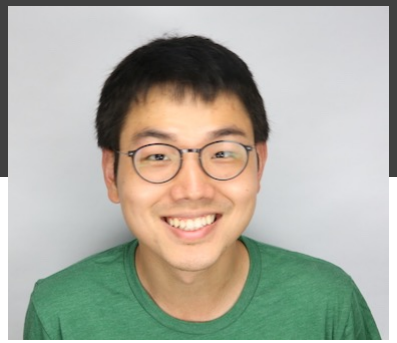
UC San Diego

DSC 102
Systems for Scalable Analytics

Spring 2024

Haojian Jin

About Myself

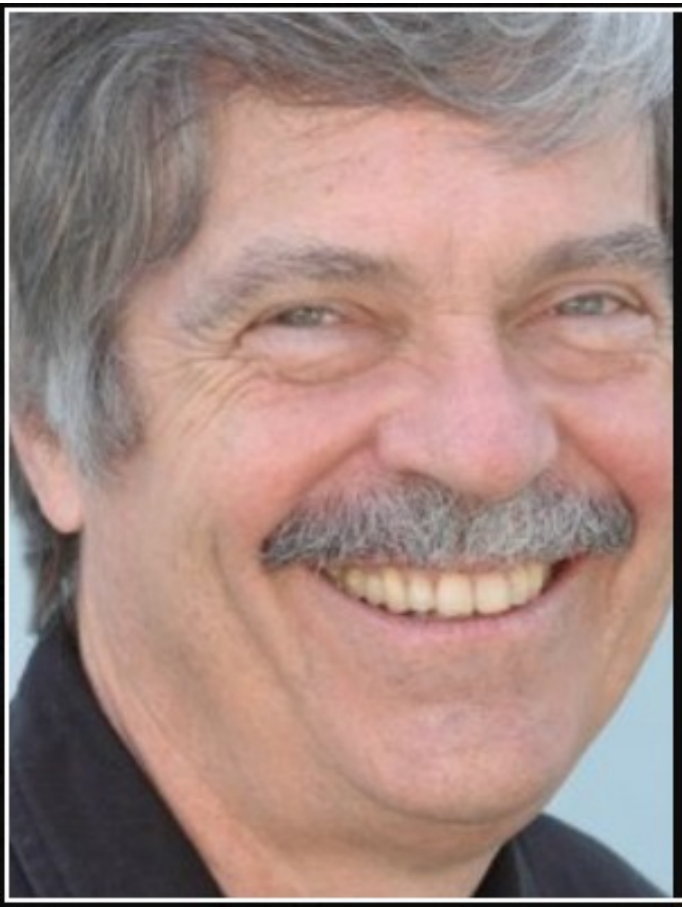


Haojian Jin (http://haojianj.in/)
Asst. Prof @ UCSD-HDSI

Data Smith Lab:
We study the security and privacy of data systems by researching the people who design, implement, and use these systems.

Ph.D. from CMU Human-Computer Interaction Institute
Ph.D. Thesis: Modular Privacy Flow

Before Ph.D.: worked at Yahoo Research, ran a startup
HCI, Software Engineering, Mobile Computing, AI.



The best way to predict the future is to invent it.

— Alan Kay —

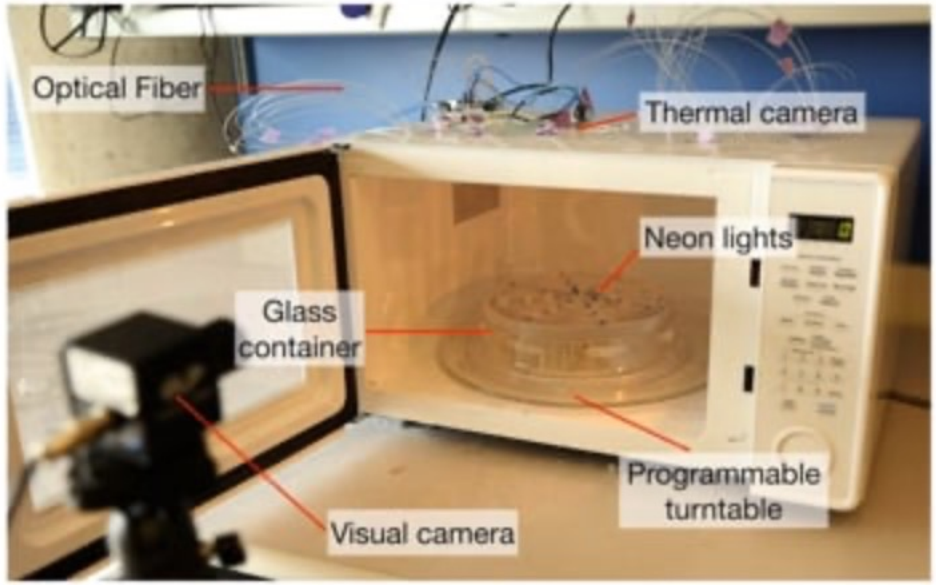
AZ QUOTES

Working Code Trumps All Hype!

Software Defined Cooking (SDC) using a microwave oven

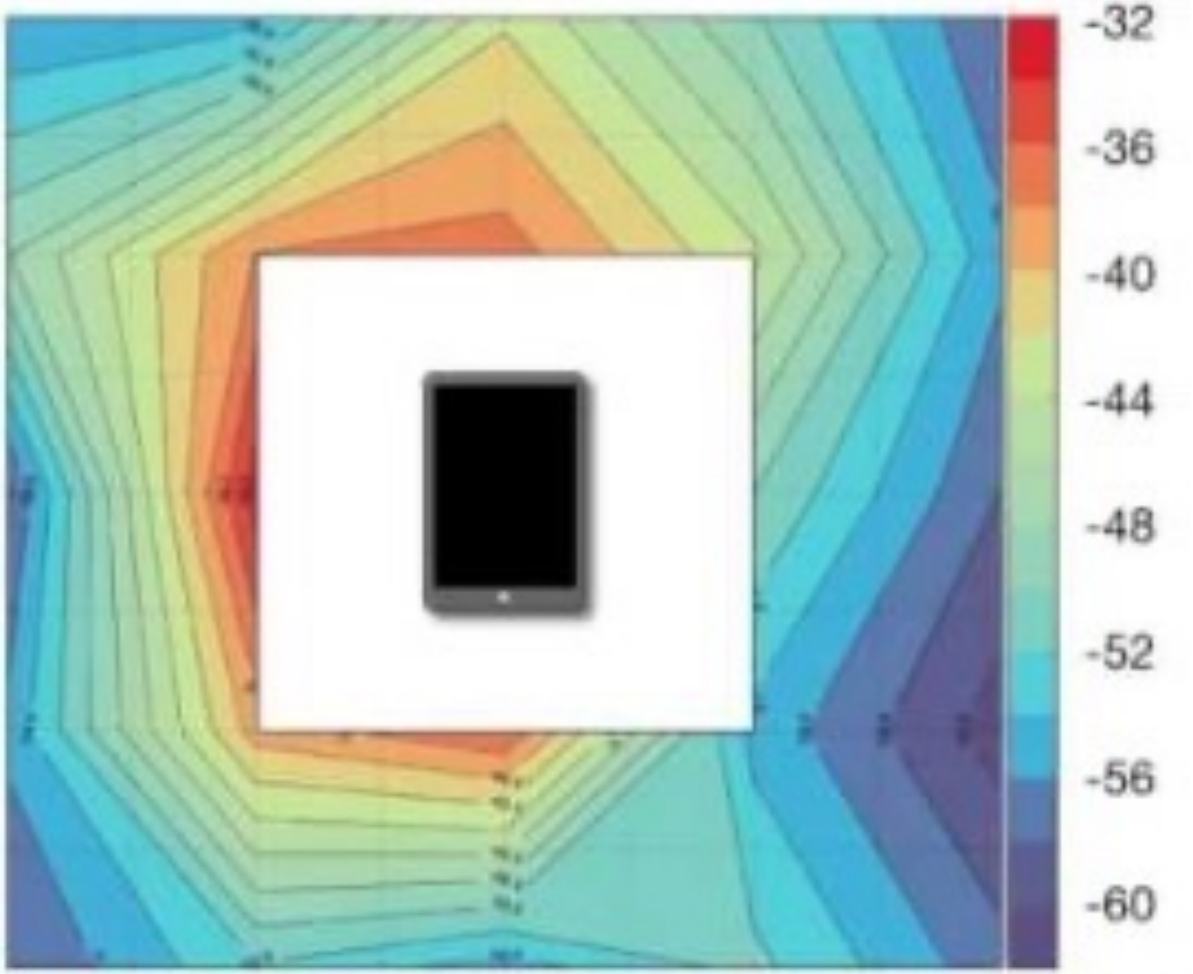
Haojian Jin
Jingxian Wang
Swarun Kumar
Jason Hong

