CS w186
Introduction to Database Systems

Prof. Joe Hellerstein

Essential Queries

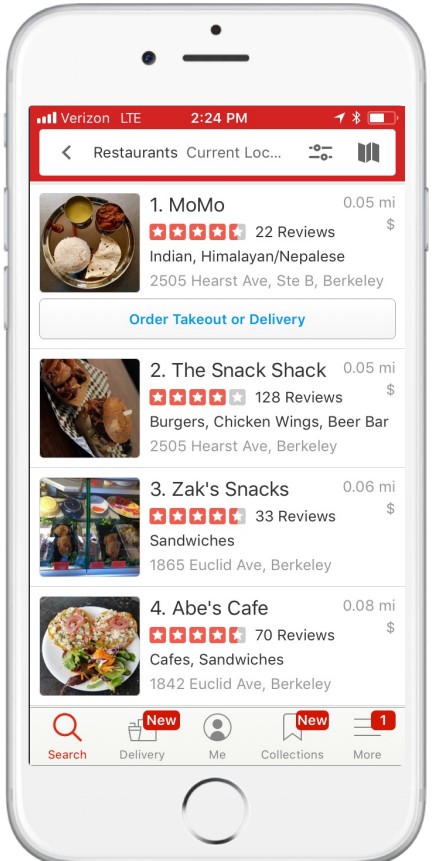
- **Why** take this class?
- **What** is this class all about?
- **Who** is running this?
- **How** will this class work?

WHY?

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?

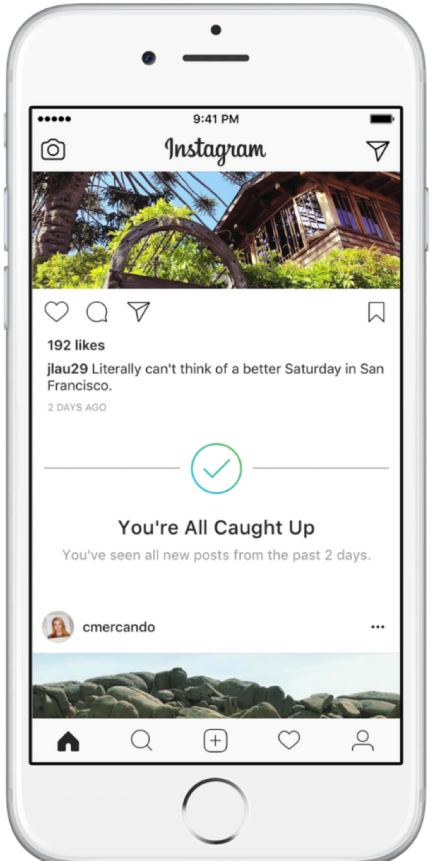


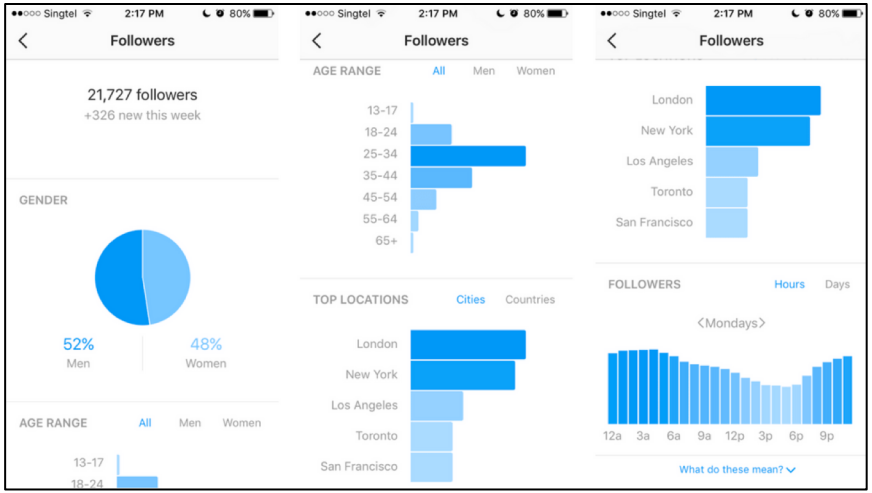
Each ratings star added on a Yelp restaurant review translated to anywhere from a 5% to a 9% effect on revenues.

—Harvard Business School, 2011

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

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-feed/

What am 2

Hey Instagram: that's Chez Panisse in Berkeley, CA!

Berkeley cs186

Who should I be with, Database?

1.6b
SWIPES PER DAY

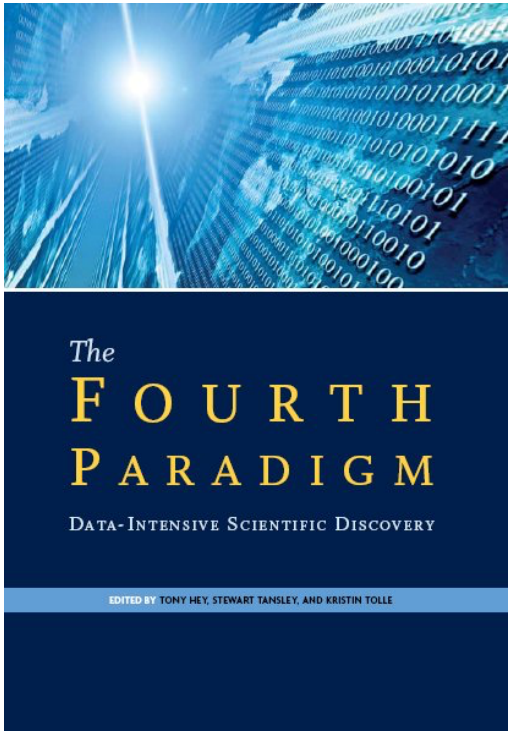
1m
DATES PER WEEK

20b+
TOTAL MATCHES

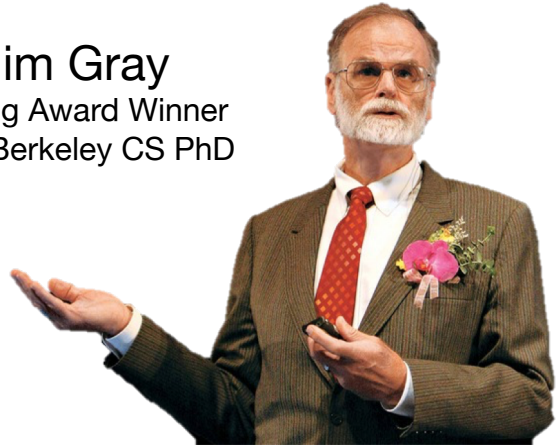
https://www.gotinder.com/press

https://www.awesomeinventions.com/funny-tinder-profiles/

How does Science work? Database.

