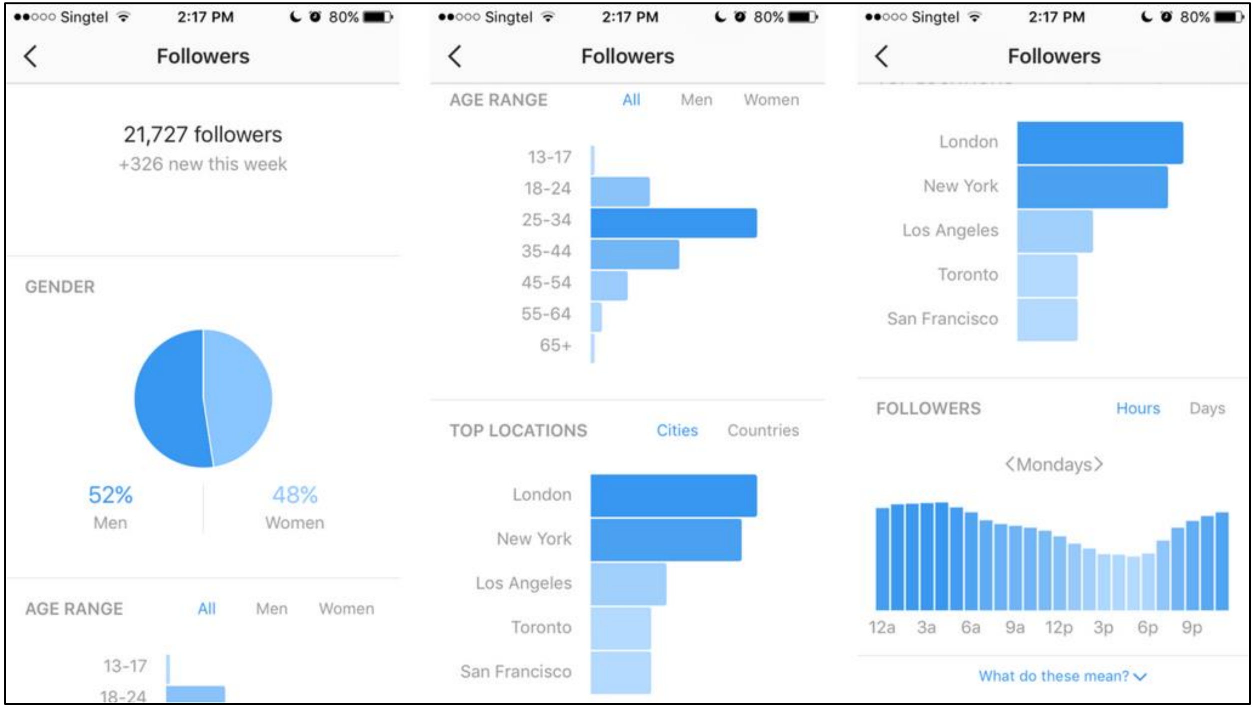
jlau29 Literally can't think of a better Saturday in San Francisco.

2 DAYS AGO

You're All Caught Up

You've seen all new posts from the past 2 days.

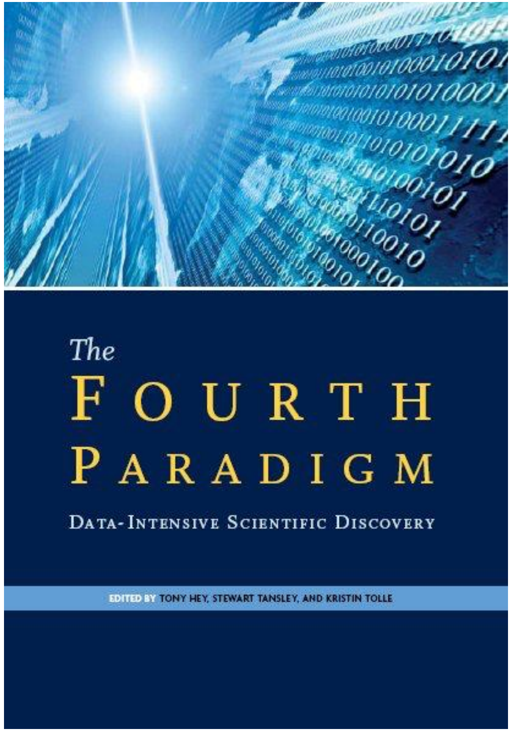
cmercando



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.



The

FOURTH

PARADIGM

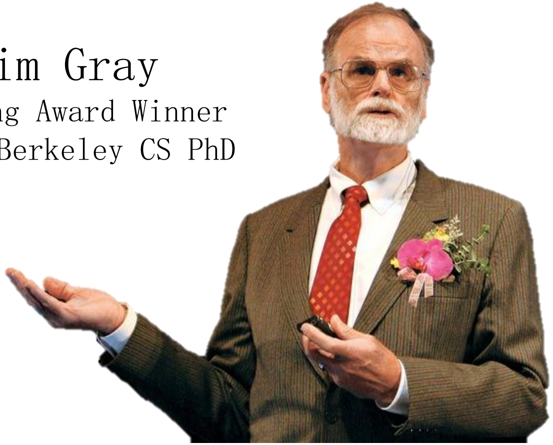
DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

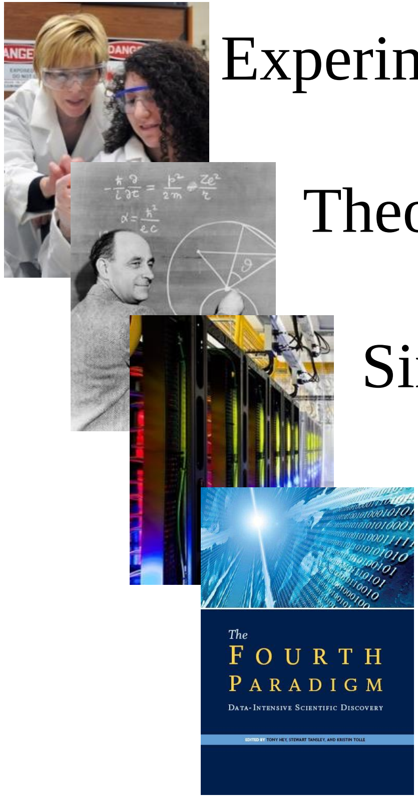
Jim Gray

Turing Award Winner

First Berkeley CS PhD



How does Science work? Database. Pt 2


The
FOURTH
PARADIGM
Data-Intensive Scientific Discovery
Edited by Tony Hey, Stewart Tansley, and Kristin Tolle