Carnegie Mellon University



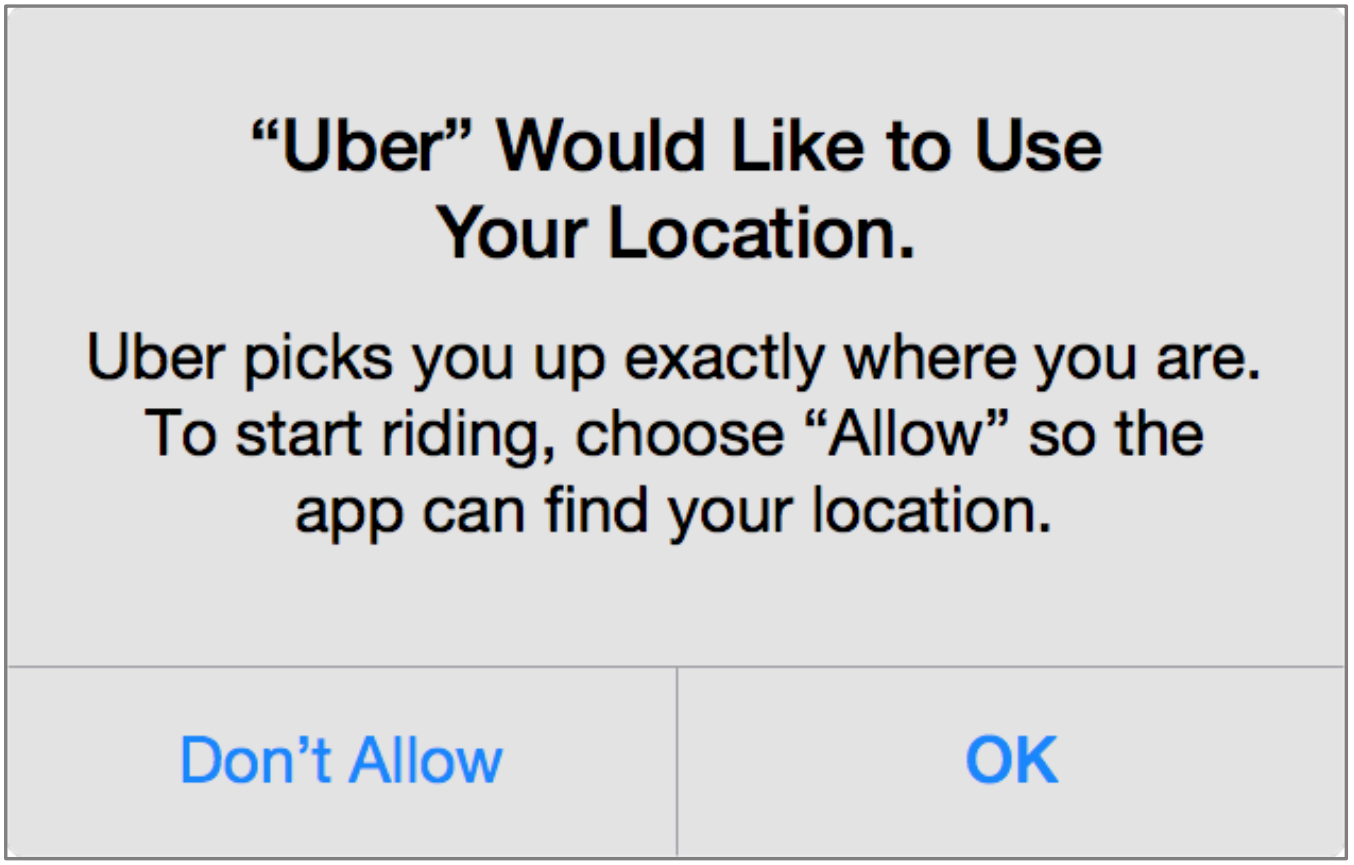
Working Code Trumps All Hype!

Working Code Trumps All Hype!

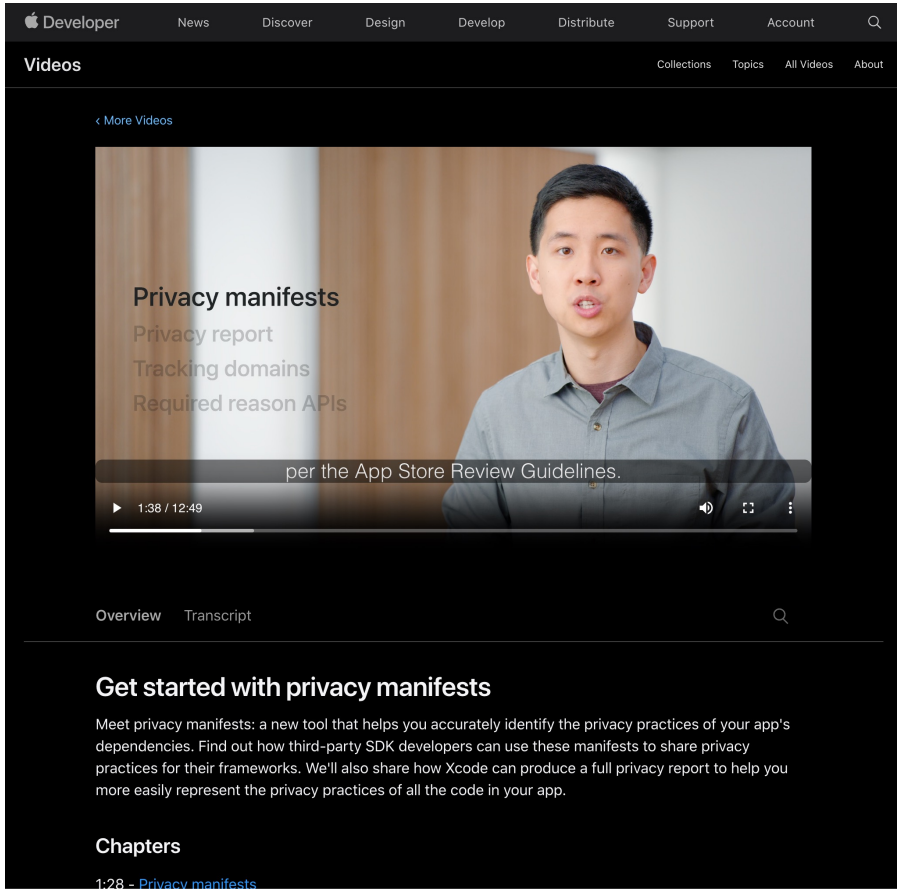


And mapped the asymmetric RSSI distribution around the iPad.

Working Code Trumps All Hype!

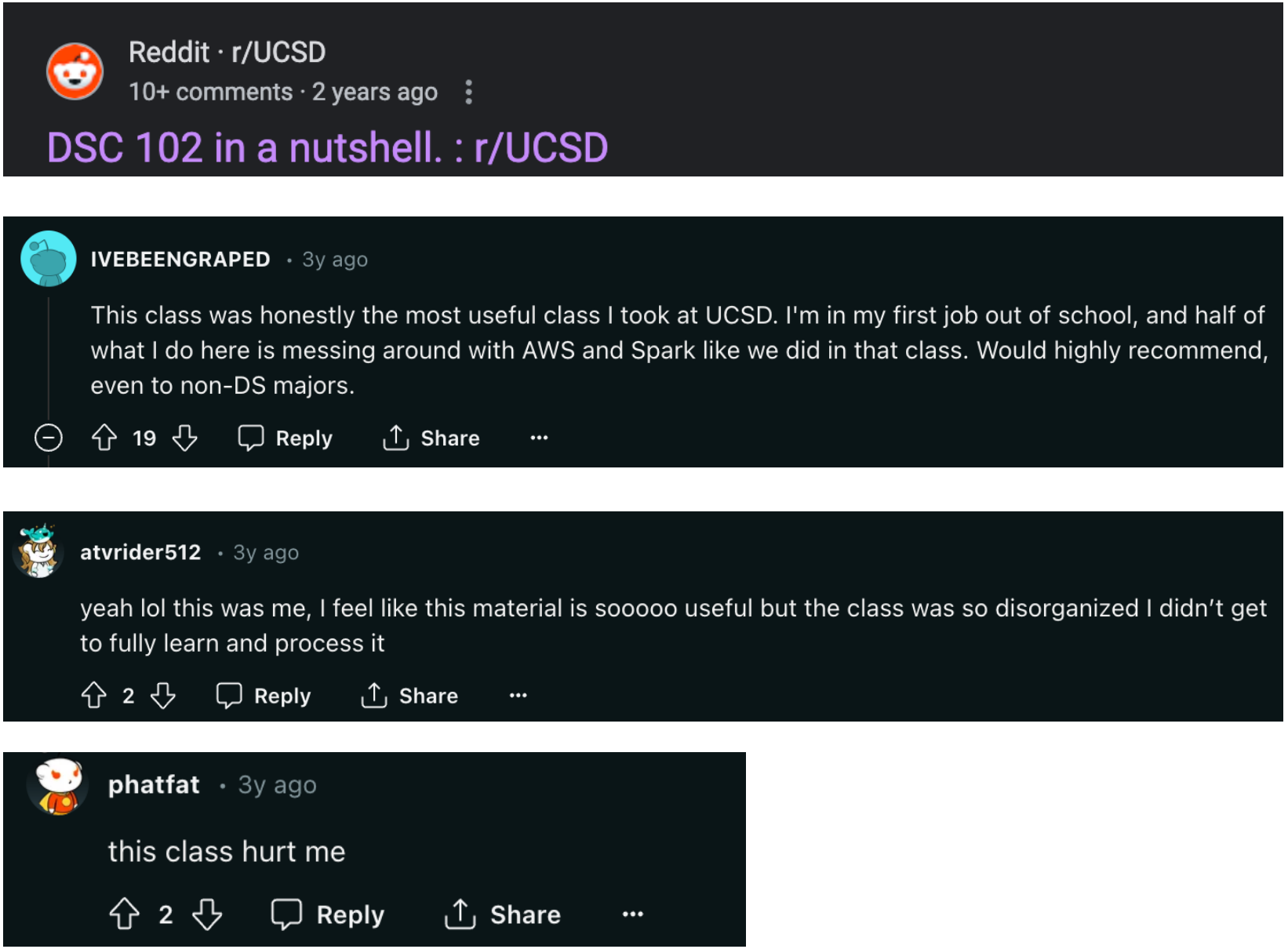


Some of them are getting adopted.



- ✓ Categorized Purpose string (2017 -> 2022)
- ✓ Declared manifests (2020->2024)
- ✓ Operator-based API (??)

What is this course about? Why take it?



https://www.reddit.com/r/UCSD/comments/npqcdm/dsc_102_in_a_nutshell/

glassdoor

statistician | Location

Statistician Salaries United States

Overview | **Salaries** | Interviews | Insights | Career Path

How much does a Statistician make?