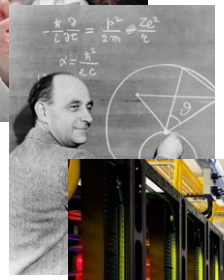


Jim Gray
Turing Award Winner
First Berkeley CS PhD



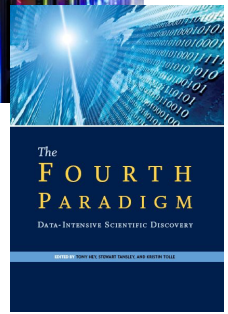
How does Science work? Database. Pt 2

Experimental



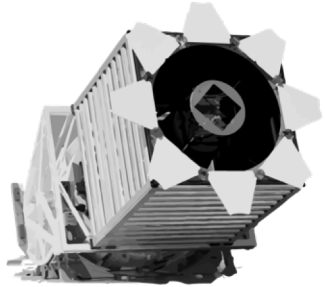
Theoretical

Simulation



Data
Intensive

Astronomy in the 4th Paradigm



Sloan Digital
Sky Survey (SDSS)

+

Database
Systems

Sky Server

Berkeley cs186

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
  - Evolution of IT
- 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

Berkeley cs186

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

Summon:

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

BBC Home News Sport Weather TV Radio

ONE-MINUTE WORLD NEWS

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.

HM Revenue and Customs gives the National Audit Office HMRC's child benefit data, in breach of security. The information is later safely returned.

00 people's details go missing after being sent

The New Yo

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

The Sunday

Facebook was warned of data risks 7 years ago

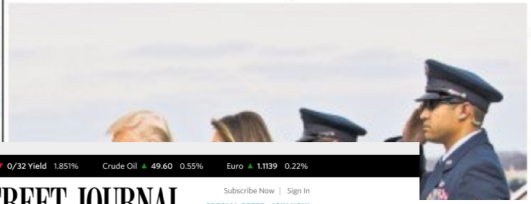
Winning feeling

By James Titcomb said the company had no way of know-

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

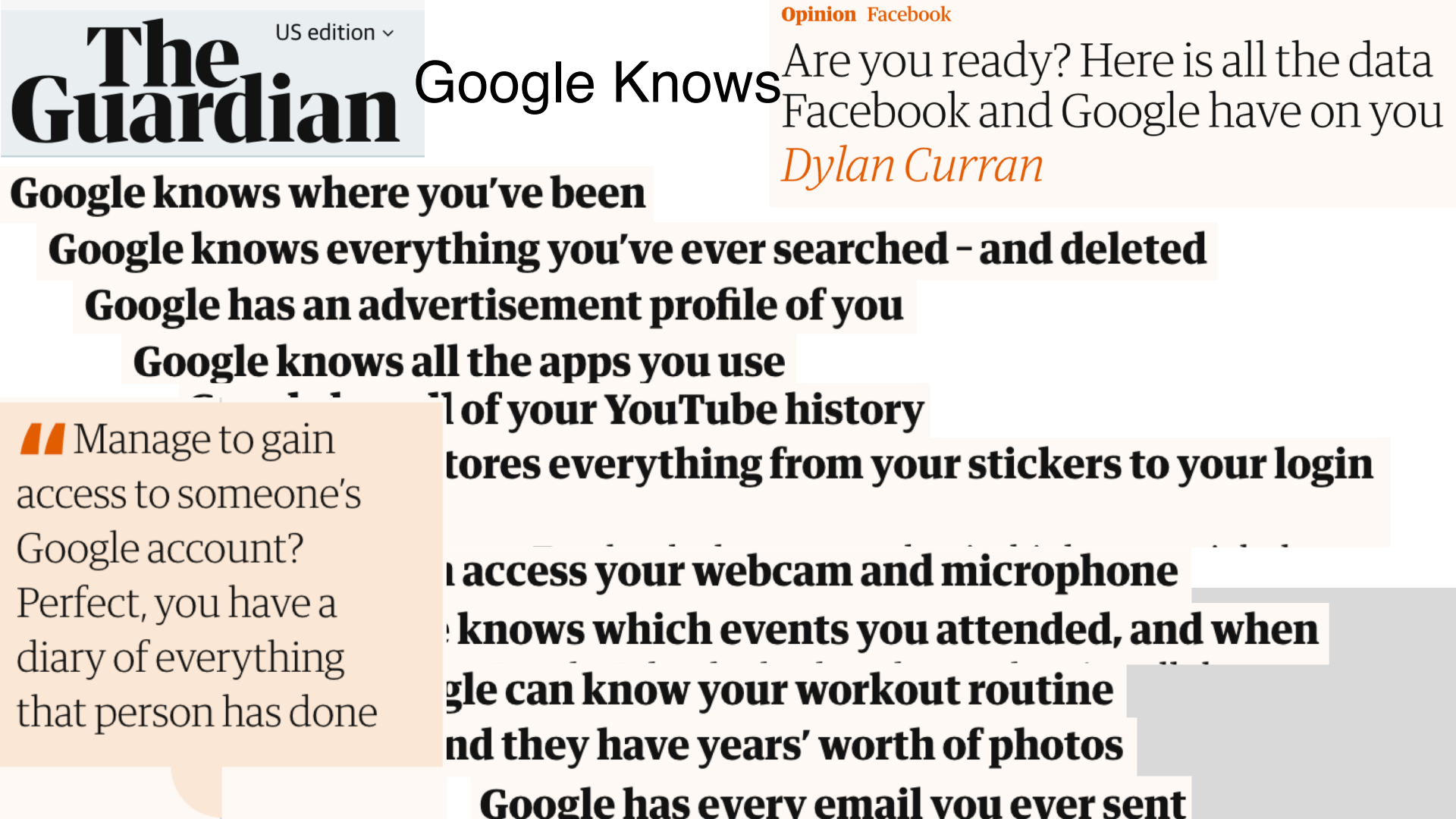
Verizon Workers Win Concessions in Deal to End Strike