Experimental

Theoretical

Simulation

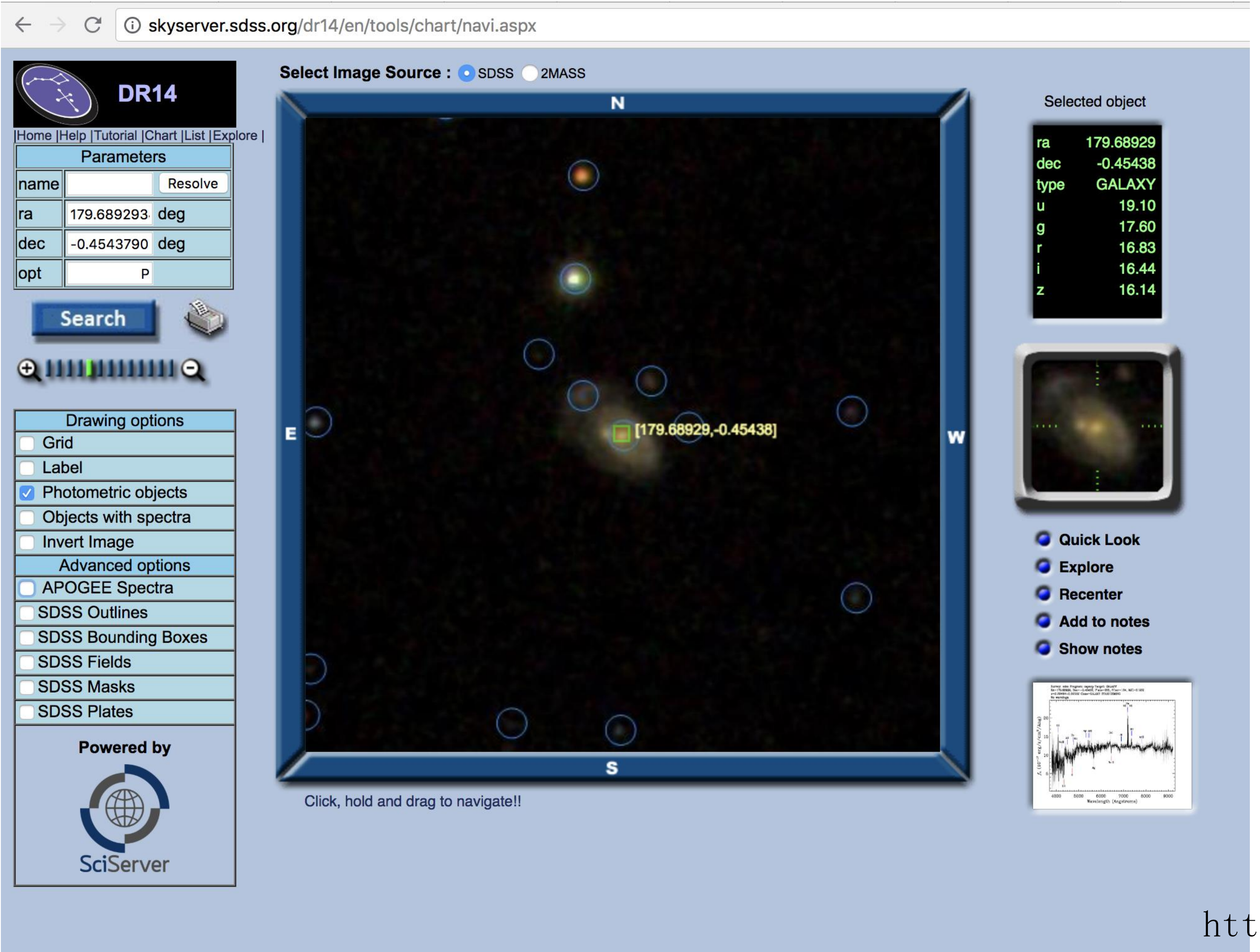
Data
Intensive

Astronomy in the 4th Paradigm

Sloan Digital
Sky Survey (SDSS)

Database
Systems

Sky Server



skyserver.sdss.org/dr14/en/tools/chart/navi.aspx

Parameters

name		Resolve
ra	179.689293	deg
dec	-0.4543790	deg
opt	P	

Drawing options

- □ Grid
- □ Label
- ☑ Photometric objects
- □ Objects with spectra
- □ Invert Image

Advanced options

- □ APOGEE Spectra
- □ SDSS Outlines
- □ SDSS Bounding Boxes
- □ SDSS Fields
- □ SDSS Masks
- □ SDSS Plates

Select Image Source : ● SDSS ○ 2MASS

Click, hold and drag to navigate!!

Selected object

ra	179.68929
dec	-0.45438
type	GALAXY
u	19.10
g	17.60
r	16.83
i	16.44
z	16.14

http://skyserver.sdss.org

Science in the 4th Paradigm

http://www.sciserver.org

Your career…

- 2000’s:
  - Shift from “programs” to apps over data-centric services
- More recently:
  - End of the full-stack programmer
  - New, ubiquitous professions:
    - Data Scientist
    - Data Engineer
    - Machine Learning Engineer
- Two things to acknowledge:
  - The fundamentals of this class will stay central
  - Other things will change
    - Be prepared to generalize from what you learn here
    - Keep learning new things

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.

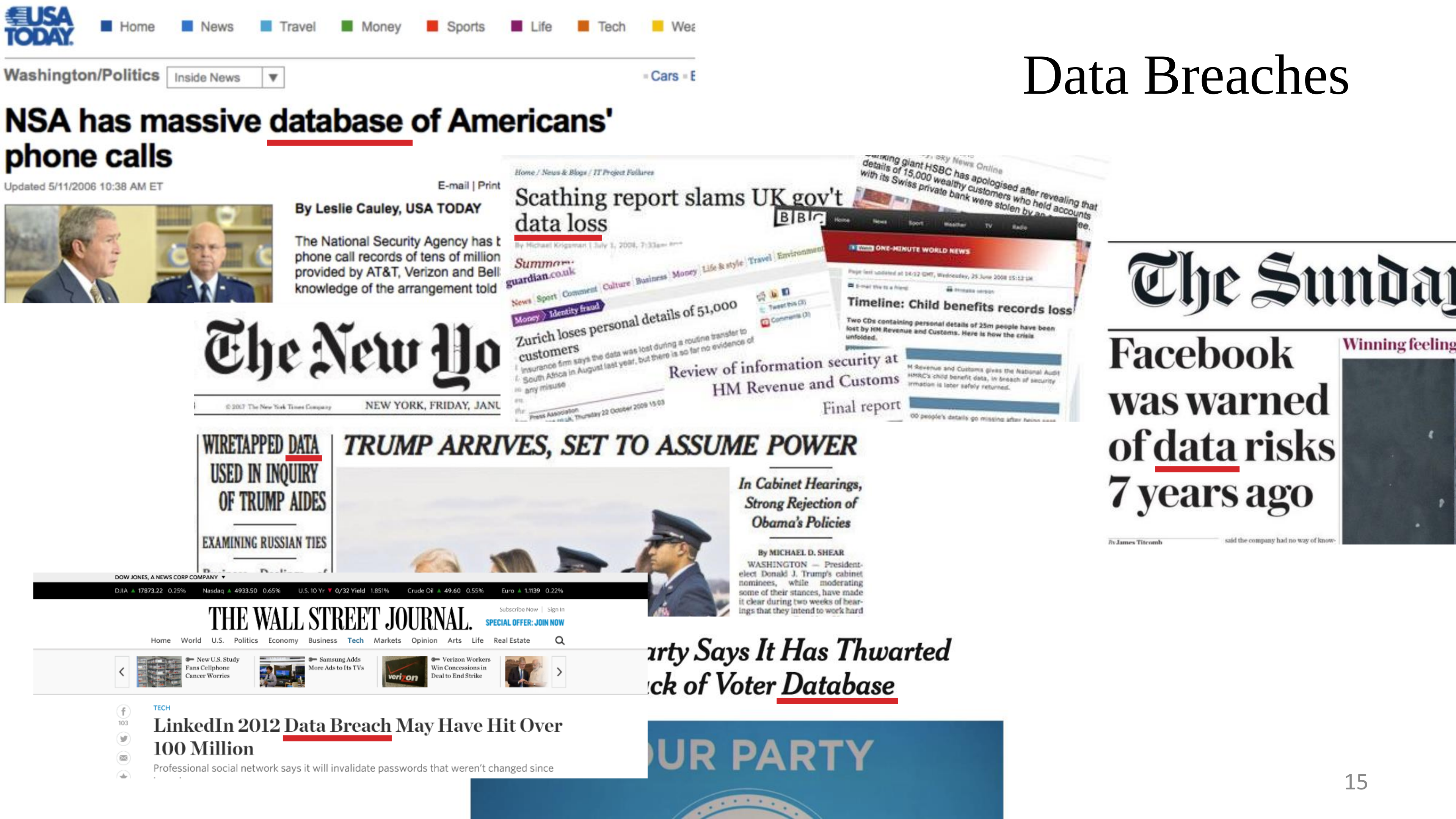
Why? Reason #2: Centrality

- 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

At the center of major issues

- Privacy
- National Security
- Fake News

Data Breaches



USA TODAY

Home News Travel Money Sports Life Tech Wea

Washington/Politics Inside News ▼ · Cars · E

NSA has massive database of Americans' phone calls

Updated 5/11/2006 10:38 AM ET E-mail | Print

By Leslie Cauley, USA TODAY

The National Security Agency has b phone call records of tens of million provided by AT&T, Verizon and Bell knowledge of the arrangement told