Updated Jan 4, 2022

Industry: All industries

Employer Size: All company sizes

Experience: All years of Experience

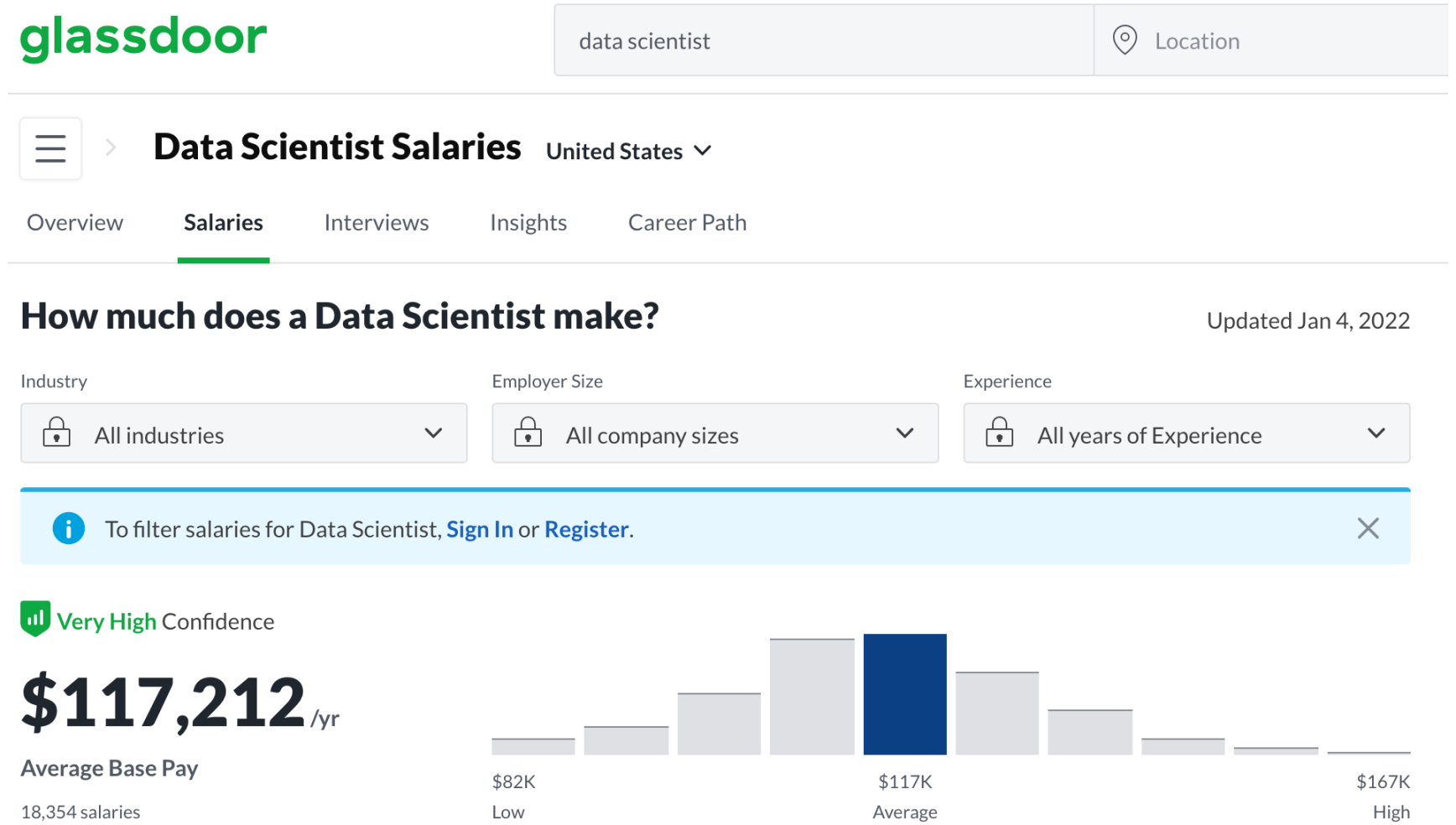
Very High Confidence

$88,989 /yr

Average Base Pay

2,398 salaries

$61K Low | $89K Average | $131K High


glassdoor
data scientist
Location
Data Scientist Salaries
United States
Overview
Salaries
Interviews
Insights
Career Path
How much does a Data Scientist make?
Updated Jan 4, 2022
Industry
All industries
Employer Size
All company sizes
Experience
All years of Experience
To filter salaries for Data Scientist, Sign In or Register.
Very High Confidence
$117,212 /yr
Average Base Pay
18,354 salaries
$82K
Low
$117K
Average
$167K
High

— 88,989

= 28,223!

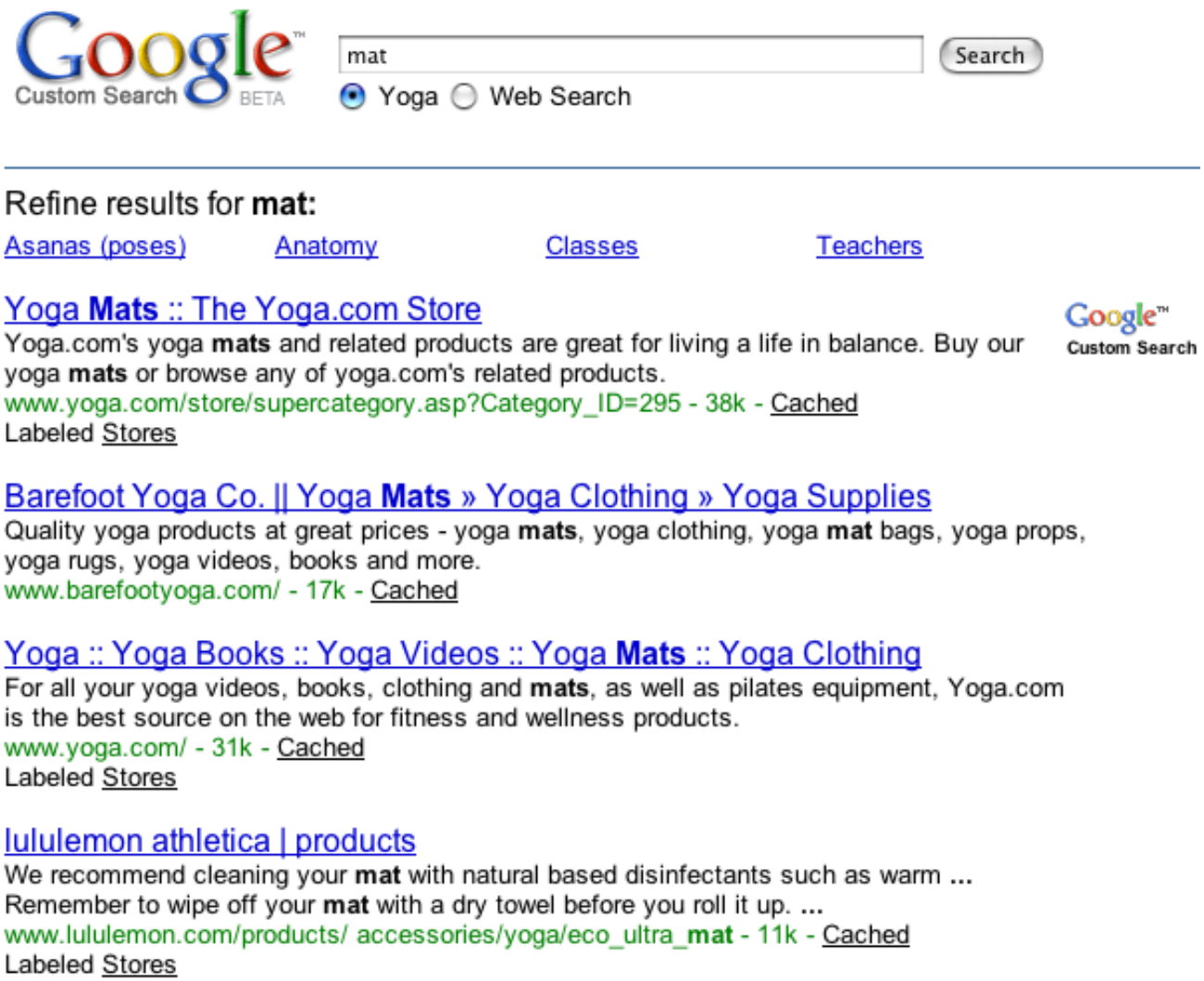
CHARLES GOODWIN
University of South Carolina

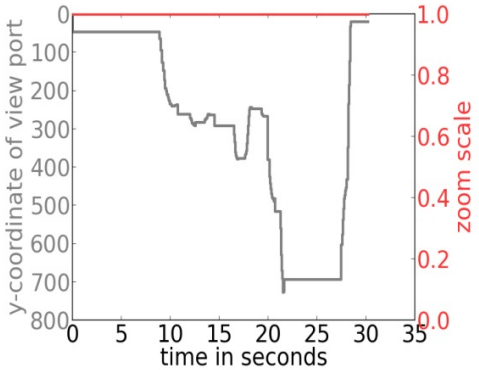
Professional Vision

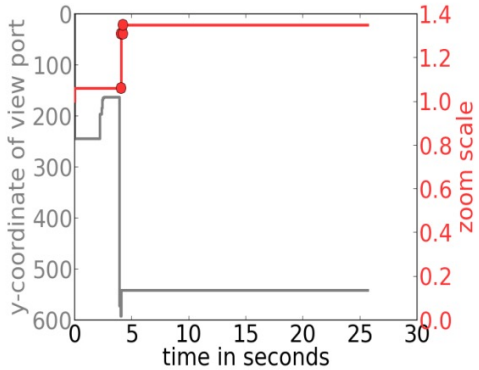
DISCURSIVE PRACTICES are used by members of a profession to shape events in the domains subject to their professional scrutiny. The shaping process creates the objects of knowledge that become the insignia of a profession's craft: the theories, artifacts, and bodies of expertise that distinguish it from other professions. Analysis of the methods used by members of a community to build and contest the events that structure their lifeworld contributes to the development of a practice-based theory of knowledge and action.[1] In this article, I examine two contexts of professional activity: archaeological field excavation and legal argumentation. In each of these contexts, I investigate three practices: (1) *coding,* which transforms phenomena observed in a specific setting into the objects of knowledge that animate the discourse of a profession; (2) *highlighting,* which makes specific phenomena in a complex perceptual field salient by marking them in some fashion; and (3) *producing and articulating material representations.* By applying such practices to phenomena in the domain of scrutiny, participants build and contest *professional vision,* which consists of socially or seeing and understanding events that are answerable to the distincti particular social group.