…arty Says It Has Thwarted …ck of Voter Database

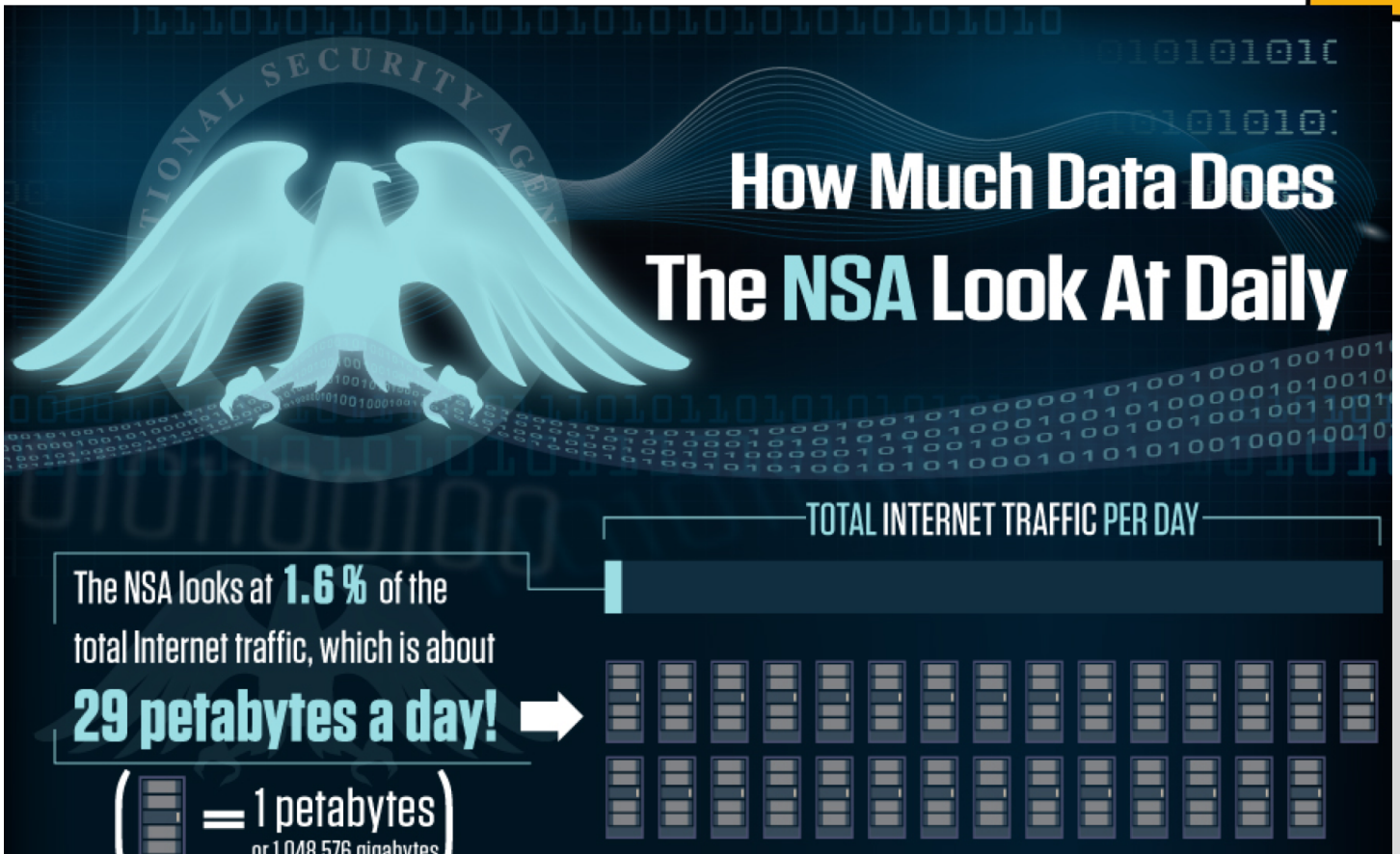
TECH

LinkedIn 2012 Data Breach May Have Hit Over 100 Million

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


The Guardian
US edition
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
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



National Security Data: 2018

POLITICS

The New York Times

GIFT THE TIMES Account

N.S.A. Triples Collection of Data From U.S. Phone Companies

By Charlie Savage

May 4, 2018

WASHINGTON — The National Security Agency vacuumed up more than 534 million records of phone calls and text messages from American telecommunications providers like AT&T and Verizon last year — more than three times what it collected in 2016, a new report revealed on Friday.

In 2016, the first full year for which that replacement system was in operation, the government obtained orders to target 42 people and collected just over 151 million call detail records. In 2017, the government obtained orders for 40 targets. (The orders generate data for 180 days, so some of the 2016 orders kept generating additional data in 2017, and some of the 2017 orders may have been reauthorizations of expiring 2016 orders pegged to the same targets.)

National Security Data: Yesterday

TECHNOLOGY

The New York Times

READ THE GUIDE Account

Facebook Identifies New Influence Operations Spanning Globe

By Sheera Frenkel and Nicholas Fandos

Aug. 21, 2018

136

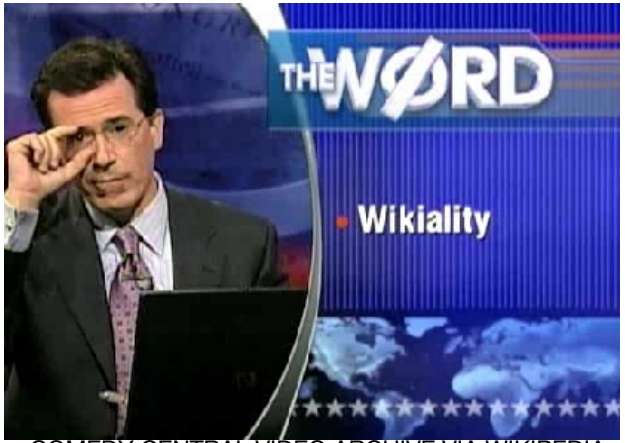
SAN FRANCISCO — Facebook said on Tuesday that it had identified multiple new influence campaigns that were aimed at misleading people around the world, with the company finding and removing 652 fake accounts, pages and groups that were trying to sow misinformation.

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.



COMEDY CENTRAL VIDEO ARCHIVE VIA WIKIPEDIA

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

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

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

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

And then came Fake News

01/03/2018

A Syllogism of Quotes

*"**information** is **knowledge**"*

— Albert Einstein

*"**knowledge** is **power**"*

— Sir Francis Bacon

*"with great **power** comes great **responsibility**"*

— Uncle Ben (Spiderman)

Ethics

Medium

dj patil Follow

Making tech and data work for you

Feb 1 · 4 min read

A Code of Ethics for Data Science

"I could go on and on about all of the amazing work that is happening around the world using data to make lives better everyday, but we also have to address where data is causing more harm than good."

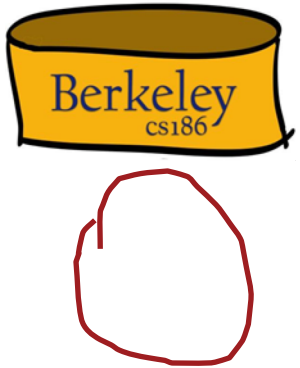
"Data is such an incredible lever arm for change, we need to make sure that the change that is coming, is the one we **all** want to see.

So how do we do it? First, there is no single voice that determines these choices. This **MUST** be community effort."

https://medium.com/@dpatil/a-code-of-ethics-for-data-science-cda27d1fac1

https://www.oreilly.com/ideas/doing-good-data-science

Berkeley's New Data Science Major



https://data.berkeley.edu/degrees/data-science-ba

Berkeley Division of Data Sciences

Home About News Degree Options Education Research Connect

Home » Degree Options » L&S Data Science Major

L&S Data Science Major

Objectives

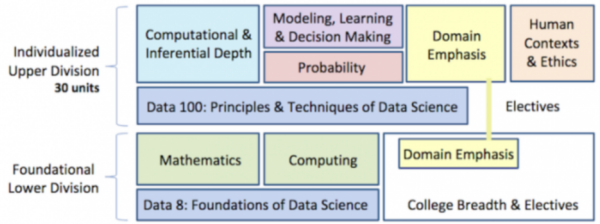
Data Science combines computational and inferential reasoning to draw conclusions based on data about some aspect of the real world. Data scientists come from all walks of life, all areas of study, and all backgrounds. They share an appreciation for the practical use of mathematical and scientific thinking and the power of computing to understand and solve problems for business, research, and societal impact.