Home / News & Blogs / IT Project Failures

Scathing report slams UK gov't data loss

By Michael Krigsman | July 1, 2008, 7:33am

Summary:

guardian.co.uk

News Sport Comment Culture Business Money Life & style Travel Environment

Money Identity fraud

Zurich loses personal details of 51,000 customers

Insurance firm says the data was lost during a routine transfer to South Africa in August last year, but there is so far no evidence of any misuse

Press Association, Thursday 22 October 2009 15.03

Review of information security at HM Revenue and Customs

Final report

Sky News Online

Banking giant HSBC has apologised after revealing that details of 15,000 wealthy customers who held accounts with its Swiss private bank were stolen by an employee.

BBC Home News Sport Weather TV Radio

ONE-MINUTE WORLD NEWS

Page last updated at 14:12 GMT, Wednesday, 25 June 2008 15:12 UK

E-mail this to a friend Printable version

Timeline: Child benefits records loss

Two CDs containing personal details of 25m people have been lost by HM Revenue and Customs. Here is how the crisis unfolded.

The Sunday

Facebook was warned of data risks 7 years ago

Winning feeling

By James Titcomb

said the company had no way of know-

The New Yo

© 2017 The New York Times Company NEW YORK, FRIDAY, JANU

WIRETAPPED DATA USED IN INQUIRY OF TRUMP AIDES

EXAMINING RUSSIAN TIES

TRUMP ARRIVES, SET TO ASSUME POWER

In Cabinet Hearings, Strong Rejection of Obama's Policies

By MICHAEL D. SHEAR

WASHINGTON — President-elect Donald J. Trump's cabinet nominees, while moderating some of their stances, have made it clear during two weeks of hearings that they intend to work hard

DOW JONES, A NEWS CORP COMPANY ▼

DJIA ▲ 17873.22 0.25% Nasdaq ▲ 4933.50 0.65% U.S. 10 Yr ▼ 0/32 Yield 1.851% Crude Oil ▲ 49.60 0.55% Euro ▲ 1.1139 0.22%

THE WALL STREET JOURNAL.

Subscribe Now | Sign In

SPECIAL OFFER: JOIN NOW

Home World U.S. Politics Economy Business Tech Markets Opinion Arts Life Real Estate

New U.S. Study Fans Cellphone Cancer Worries

Samsung Adds More Ads to Its TVs

Verizon Workers Win Concessions in Deal to End Strike

TECH

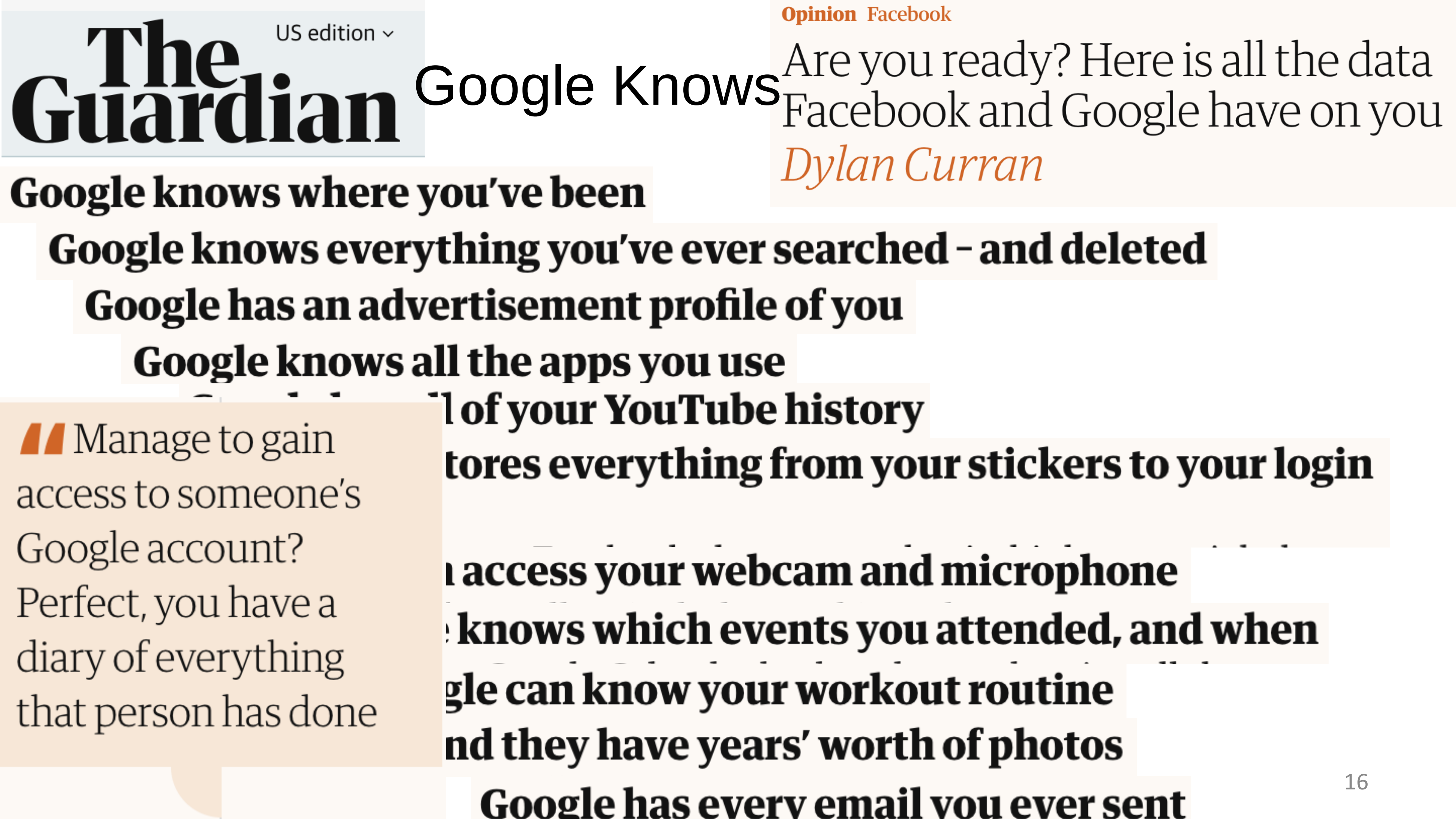
LinkedIn 2012 Data Breach May Have Hit Over 100 Million

Professional social network says it will invalidate passwords that weren't changed since