Coding
Highlighting
Reconstructing

What will happen if I click the second link?

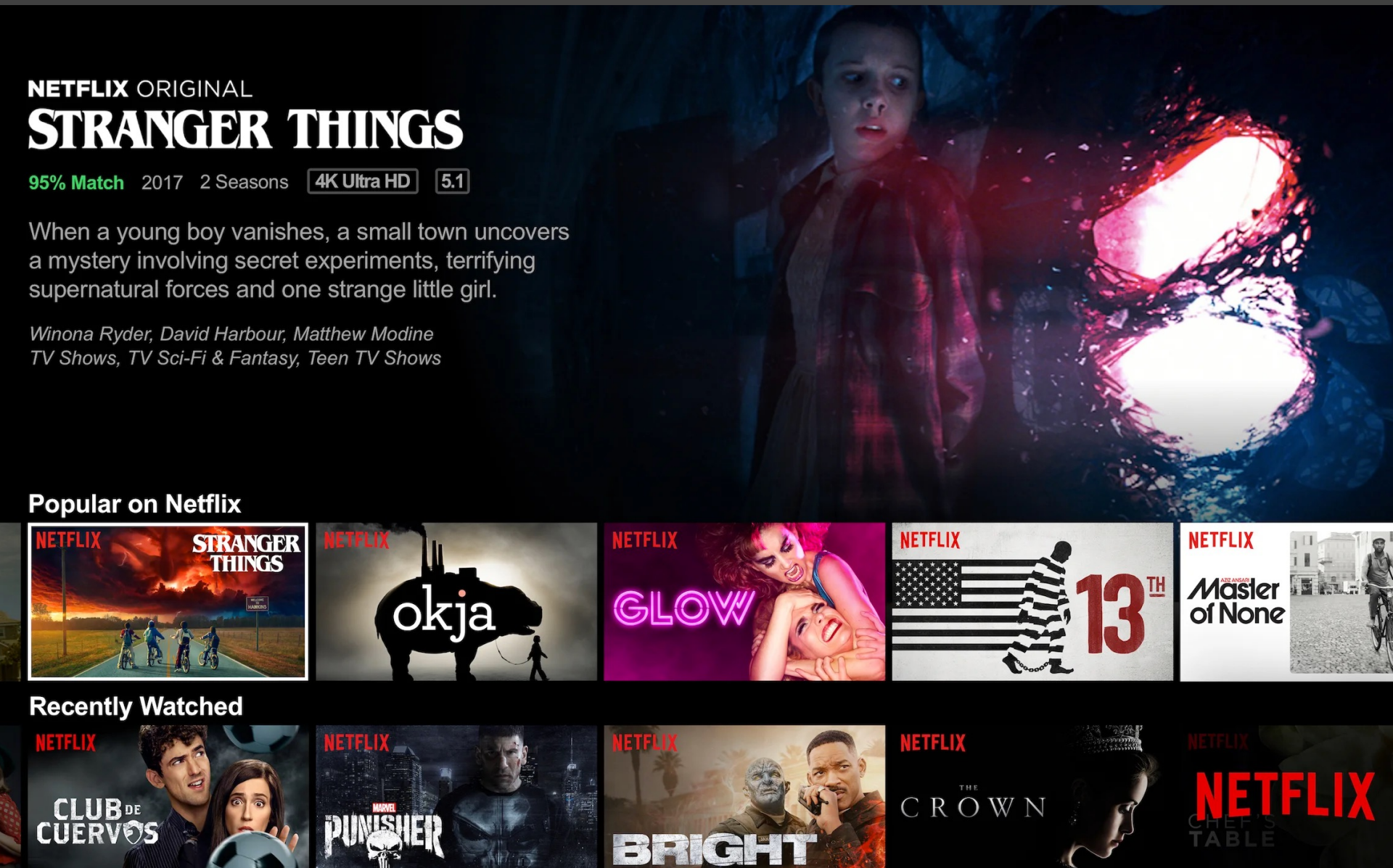






Mining Touch Interaction Data on Mobile Devices to Predict Web Search Result Relevance, SIGIR 2013

1. Netflix's "spot-on" recommendations



TV Schedule

TVLine

FALL TV GRID 2022

New series are listed in RED

SUN	8:00	8:30	9:00	9:30	10:00	10:30
ABC	7 pm AFV | CELEBRITY JEOPARDY		CELEBRITY WHEEL OF FORTUNE		THE ROOKIE	
CBS	7 pm 60 MIN | THE EQUALIZER		EAST NEW YORK		NCIS: LOS ANGELES	
FOX	To be announced					
NBC	SUNDAY NIGHT FOOTBALL					
CW	FAMILY LAW (acquired)		CORONER (acquired)			

MON	8:00	8:30	9:00	9:30	10:00	10:30
ABC	BACHELOR IN PARADISE				THE GOOD DOCTOR	
CBS	THE NEIGHBORHOOD	BOB HEARTS ABISHOLA	NCIS		NCIS: HAWAI'I	
FOX	To be announced					
NBC	THE VOICE				QUANTUM LEAP	
CW	ALL AMERICAN		ALL AMERICAN: HOMECOMING			

TUE	8:00	8:30	9:00	9:30	10:00	10:30
ABC	BACHELOR IN PARADISE				THE ROOKIE: FEDS	
CBS	FBI		FBI: INTERNATIONAL		FBI: MOST WANTED	
FOX	To be announced					
NBC	THE VOICE		LA BREA		NEW AMSTERDAM	
CW	THE WINCHESTERS		PROFESSIONALS (acquired)			

WED	8:00	8:30	9:00	9:30	10:00	10:30
ABC	THE CONNERS	THE GOLDBERGS	ABBOTT ELEMENTARY	HOME ECONOMICS	BIG SKY	
CBS	SURVIVOR		THE AMAZING RACE		THE REAL LOVE BOAT	
FOX	To be announced					
NBC	CHICAGO MED		CHICAGO FIRE		CHICAGO P.D.	
CW	DC'S STARGIRL		KUNG FU			

THU	8:00	8:30	9:00	9:30	10:00	10:30
ABC	STATION 19		GREY'S ANATOMY		ALASKA	
CBS	YOUNG SHELDON	GHOSTS	SO HELP ME TODD		CSI: VEGAS	
FOX	To be announced					
NBC	LAW & ORDER		LAW & ORDER: SVU		LAW & ORDER: ORGANIZED CRIME	
CW	WALKER		WALKER INDEPENDENCE			

FRI	8:00	8:30	9:00	9:30	10:00	10:30
ABC	SHARK TANK		20/20			
CBS	S.W.A.T.		FIRE COUNTRY		BLUE BLOODS	
FOX	FRIDAY NIGHT SMACKDOWN (presumably!)					
NBC	COLLEGE BOWL (until November) LOPEZ VS LOPEZ | YOUNG ROCK		DATELINE NBC			
CW	PENN & TELLER: FOOL US		WHOSE LINE IS IT ANYWAY? X2			

SAT	8:00	8:30	9:00	9:30	10:00	10:30
ABC	COLLEGE FOOTBALL					
CBS	Drama encores		Drama encores		48 HOURS	
FOX	To be announced					
NBC	Drama encores		DATELINE WEEKEND MYSTERY		SNL VINTAGE	| 11:30 SNL
CW	MAGIC WITH THE STARS		WORLD'S FUNNIEST ANIMALS x2			

How does Netflix know that?

Large datasets + Machine learning!

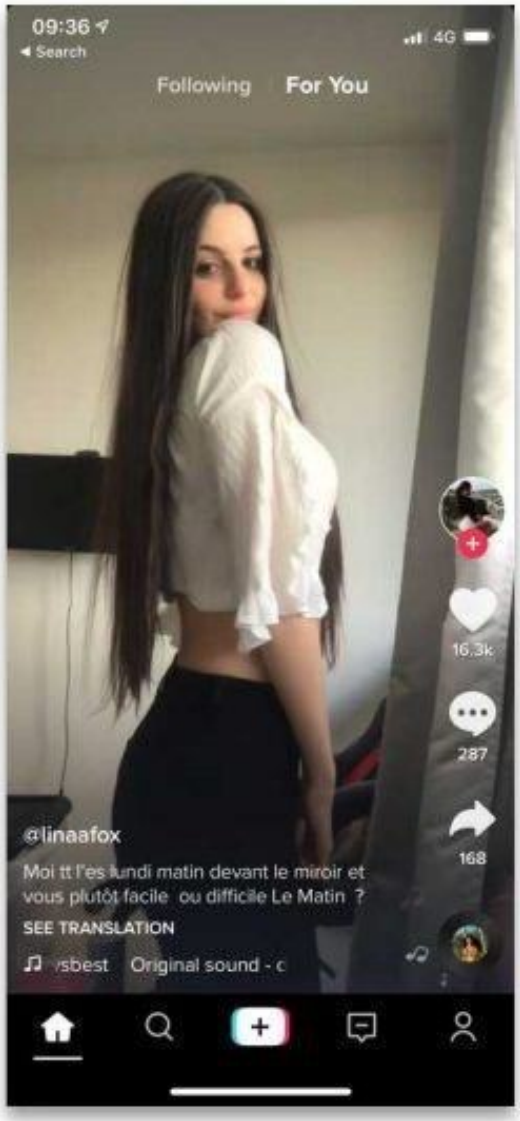
Log all user behavior (views, clicks, pauses, searches, etc.)
Recommender systems apply ML to TBs of data from all users and movies to deliver a tailored experience

Pandora v.s. Spotify

Music Genome Project

Collaborative filters (social)