The Data Science Major will equip students to draw sound conclusions from data in context, using knowledge of statistical inference, computational processes, data management strategies, domain knowledge, and theory. Students will learn to carry out analyses of data through the full cycle of the investigative process in scientific and practical contexts. Students will gain understanding of the human and ethical implications of data analysis and integrate that knowledge in designing and carrying out their work.

Description of the Undergraduate Major

The L&S undergraduate Data Science major requirements include one core lower-division (Data 8) and upper-division (Data 100) course, along with required courses from each of the following groups

- Foundations in Mathematics and Computing
- Computational and Inferential Depth
- Modeling, Learning and Decision Making
- Probability
- Domain Emphasis
- Human Contexts and Ethics



Why? Reason #2: Centrality (again)

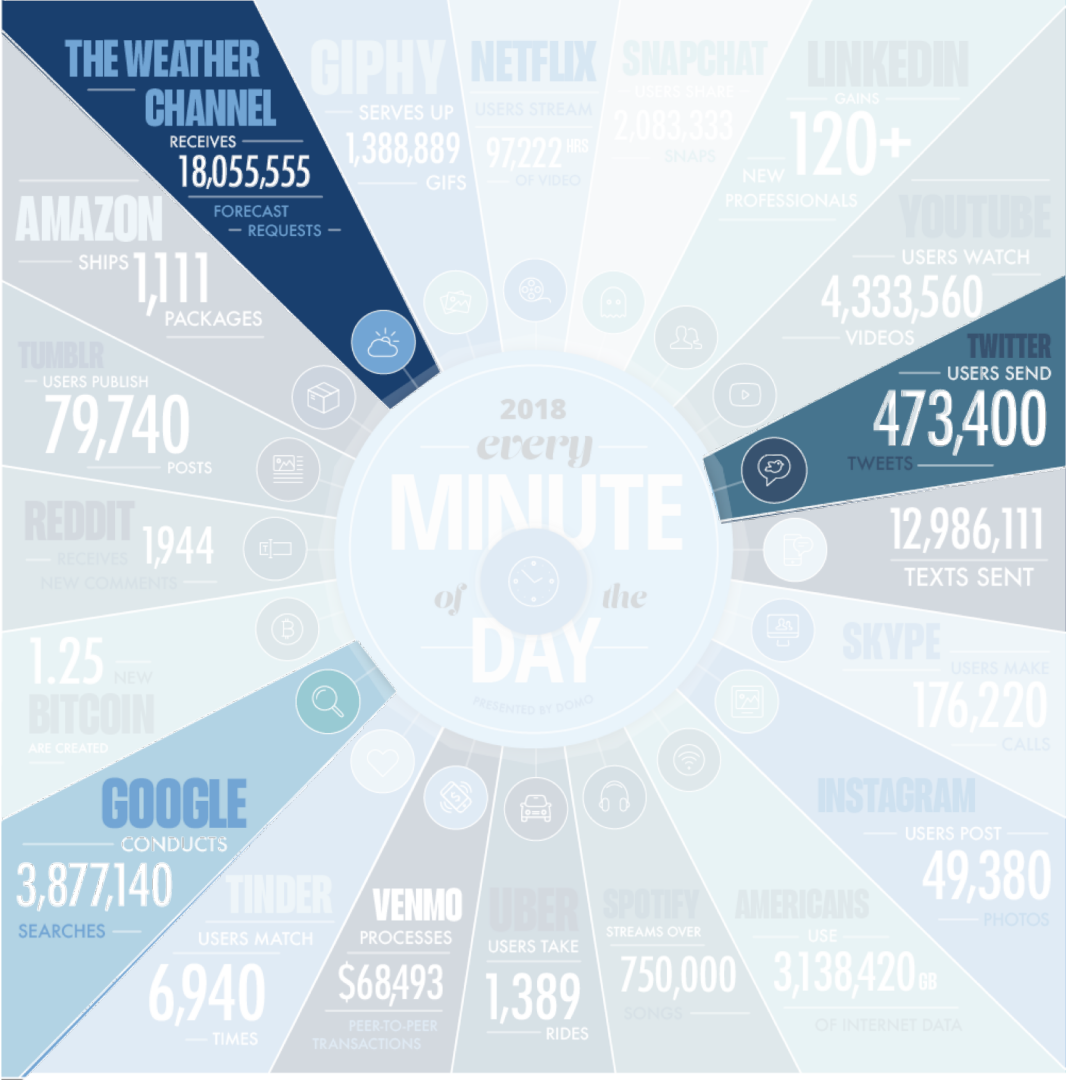
- Data is at the center of modern society.
- Unprecedented in its nature and significance
  - *Particular* and *voluminous*
  - Often asymmetric
    - low value in isolation, high value when aggregated
  - Difficult to protect
- The infrastructure determines what's possible

Why #3? The Core of Computing

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

Data Will Continue to Grow

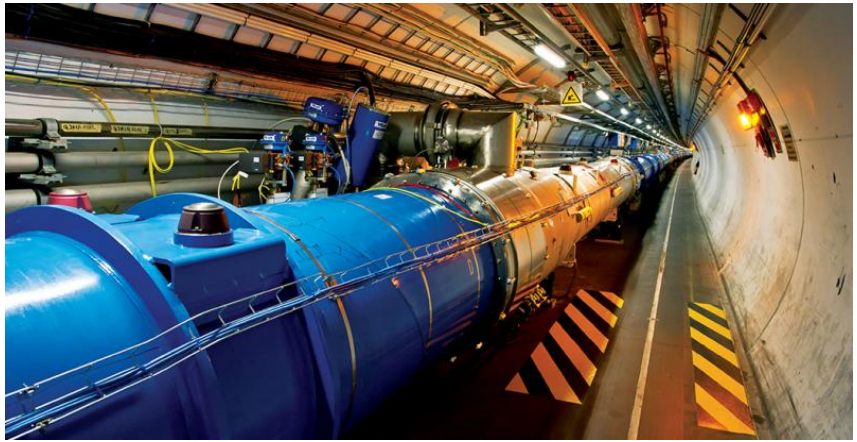
- And outpace computing power



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

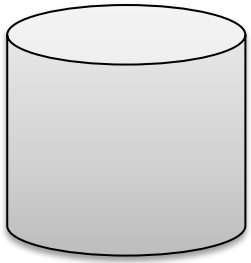
Essential Queries, Pt 2

- **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?