103

arty Says It Has Thwarted ck of Voter Database

Google Knows



The Guardian

US edition

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

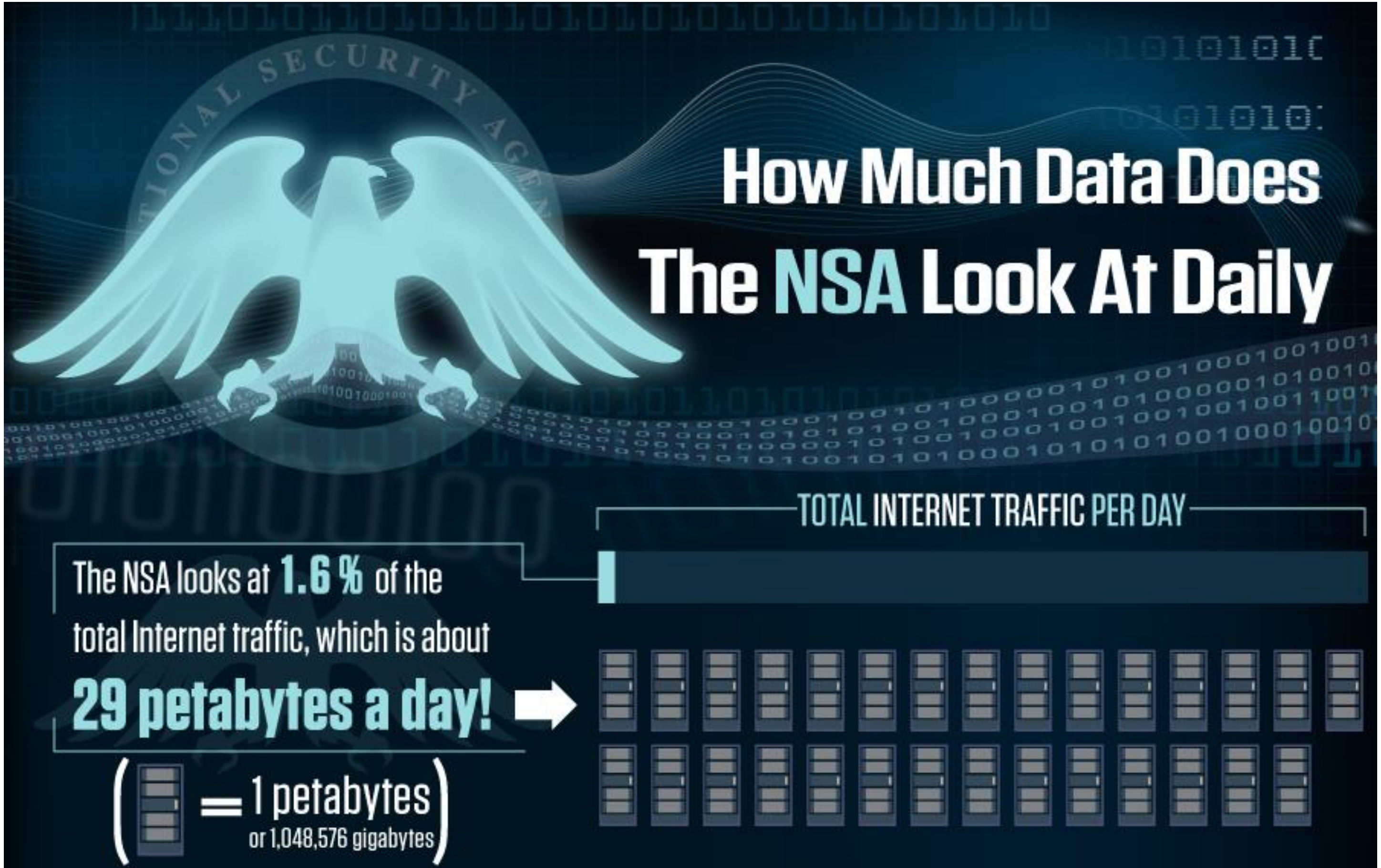
gle can know your workout routine

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

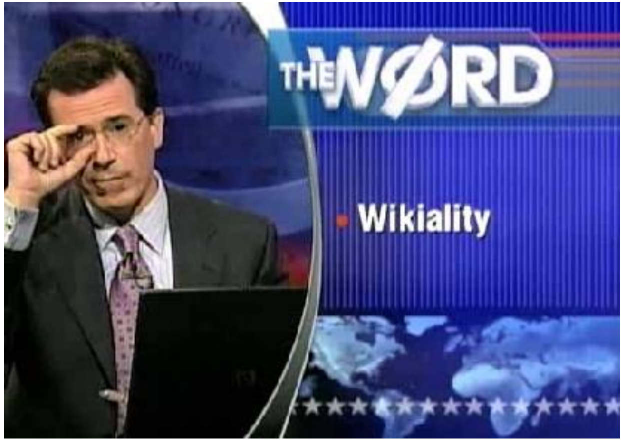


Data Integrity: Not all Data is Correct

"Any user can change any entry, and if enough users agree with them, ***it becomes true.****"*
– Colbert Report 7/31/2007

Asked users to update the page on Elephants to reflect a tripling population, forcing Wikipedia to lock the page.

Yet a 2005 Nature study found **Wikipedia science** articles to be ***similar in accuracy*** to