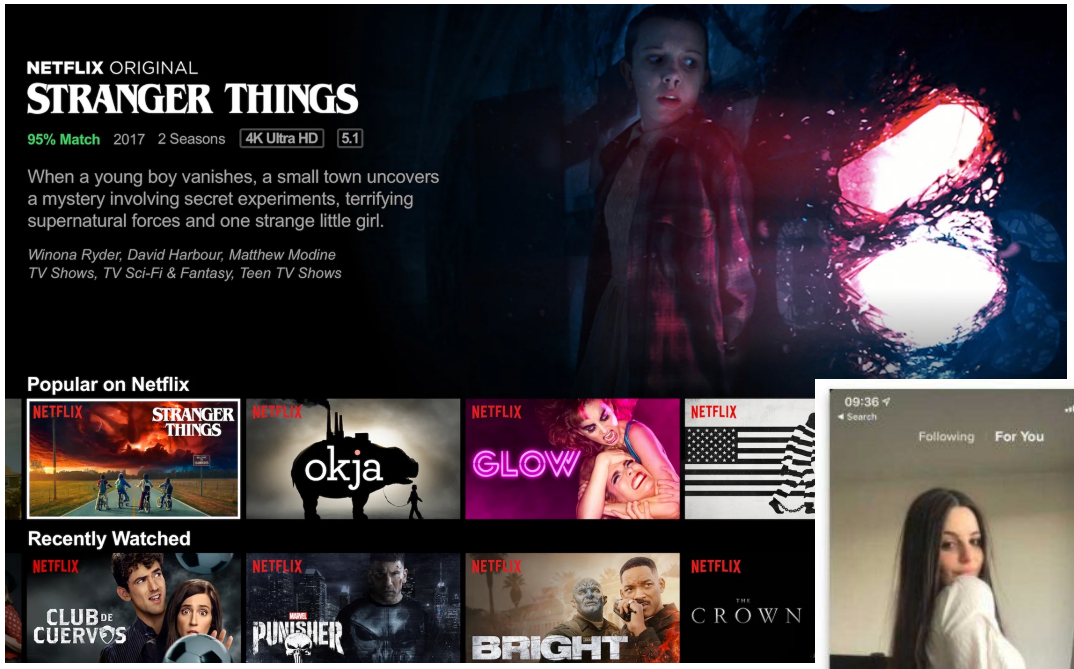
Tiktok

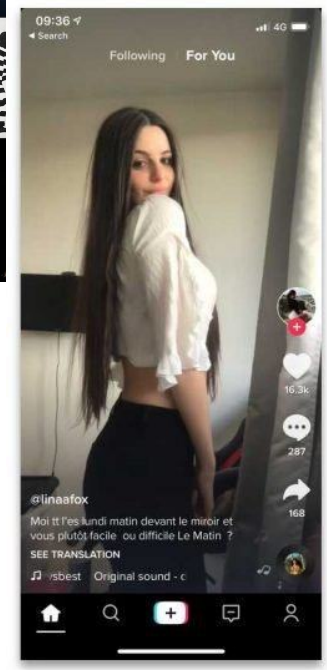

09:36
Search
4G
Following
For You
16.3k
287
168
@linaafox
Moi tt l'es lundi matin devant le miroir et vous plutôt facile ou difficile Le Matin ?
SEE TRANSLATION
Original sound - c


09:37
Search
4G
Following
For You
134.7k
962
1621
@fabianfx
#bratz #makeup #bratzchallenge
SEE TRANSLATION
Original s

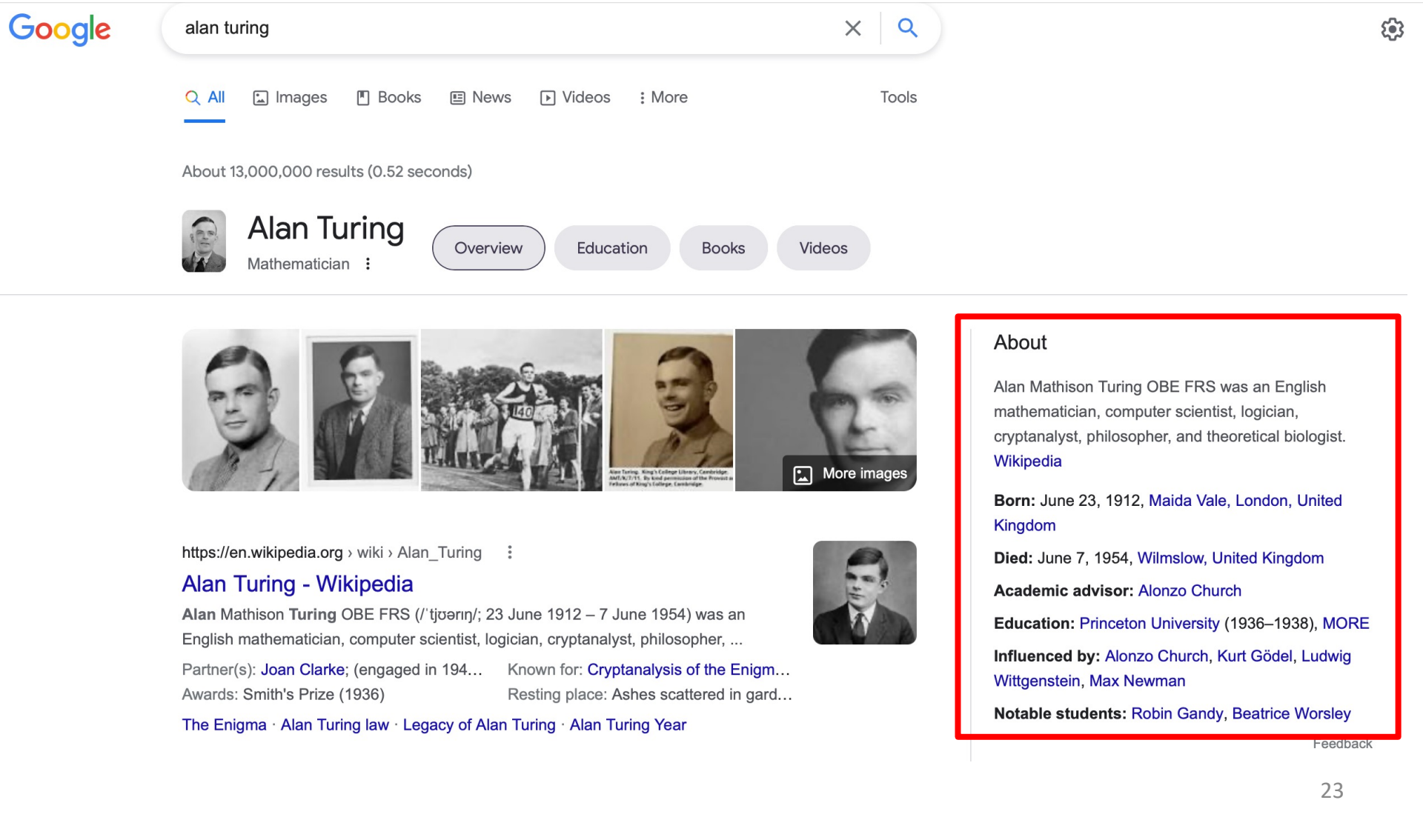

09:46
Search
4G
Following
For You
532.9k
1774
4913
@ilonaaln
MERCI POUR LE MILLION jpp nos tête
on a tellement eu peur #fail
#slowmotion @lea_spk
SEE TRANSLATION
Original sound - itsofficia

From Netflix to TikTok: More than algorithms



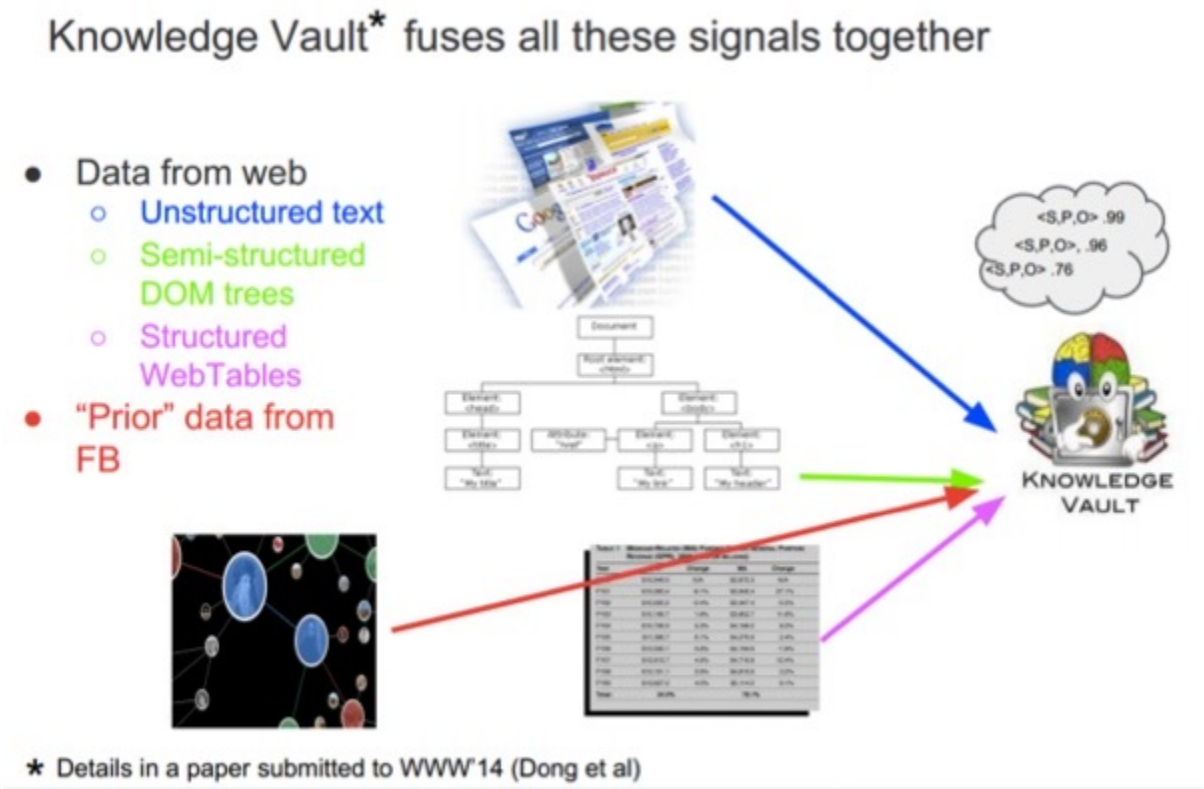


2. Structured data with search results



How does Google know that?