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

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

*"...We must immediately...**attack accounting problems** under the philosophy of handling each business **transaction as it occurs**, rather than under the present condition of **batching techniques**...."*

-- F. J. Wesley IBM Senior Manager

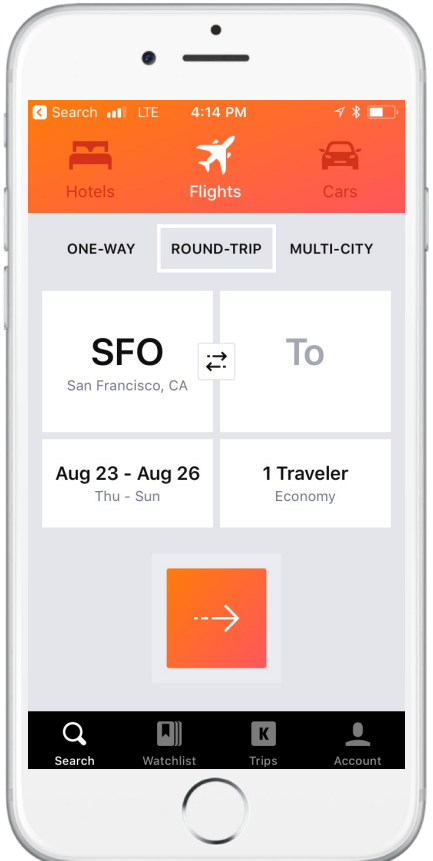
Is This a Database?

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

Is This a Database?, cont

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

Is This a Database? Part 3



- 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?

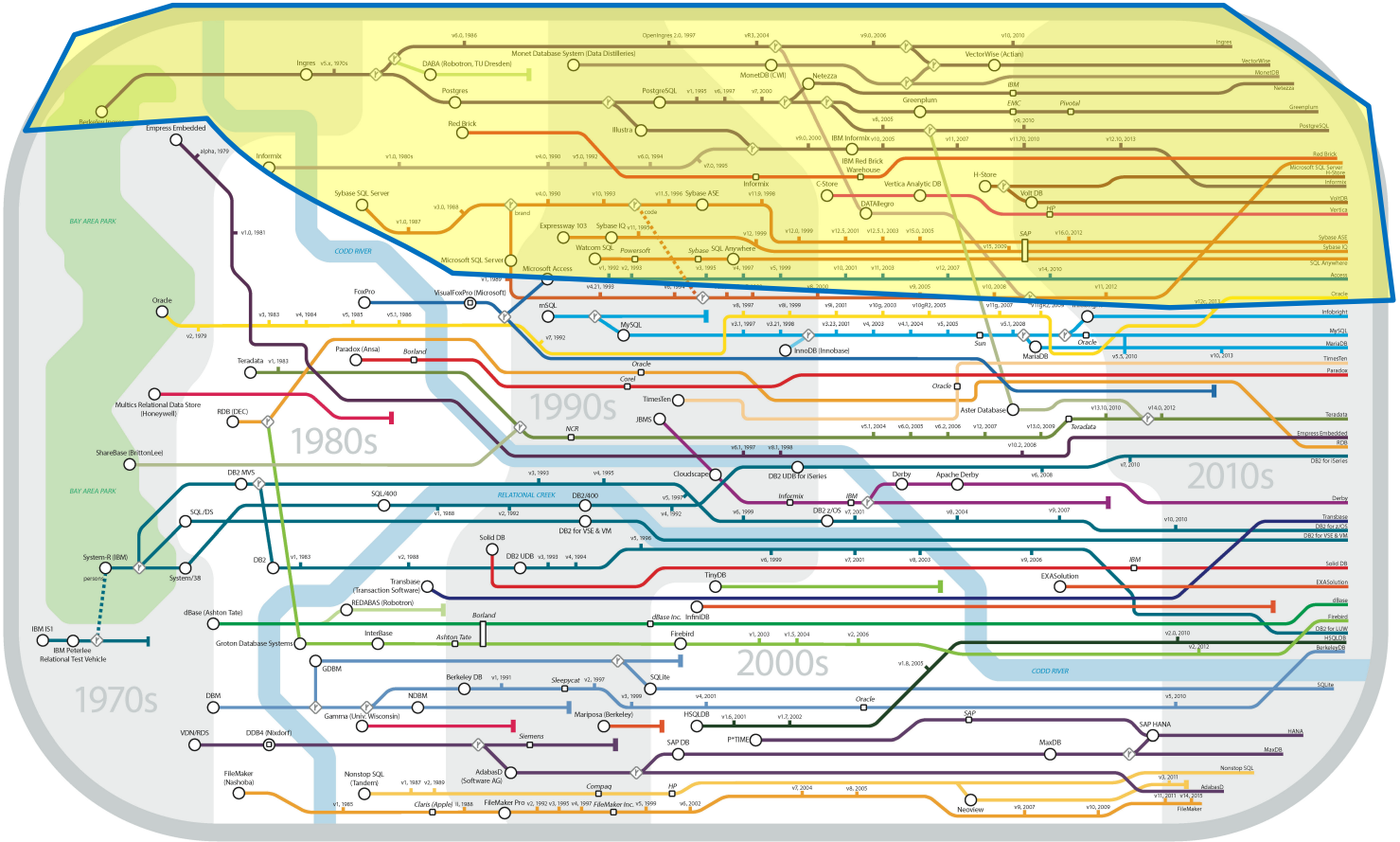
- Let's not split hairs.
  - *A database is a large, organized collection of data.*
- Sometimes confused with a Database Management System (DBMS)
  - *A DBMS is software that **stores, manages,** and facilitates **access** to data.*