COMEDY CENTRAL VIDEO ARCHIVE VIA WIKIPEDIA

http://www.cc.com/video-clips/z1aahs/the-colbert-report-the-word---wikiality

https://en.wikipedia.org/wiki/Reliability_of_Wikipedia

http://www.nature.com/nature/journal/v438/n7070/full/438900a.html

A Syllogism of Quotes

*“**information** is **knowledge**”*

— Albert Einstein

*“**knowledge** is **power**”*

— Sir Francis Bacon

*“with great **power** comes great **responsibility**”*

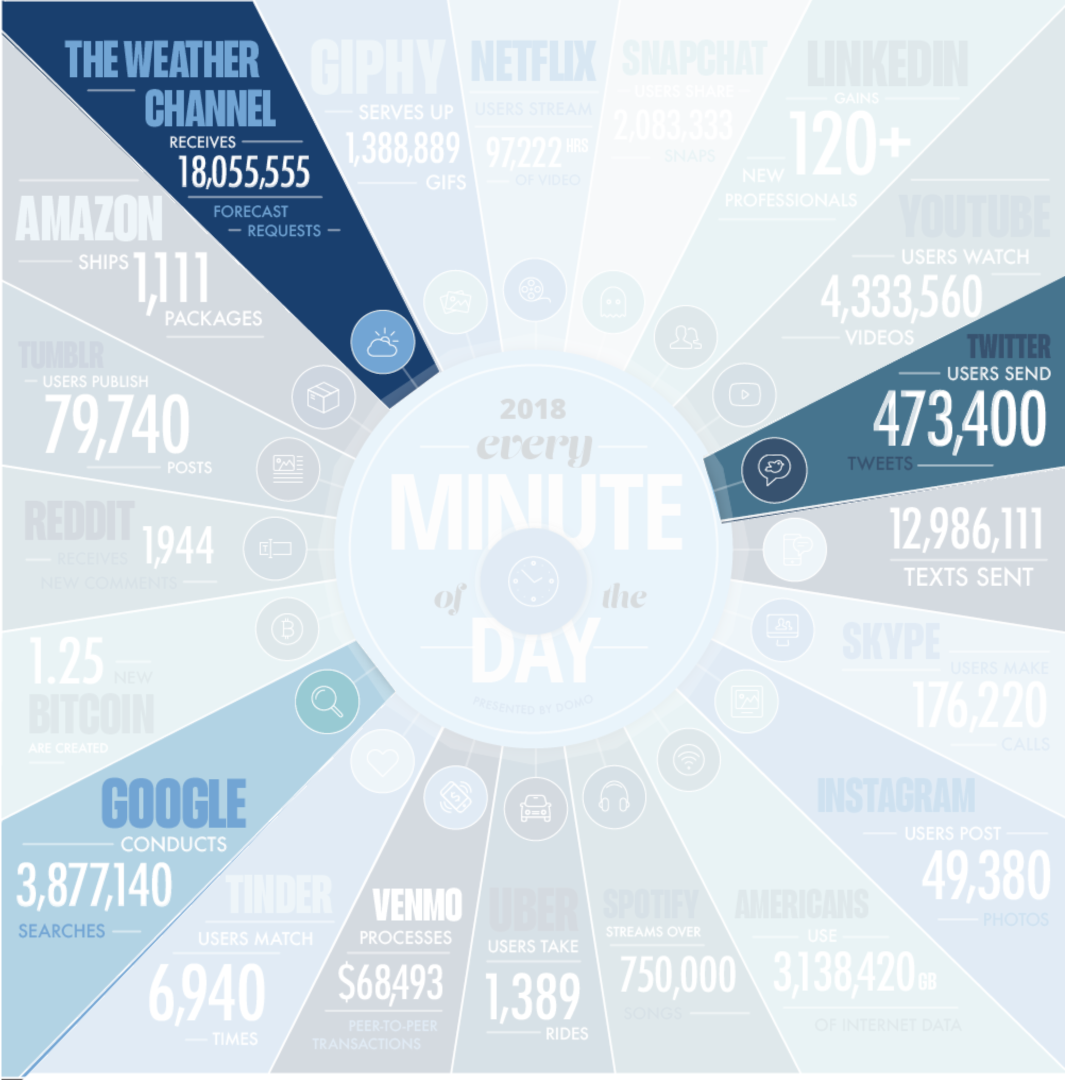
— Uncle Ben (Spiderman)

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

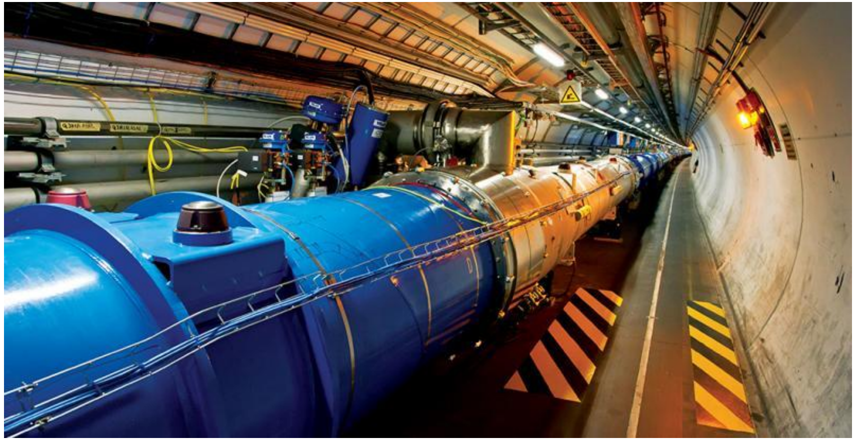
- Data growth will continue to outpace computation
- Systems for Data at Scale: the core of modern computing

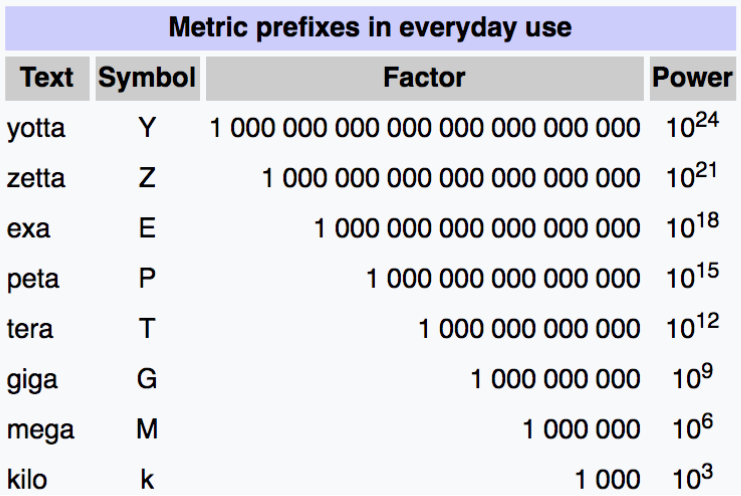


Every Minute!

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

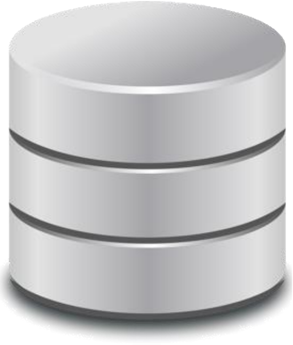
- **Why** take this class?
- **What** is this class all about?
- **Who** is running this?
- **How** will this class work?

What is this class all about?

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

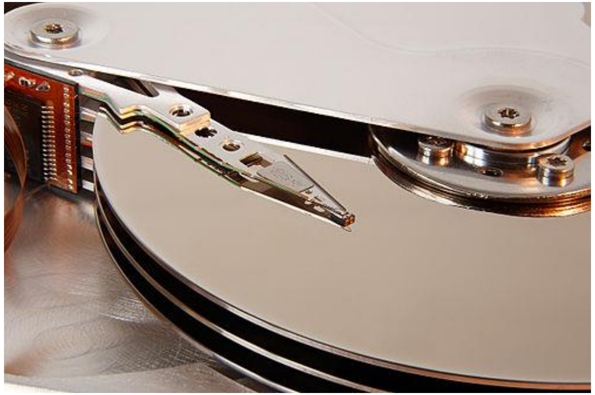
Universal Symbol for a Database

Why the Symbol?



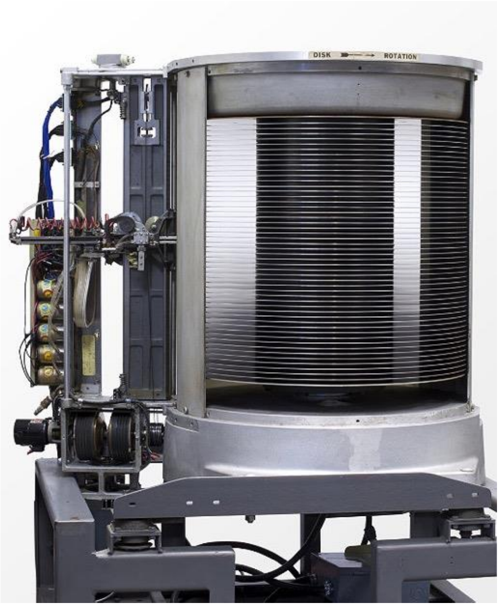
Looks Like?

Platters on a Disk Drive



Why the Symbol?, cont

Looks Like?