Large datasets + Machine learning!



Knowledge Base Construction (KBC) process extracts tabular/relational data from large amounts of text data

Innumerable “enterprise” applications

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

307 comments, 167 called-out + Comment Now + Follow Comments

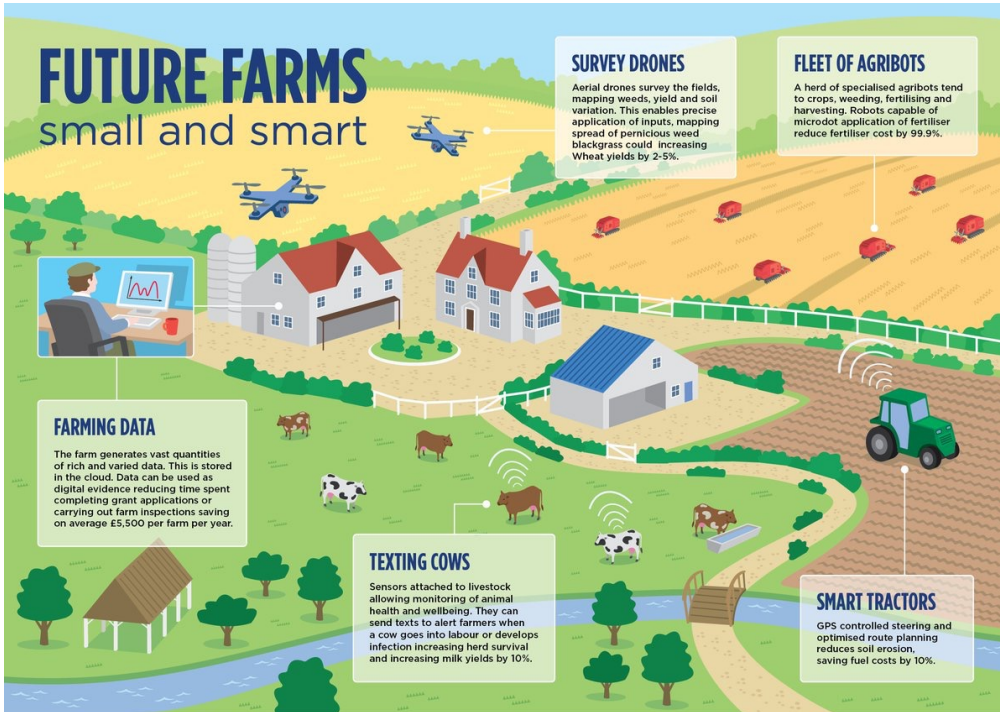
Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Charles Duhigg outlines in the New York Times how Target tries to hook parents-to-be at that crucial moment before they turn into

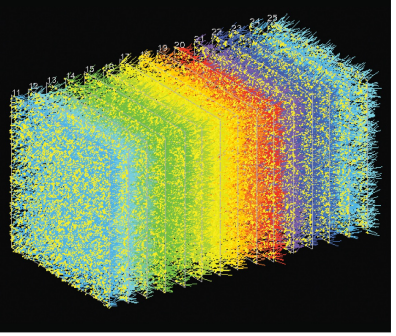
Target has got you in its aim

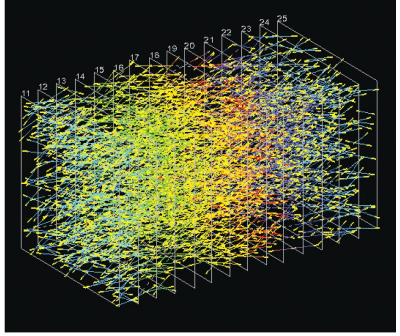

Enter

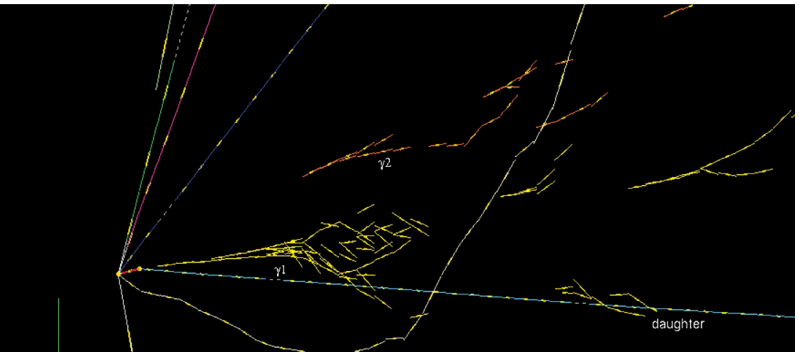

I ♥ INSURANCE
FLO

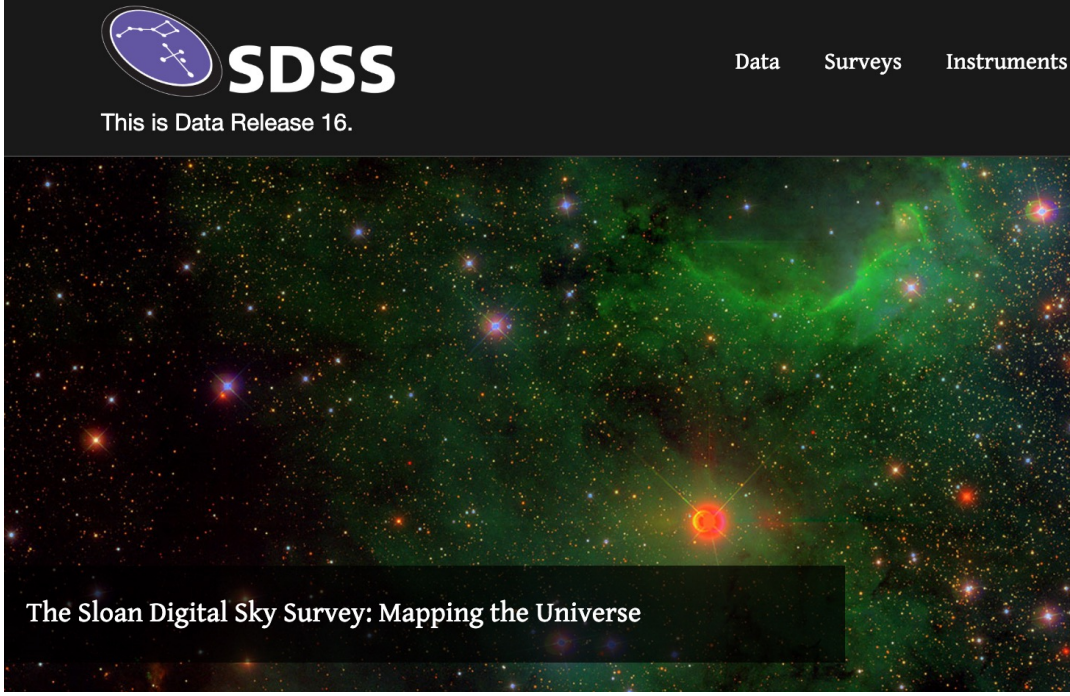

FUTURE FARMS
small and smart
SURVEY DRONES
Aerial drones survey the fields, mapping weeds, yield and soil variation. This enables precise application of inputs, mapping spread of pernicious weed blackgrass could increasing Wheat yields by 2-5%.
FLEET OF AGRIBOTS
A herd of specialised agribots tend to crops, weeding, fertilising and harvesting. Robots capable of microdot application of fertiliser reduce fertiliser cost by 99.9%.
FARMING DATA
The farm generates vast quantities of rich and varied data. This is stored in the cloud. Data can be used as digital evidence reducing time spent completing grant applications or carrying out farm inspections saving on average £5,500 per farm per year.
TEXTING COWS
Sensors attached to livestock allowing monitoring of animal health and wellbeing. They can send texts to alert farmers when a cow goes into labour or develops infection increasing herd survival and increasing milk yields by 10%.
SMART TRACTORS
GPS controlled steering and optimised route planning reduces soil erosion, saving fuel costs by 10%.