Relational DBMSs

- Traditionally DBMS referred to relational databases

- **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

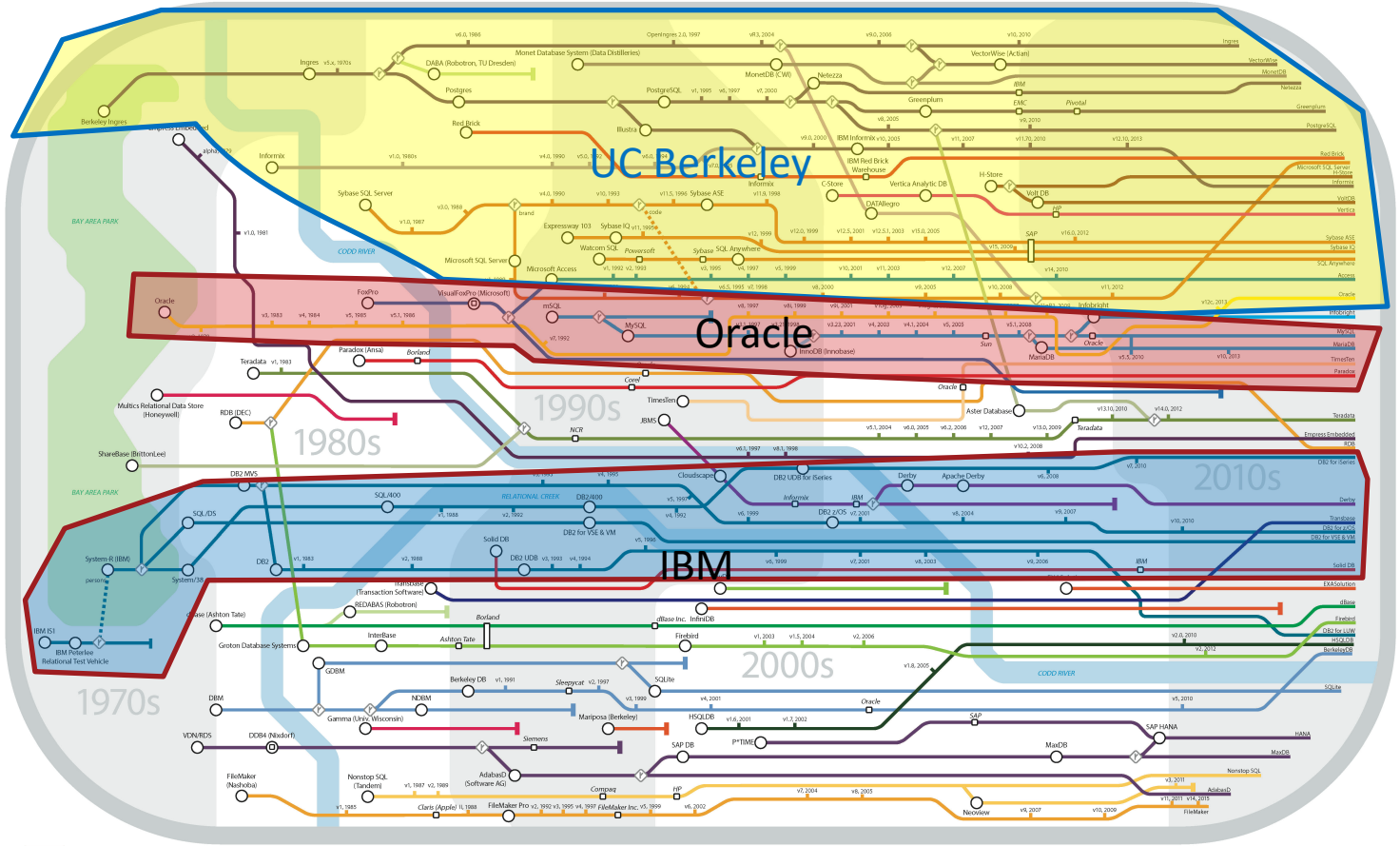
Key to lines and symbols

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

Genealogy of Relational Database Management Systems



Berkeley Roots!

- Ingres / Postgres
- Sybase
- Informix

Key to lines and symbols

Publishing Date | Acquisition | Versions | Discontinued | Branch (intellectual and/or code) | Crossing lines have no special semantics