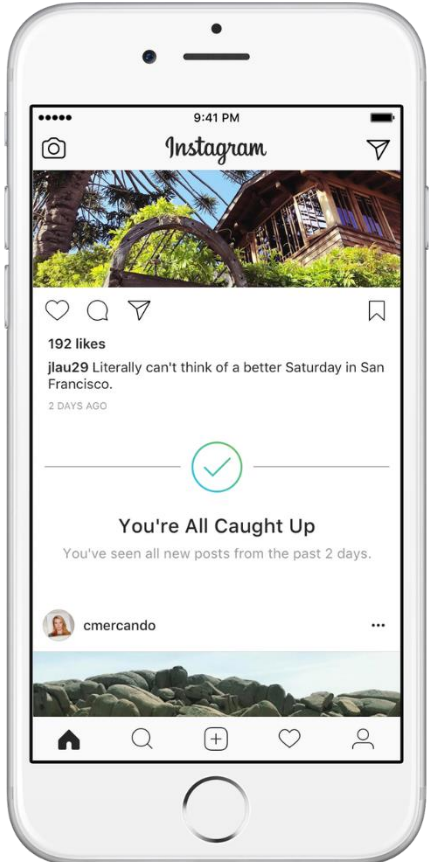
1956: IBM MODEL
350 RAMAC
First Commercial
Disk Drive
5MB @ 1 ton

http://www.computerhistory.org/storageengine/first-commercial-hard-disk-drive-shipped

Is This a Database?

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

Is This a Database?, cont



192 likes

jlau29 Literally can't think of a better Saturday in San Francisco.

2 DAYS AGO

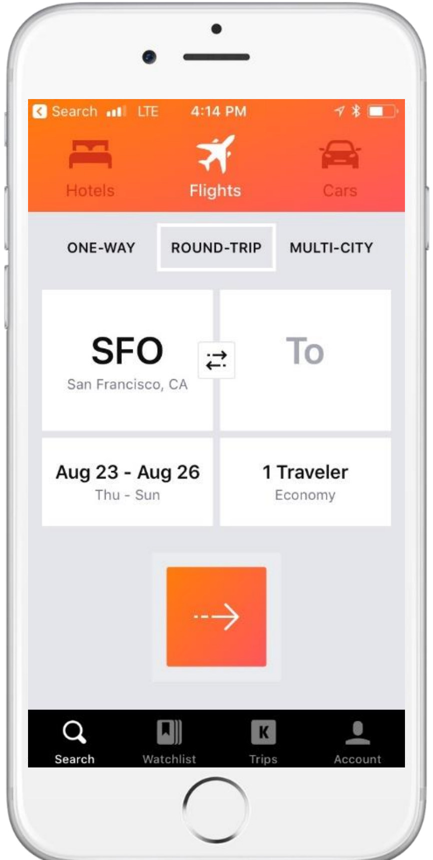
You're All Caught Up

You've seen all new posts from the past 2 days.

cmercando

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

Is This a Database? Part 3



SFO
San Francisco, CA

Aug 23 - Aug 26
Thu - Sun

1 Traveler
Economy

- Airline reservation systems were one of the earliest pervasive consumer uses of database systems.
  - IBM/American Airlines’ SABRE system, 1964.
    - “Semi-Automated Business Research Environment”
  - Travelocity.com a direct descendant of SABRE
    - Acquired by Expedia, 1/2015