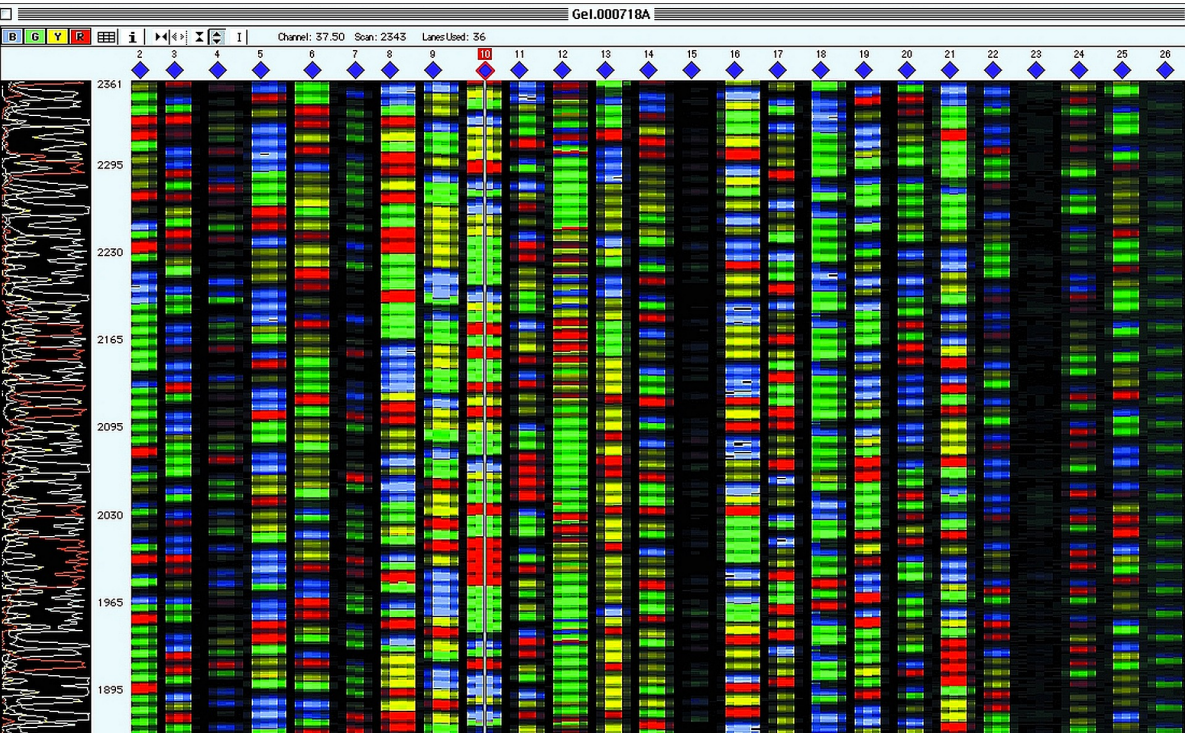
“Domain sciences” and healthcare tech are also becoming data+ML intensive

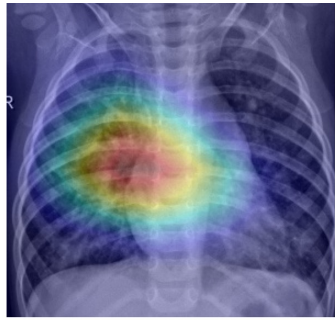





γ2
γ1
daughter

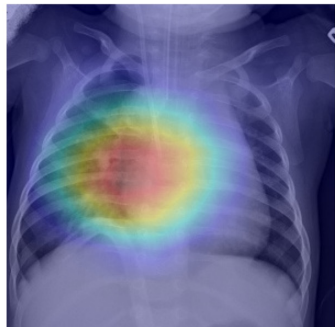

SDSS
Data
Surveys
Instruments
This is Data Release 16.
The Sloan Digital Sky Survey: Mapping the Universe


Gel.000718A
Channel: 37.50 Scan: 2343 Lanes Used: 36



(a)

(b)



Software systems for data analytics and ML over large and complex datasets are now critical for digital applications in many domains

The Age of “Big Data”/“Data Science”

The New York Times

SundayReview | NEWS ANALYSIS

The Age

By STEVE LOHR

Email

Share

Tweet

Save

Forbes / Entrepreneurs

Forbes

MAR 25, 2015 @ 7:33 PM 4,407 VIEWS

Drowning In Big Data - Finding Insight In A Digital

Josh Steimle, CONT

For roughly a deca
information about
Big Data. The IDC
industry will exper
by 2018. What this

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

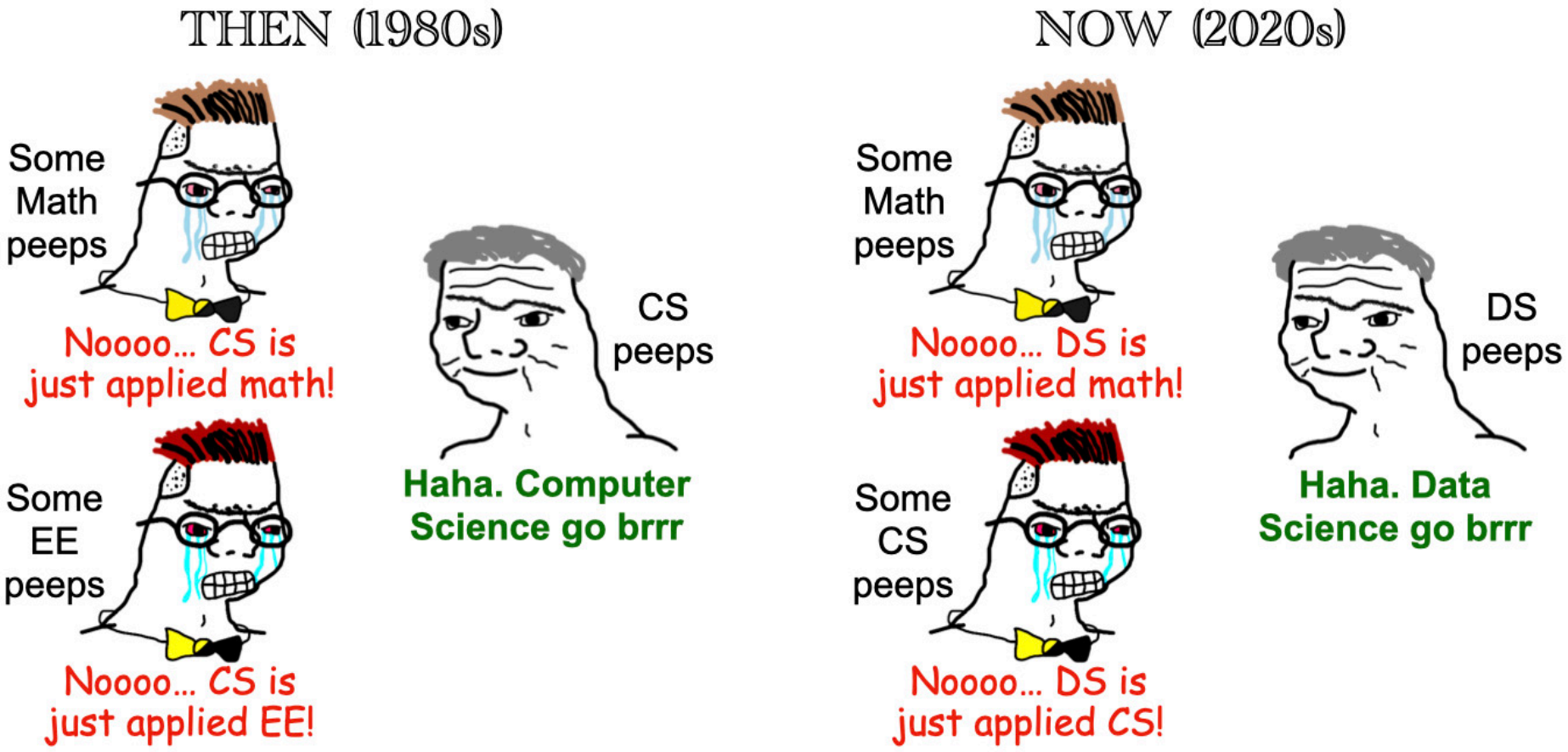
FROM THE OCTOBER 2012 ISSUE

Harvard Business Review

SUMMARY SAVE SHARE COMMENT 5 TEXT SIZE PRINT $8.95 BUY COPIES

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—

Meme from Previous DSC 102



The Discipline of Machine Learning

Tom M. Mitchell

July 2006
CMU-ML-06-108

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

*Machine Learning Department
†School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA

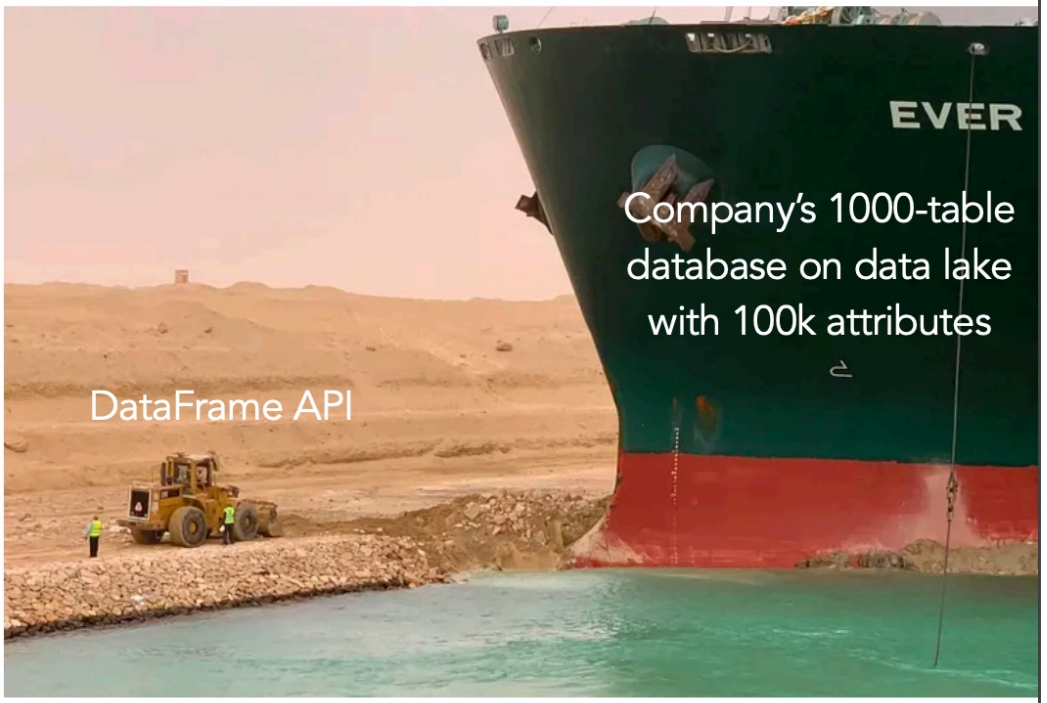
Abstract

Over the past 50 years the study of Machine Learning has grown from the efforts of a handful of computer engineers exploring whether computers could learn to play games, and a field of Statistics that largely ignored computational considerations, to a broad discipline that has produced fundamental statistical-computational theories of learning processes, has designed learning algorithms that are routinely used in commercial systems for speech recognition, computer vision, and a variety of other tasks, and has spun off an industry in data mining to discover hidden regularities in the growing volumes of online data. This document provides a brief and personal view of the discipline that has emerged as Machine Learning, the fundamental questions it addresses, its relationship to other sciences and society, and where it might be headed.