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

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

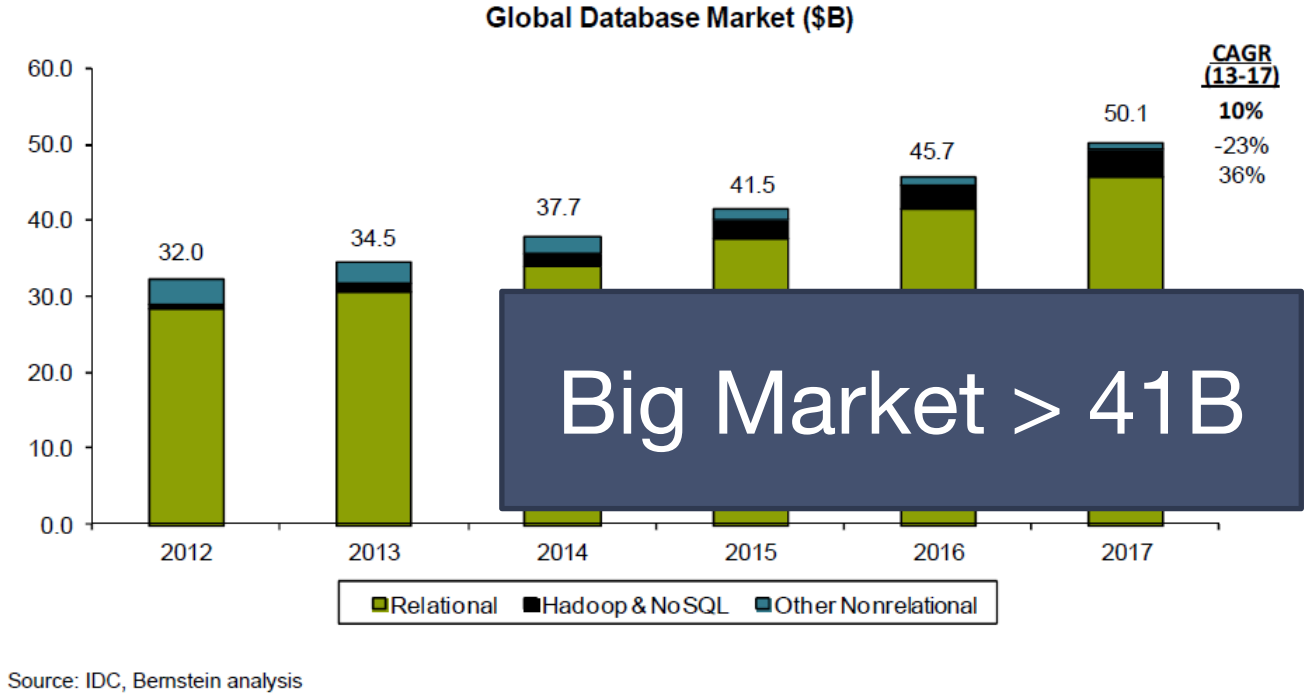
Ranking of DBMS Technologies 2018

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

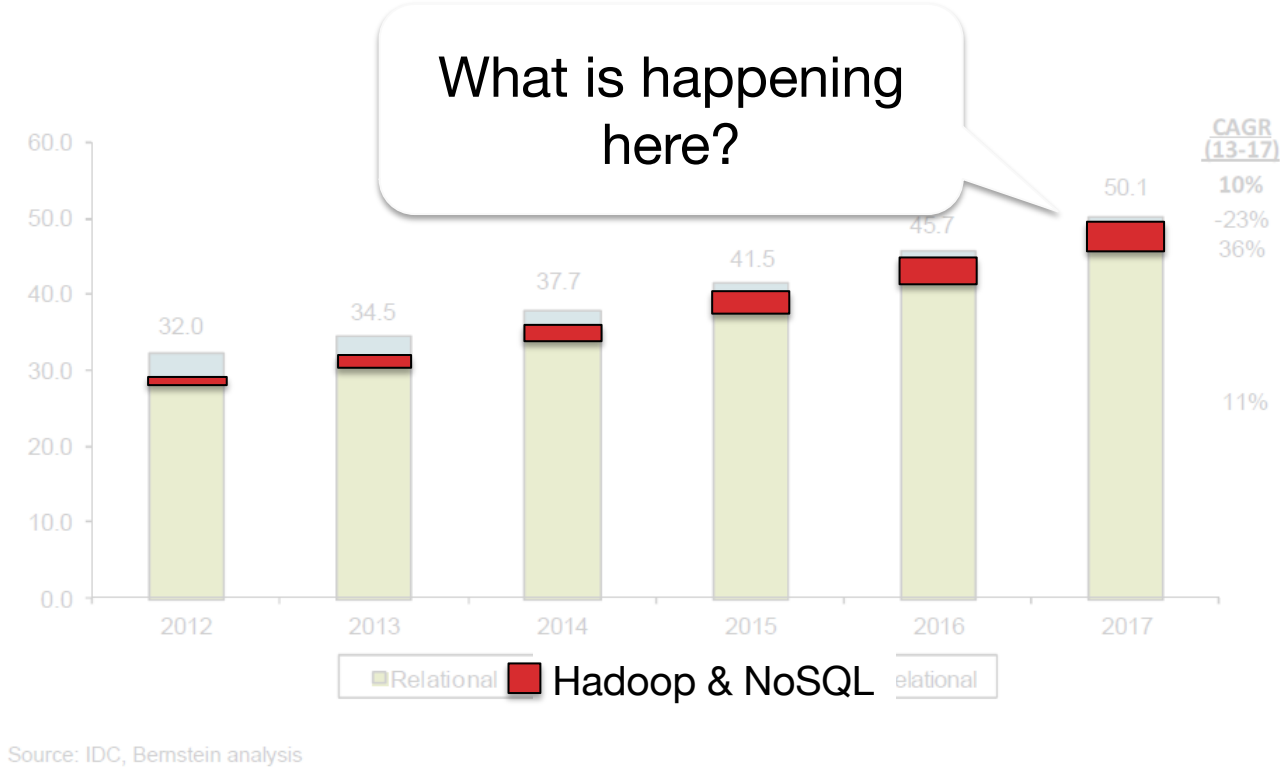
Relational Database Market



Source: IDC, Bernstein analysis

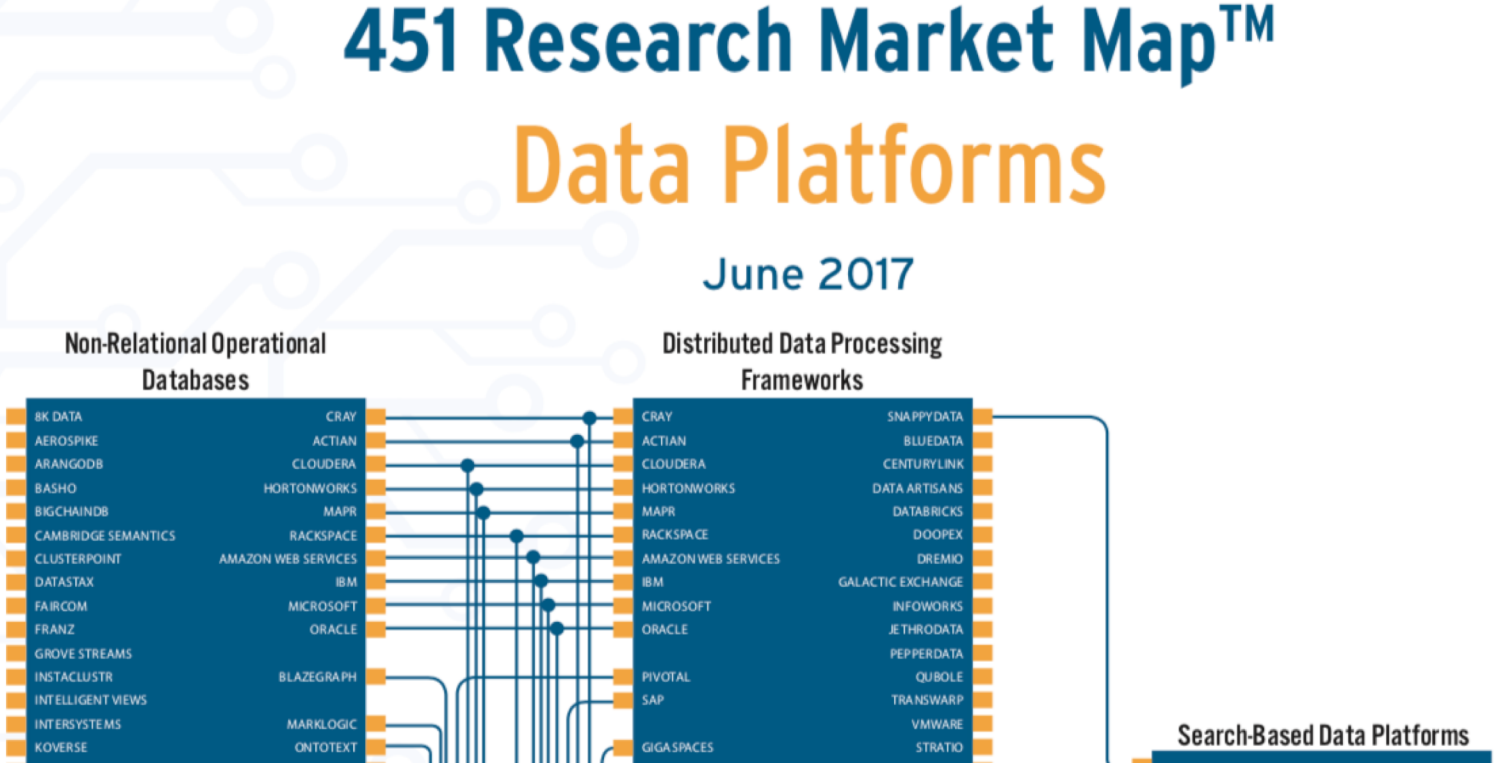
http://www.infoworld.com/article/2916057/open-source-software/open-source-threatens-to-eat-the-database-market.html

Relational Database Market, cont



Source: IDC, Bernstein analysis

Research Market Gap



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 learn the textbook.

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...