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.*

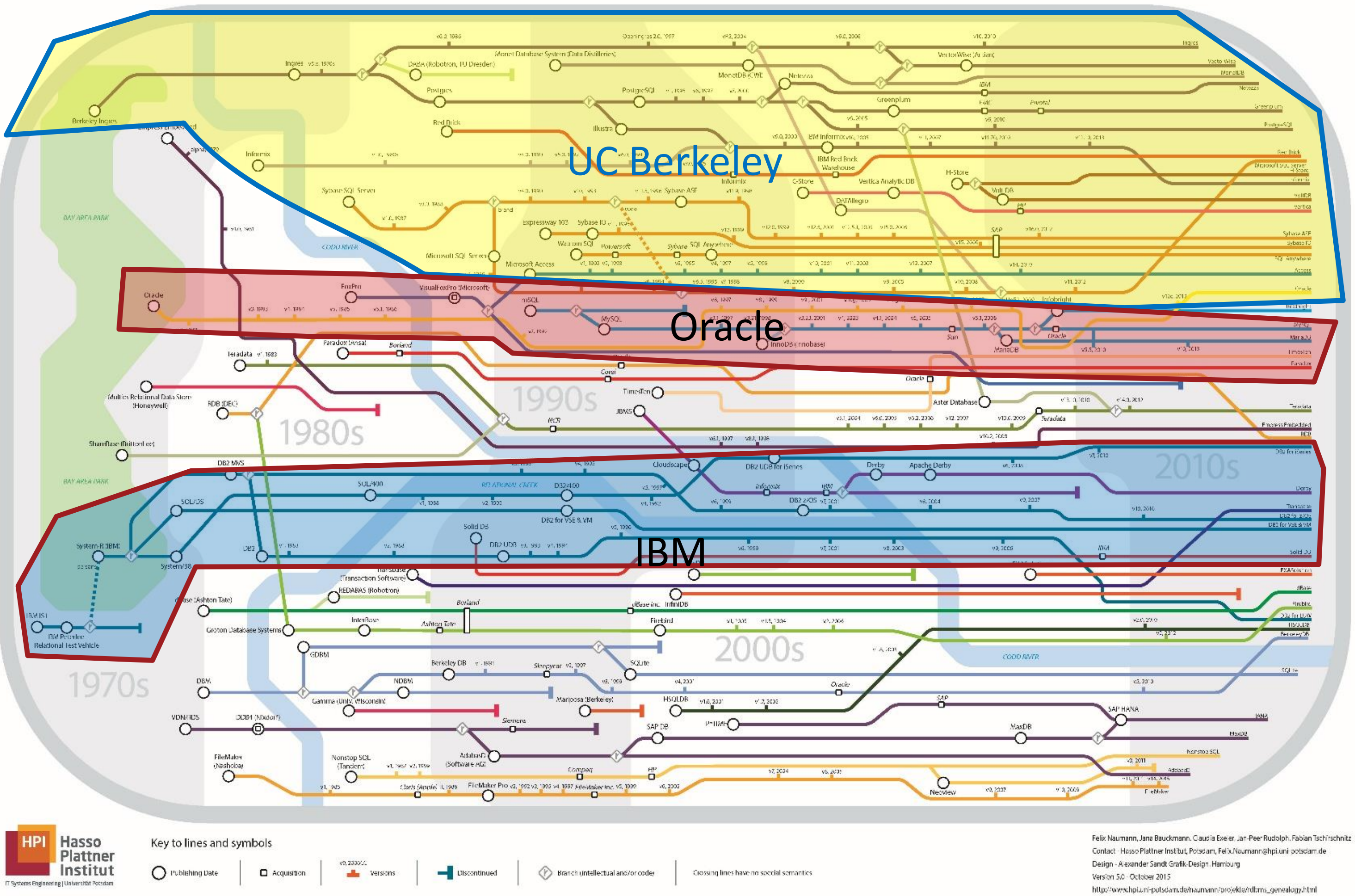
Relational DBMSs

- Traditionally DBMS referred to relational databases

ORACLE TERADATA Microsoft SQL Server PostgreSQL SQLite

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

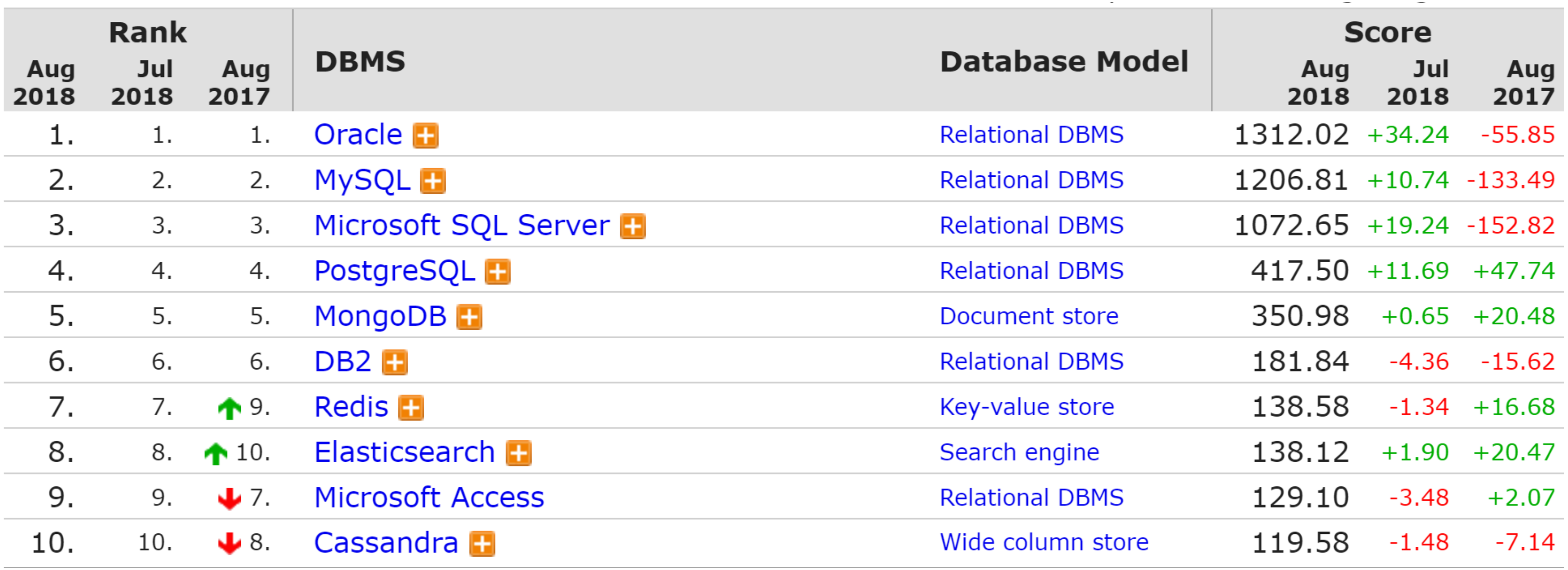
Genealogy of Relational Database Management Systems


UC Berkeley
Oracle
IBM
1970s
1980s
1990s
2000s
2010s
Berkeley Ingres
Ingres
Postgres
PostgreSQL
Informix
Sybase SQL Server
Microsoft SQL Server
Oracle
MySQL
MariaDB
System-R (IBM)
DB2
SQL/DS
Teradata
HPI Hasso Plattner Institut
Key to lines and symbols
Publishing Date
Acquisition
Versions
Discontinued
Branch (intellectual and/or code)
Crossing lines have no special semantics
Felix Naumann, Jana Bauckmann, Claudia Exeler, Jan-Peer Rudolph, Fabian Tschirschnitz
Contact - Hasso Plattner Institut, Potsdam, felix.naumann@hpi.uni-potsdam.de
Design - Alexander Sandt Grafik-Design, Hamburg
Version 5.0 - October 2015
http://www.hpi.uni-potsdam.de/naumann/projekte/rdbms_genealogy.html

Berkeley Roots!

- Ingres / Postgres
- Sybase
- Informix

Ranking of DBMS Technologies 2018

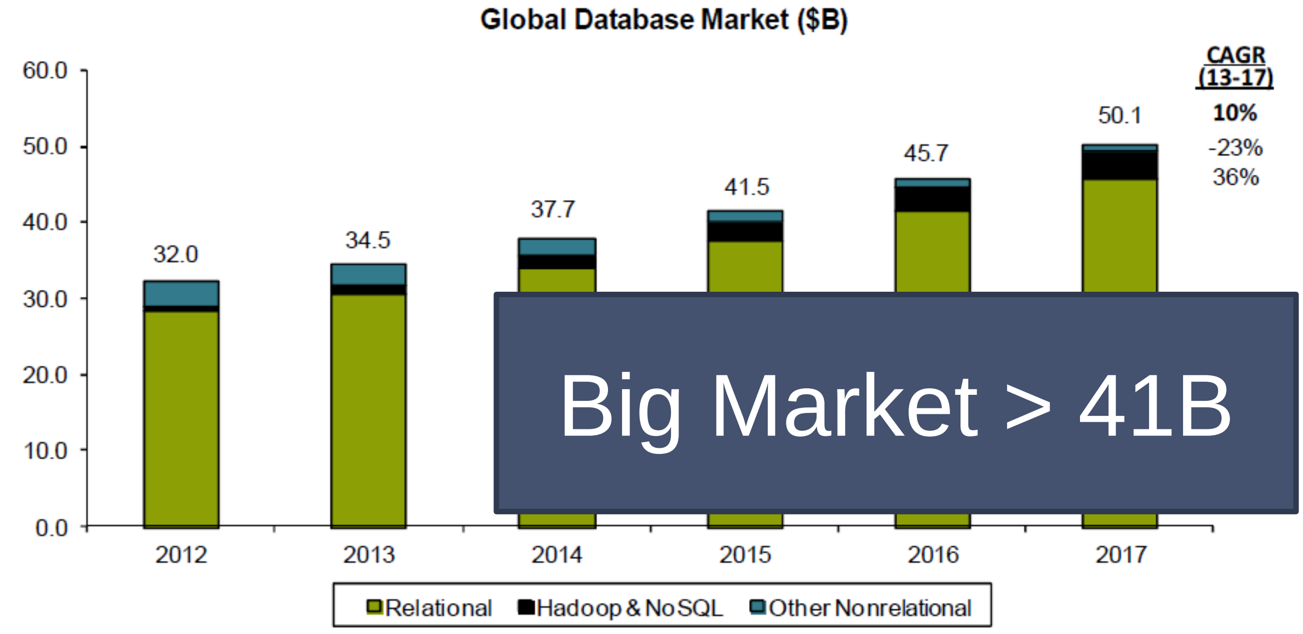


Rank			DBMS	Database Model	Score		
Aug 2018	Jul 2018	Aug 2017			Aug 2018	Jul 2018	Aug 2017
1.	1.	1.	Oracle	Relational DBMS	1312.02	+34.24	-55.85
2.	2.	2.	MySQL	Relational DBMS	1206.81	+10.74	-133.49
3.	3.	3.	Microsoft SQL Server	Relational DBMS	1072.65	+19.24	-152.82
4.	4.	4.	PostgreSQL	Relational DBMS	417.50	+11.69	+47.74
5.	5.	5.	MongoDB	Document store	350.98	+0.65	+20.48
6.	6.	6.	DB2	Relational DBMS	181.84	-4.36	-15.62
7.	7.	↑ 9.	Redis	Key-value store	138.58	-1.34	+16.68
8.	8.	↑ 10.	Elasticsearch	Search engine	138.12	+1.90	+20.47
9.	9.	↓ 7.	Microsoft Access	Relational DBMS	129.10	-3.48	+2.07
10.	10.	↓ 8.	Cassandra	Wide column store	119.58	-1.48	-7.14