DSC 204a

DSC 204a Scalable Data Systems

- Haojian Jin



Meme idea credit: https://datasystemsfun.tumblr.com/

Vision

Data science professionals ought to be familiarized with data systems from a user's standpoint, as opposed to the conventional approach of a system implementer.

15-213/15-513/14-513 Introduction to Computer Systems (ICS)

Fall 2023

- 15-213 Pittsburgh: Tue, Thu 12:30 PM–01:50 PM, GHC 4401, Brian Railing and Phillip Gibbons
- 14-513 Pittsburgh: Tue, Thu 12:30 PM–01:50 PM, CIC 1202, David Varodayan

12 units

The ICS course provides a programmer's view of how computer systems execute programs, store information, and communicate. It enables students to become more effective programmers, especially in dealing with issues of performance, portability and robustness. It also serves as a foundation for courses on compilers, networks, operating systems, and computer architecture, where a deeper understanding of systems-level issues is required. Topics covered include: machine-level code and its generation by optimizing compilers, performance evaluation and optimization, computer arithmetic, memory organization and management, networking technology and protocols, and supporting concurrent computation.

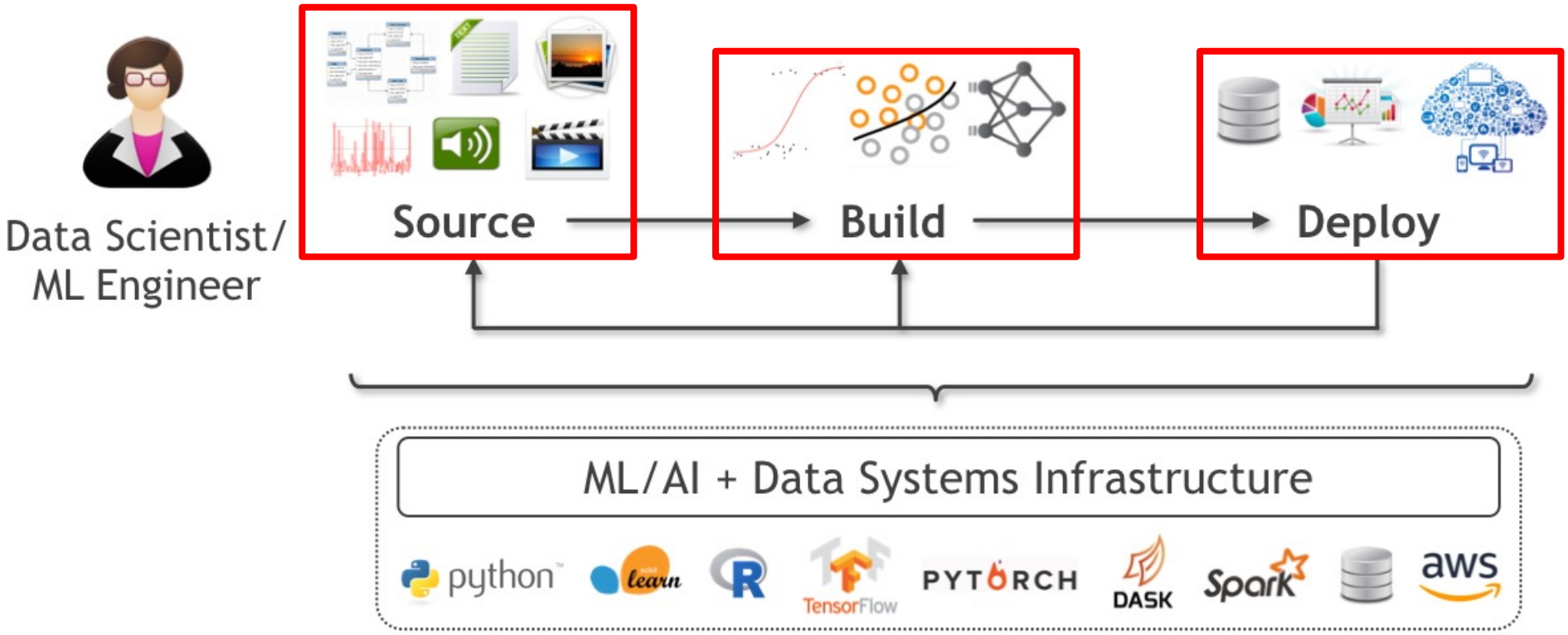
Course Syllabus

Prerequisites: 15-122

DSC 102 will get you thinking about the fundamentals of systems for scalable analytics

1. "**Systems**": What resources does a computer have? How to store and efficiently compute over large data? What is cloud?
2. "**Scalability**": How to scale and parallelize data-intensive computations?
3. **For "Analytics"**:
    1. **Source**: Data acquisition & preparation for ML
    2. **Build**: Model selection & deep learning systems
    3. **Deploying** ML models
4. Hands-on experience with scalable analytics tools

The Lifecycle of ML-based Analytics



Data acquisition
Data preparation

Feature Engineering
Training & Inference
Model Selection

Serving
Monitoring

ML Systems

***Q:** What is a Machine Learning (ML) System?*

- A data processing system (aka *data system*) for mathematically advanced data analysis operations (inferential or predictive):
  - Statistical analysis; ML, deep learning (DL); data mining (domain-specific applied ML + feature eng.)
  - *High-level APIs* to express ML computations over (large) datasets
  - *Execution engine* to run ML computations efficiently

Categorizing ML Systems

- **Orthogonal Dimensions of Categorization**:
  1. **Scalability:** In-memory libraries v. Scalable ML system (works on larger-than-memory datasets)
  2. **Target Workloads:** General ML library v. Decision tree-oriented v. Deep learning, etc.
  3. **Implementation Reuse:** Layered on top of scalable data system v. Custom from-scratch framework

Major Existing ML Systems

General ML libraries:

In-memory:

Disk-based files:

Layered on RDBMS/Spark:

Cloud-native:

Azure Machine Learning

Amazon SageMaker

"AutoML" platforms:

DataRobot

Decision tree-oriented:

Microsoft
LightGBM

Deep learning-oriented:

Data Systems Concerns in ML

Key concerns in ML:

Accuracy

Runtime efficiency (sometimes)

Additional key *practical* concerns in ML Systems:

Scalability (and efficiency at scale)

Usability

Manageability

Developability

Q: How do "ML Systems" relate to ML?

ML Systems : ML :: Computer Systems : TCS

*Long-standing concerns in the **DB systems** world!*

Q: [illegible] ML systems?

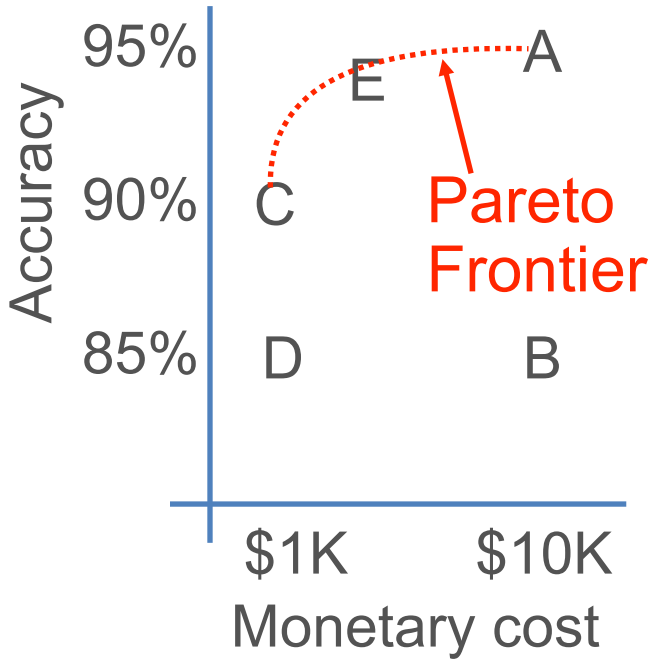
Conceptual System Stack Analogy

	Relational DB Systems	ML Systems
Theory	First-Order Logic Complexity Theory	Learning Theory Optimization Theory
Program Formalism	Relational Algebra	Tensor Algebra Gradient Descent
Program Specification	SQL	TensorFlow? Scikit-learn?
Program Modification	Query Optimization	???
Execution Primitives	Parallel Relational Operator Dataflows	Depends on ML Algorithm
Hardware	CPU, GPU, FPGA, NVM, RDMA, etc.	

Real-World ML: Pareto Surfaces

Q: *Suppose you are given ad click-through prediction models A, B, C, and D with accuracies of 95%, 85%, 90%, and 85%, respectively. Which one will you pick?*

Q: *What about now?*



- Real-world ML users must grapple with multi-dimensional *Pareto surfaces*: accuracy, monetary cost, training time, scalability, inference latency, tool availability, interpretability, fairness, etc.
- *Multi-objective optimization* criteria set by application needs / business policies.

Learning Outcomes of this course

- **Explain** the basic principles of the memory hierarchy, parallelism paradigms, scalable data systems, and cloud computing.
- **Identify** the abstract data access patterns of, and opportunities for parallelism and efficiency gains in, data processing and ML algorithms at scale.
- **Outline** how to use cluster and cloud services, dataflow ("Big Data") programming with MapReduce and Spark, and ML tools at scale.
- **Apply** the above programming skills to create end-to-end pipelines for data preparation, feature engineering, and model selection on large-scale datasets.
- **Reason** critically about practical tradeoffs between accuracy, runtimes, scalability, usability, and total cost.

What this course is NOT about

- NOT a course on databases, relational model, or SQL
  - Take DSC 100 instead (pre-requisite)
- NOT a course on internal details of RDBMSs
  - Take CSE 132C instead
- NOT a training module for how to use Spark
- NOT a course on ML or data mining *algorithmics*; instead, we focus on ML *systems*