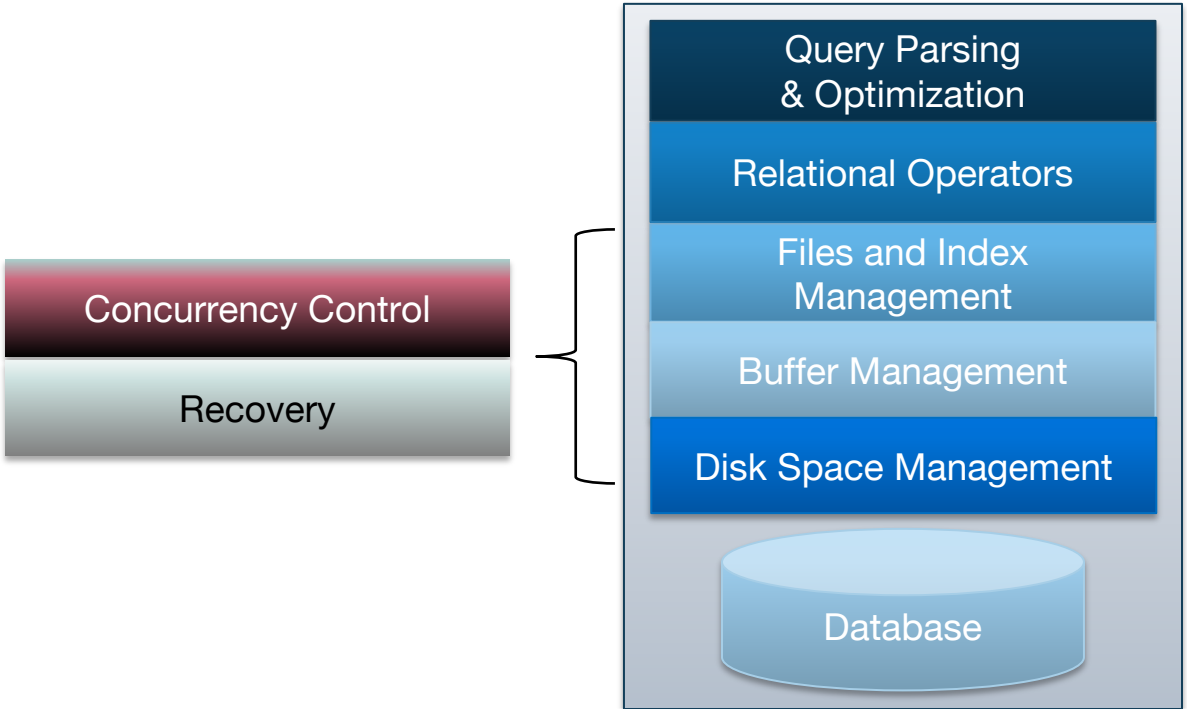
- Data Oriented Programming with SQL
- Foundations of Data System Design
  - Storage, indexing
  - Query processing and optimization
- Transactions
  - Concurrency, Consistency, Recovery
- Data Modeling
  - Application-level representations of data

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*

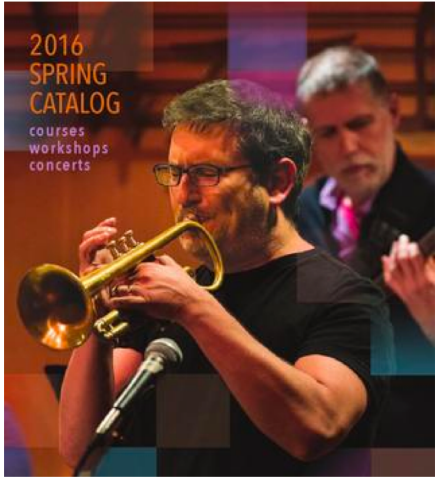
Essential Queries, Pt 3

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

Who Am I?

- 23 years CS faculty at UC Berkeley
  - Recently: RISElab, Data Science Governance Committee
- Co-Founder and Chief Strategy Officer, Trifacta
  - Founded 2012 based on research at Berkeley and Stanford
  - Leading vendor of Data Preparation (a.k.a Data Wrangling) softwar
  - Got messy data? Try it out!
    - https://www.trifacta.com/start-wrangling/
    - https://cloud.google.com/dataprep/
- Advisor to tech companies
  - Notably Greenplum, acquired by EMC
  - Also SurveyMonkey, DellEMC, various startups
- Editor-in-Chief Foundations and Trends in Databases

Who Am I? Outside of work



- Midwest raised
- Father of 2 teenagers
- Occasional musician
  - Fortunate to have performed on stage behind many jazz luminaries
  - Eclipsed by class/bandmates like Berkeley's own Josh Redman and Vijay Iyer
  - Brought jazz onto the internet
- Occasional wedding officiant

"Laziness in doing stupid things can be a great virtue"
— James Hilton

Your Amazing TAs

Benjamin Kha
Brian DeLeonardis
Daphne Nhuch
Eric Sheng
Jamie Gu
Jason Dai
Kimberly Ko
Kimberly Zai
Lakshya Jain
Yawen Sun

You!

- This class is in your hands.
- Everything is doable, with steady work.
- We will help pace you
  - Weekly section worksheets, vitamins keep you on schedule
  - Weekly in-person sections and office hours
- But now more than ever, success is in your hands

Essential Queries, Part 4

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

What is different about CS w186?

- Main difference: online lecture content
  - 3-5 minute videos, followed by online exercises
  - No classes after today
- Otherwise, things are largely standard
  - Discussion sections
    - Worksheet will be online in advance
  - TA and Prof Office Hours

Main Points of Information

- Course Website: cs186berkeley.net
  - Syllabus
  - Calendar: sections and OH
  - Lecture slides
- Course content is on edX
  - Video lectures corresponding to slides
  - Exercises to reinforce lecture content
  - Vitamins to assess understanding
  - Homeworks: SQL and Java
- Piazza discussion group
- **All this info linked on website.**

Workload

- Weekly “lecture” style work
  - Lecture exercises:
    - Intermixed with lecture videos
  - Vitamins: simple weekly online quizzes
    - You can drop 2
    - Exercises must be completed to submit vitamin.
- 5 programming homeworks (next slide)
- 2 midterm exams: held during “class time” in 1 Pimentel
- 1 final exam

- Schedule for homeworks and exams on the website

Homeworks

- Real-world focus
  - SQL querying: basics and algorithmics
  - Building pieces of a DBMS
    - B+-tree indexes
    - Join Algorithms
    - Dynamic Programming Query Optimizer
    - Concurrency (2PL) and Recovery (ARIES)
- HW1 goes out next week!!

Cheating policy

- Zero Tolerance. It is uncool. Don't.
  - We have the technology to find out.
- I know that most cheating happens due to stress
  - Plan ahead and stay on schedule to minimize stress
  - You have built-in safety valves
    - Dropped vitamins
    - Slip days on homeworks: save for when you **need** them, don't micro-optimize!
    - Midterms weighted to the higher grade
    - Keep an eye on the course drop date. Don't take too many courses!
  - Feeling stressed? Reach out!
    - Campus resources
    - Course staff is here for you
    - Incompletes are appropriate for health issues of any kind
- Staff perspective
  - We want you to learn and to succeed
  - We want things to be fair, so need to stick to rules

Staying in touch

- All class communication via Piazza
  - https://piazza.com/berkeley/fall2018/cs186
  - We will go live with answers today
- Announcements and discussion
  - read it regularly
  - post all questions/comments there
  - answer each other's questions!
- Direct email to Prof or TAs is not a good idea
  - And will likely not get answered unless sensitive

See you online!