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

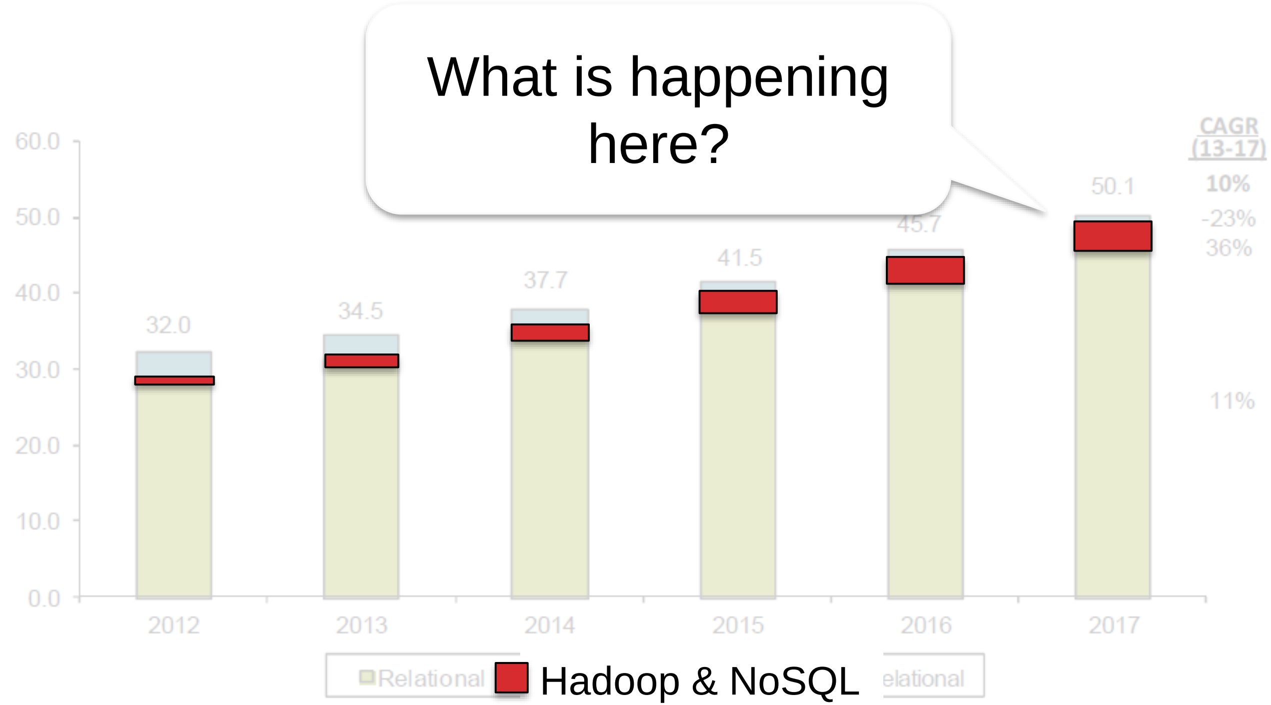


Big Market > 41B

Source: IDC, Bernstein analysis

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

Relational Database Market, cont



What is happening here?

Source: IDC, Bernstein analysis

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...

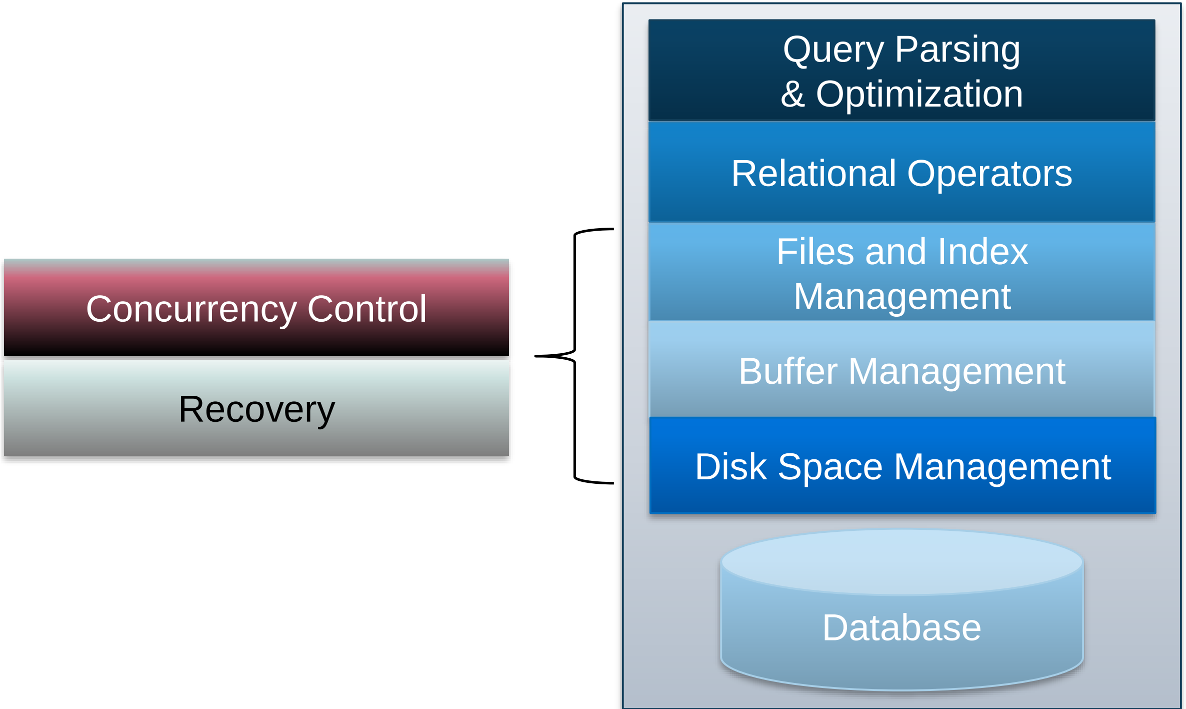
- Data Oriented Programming with SQL
- Foundations of Database System Design
    - Storage, indexing
    - Query processing and optimization
- Transactions
    - Concurrency, Consistency, Recovery
- Data Modeling
    - Application-level representations of data

Principles

- Data Independence
- Declarative Programming
- Rendezvous in Time and Space
- Isolation and consistency
- Data representations

Systems

We will examine various levels of a DBMS



What is this class all about?, cont

- Databases?
  - What is a database?
- Database Management Systems?
- Implementation?
- Big Ideas in Database Management Systems
  - Principles and Algorithms
  - System Designs
  - *The heart of scalable CS*

Essential Queries, Pt 3

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

About Me: Lu Sun（孙 露）

- Assistant Professor in SIST
    - Joined in Nov., 2019
    - PhD @ Hokkaido University
    - Postdoc @ Kyoto University
    - Email: sunlu1@shanghaitech.edu.cn

- Teaching Experience
    - CS182—Introduction to Machine Learning
        - 2022 Spring
    - CS150A—Database
        - 2021 Fall
    - SI151—Optimization and Machine Learning
        - 2020 Spring, 2021 Spring
    - CS150—Database and Data Mining
        - 2020 Fall

TAs

- Binbin Chen (陈 彬彬)
  - B4 student in CS
  - A+ in CS150A 2021 Fall
  - chenbb@shanghaitech.edu.cn

- Xingyue Peng (彭 星月)
  - B4 student in CS
  - A+ in CS150A 2021 Fall
  - pengxy@shanghaitech.edu.cn

- Jiahui Xu (徐 嘉慧)
  - M1 student in CS
  - A+ in CS150 2020 Fall
  - xujh2022@shanghaitech.edu.cn

- Ke Bian (卞 珂)
  - M1 student in CS
  - A in DB course of Xiamen University
  - bianke2022@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: Tue. & Thu., 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/class/l78qj3s2ced2az
- 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: 25%
- Quiz: 10%
- Course project: 25%
- Final exam: 40%

Highlights

- Please write your HW, project and exam in English
- Submitted to GradeScope:
  - https://www.gradescope.com/courses/429221 (Entry Code: 7GEBKG)
- For late HW or project, 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