UC San Diego

DSC 102
Systems for Scalable Analytics

Winter 2022

Arun Kumar

About Myself

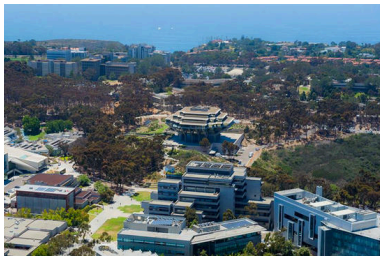
Summers: 110F!

2009: Bachelors in CSE from IIT Madras, India

Winters: —40F!

2009—16: MS and PhD in CS from UW-Madison

PhD thesis area: Data systems for ML workloads



Ahh! :)

2016-: Asst. Prof. at UC San Diego CSE

2019-: + Asst. Prof. at UC San Diego HDSI

2021: Assoc. Prof. at CSE & HDSI

My Current Research

New abstractions, algorithms, and software systems
to **"*democratize"*** ML-based data analytics from
a data management/systems standpoint

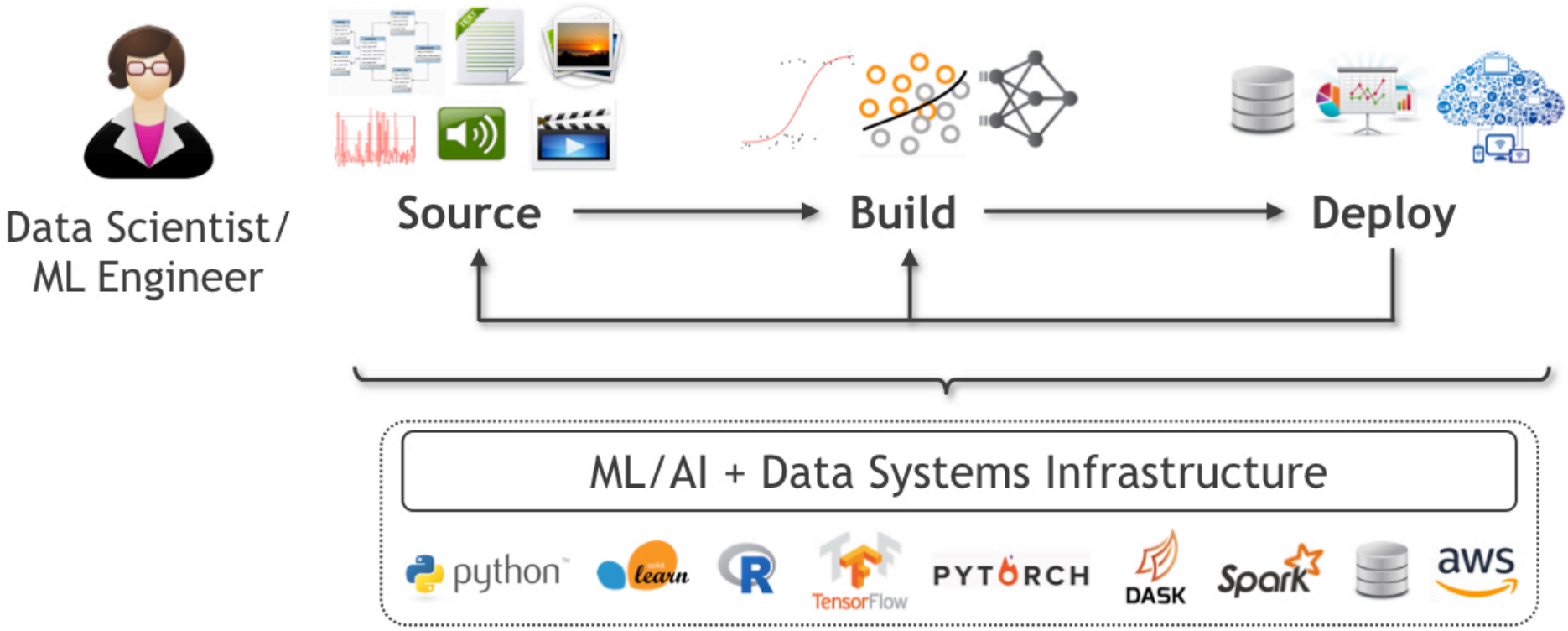
Democratization = **System Efficiency (Lower resource costs)** **+** **Human Efficiency (Higher productivity)**

Practical and scalable data systems for ML analytics

Inspired by *relational database systems* principles

Exploit insights from *learning theory* and *optimization theory*

My Current Research

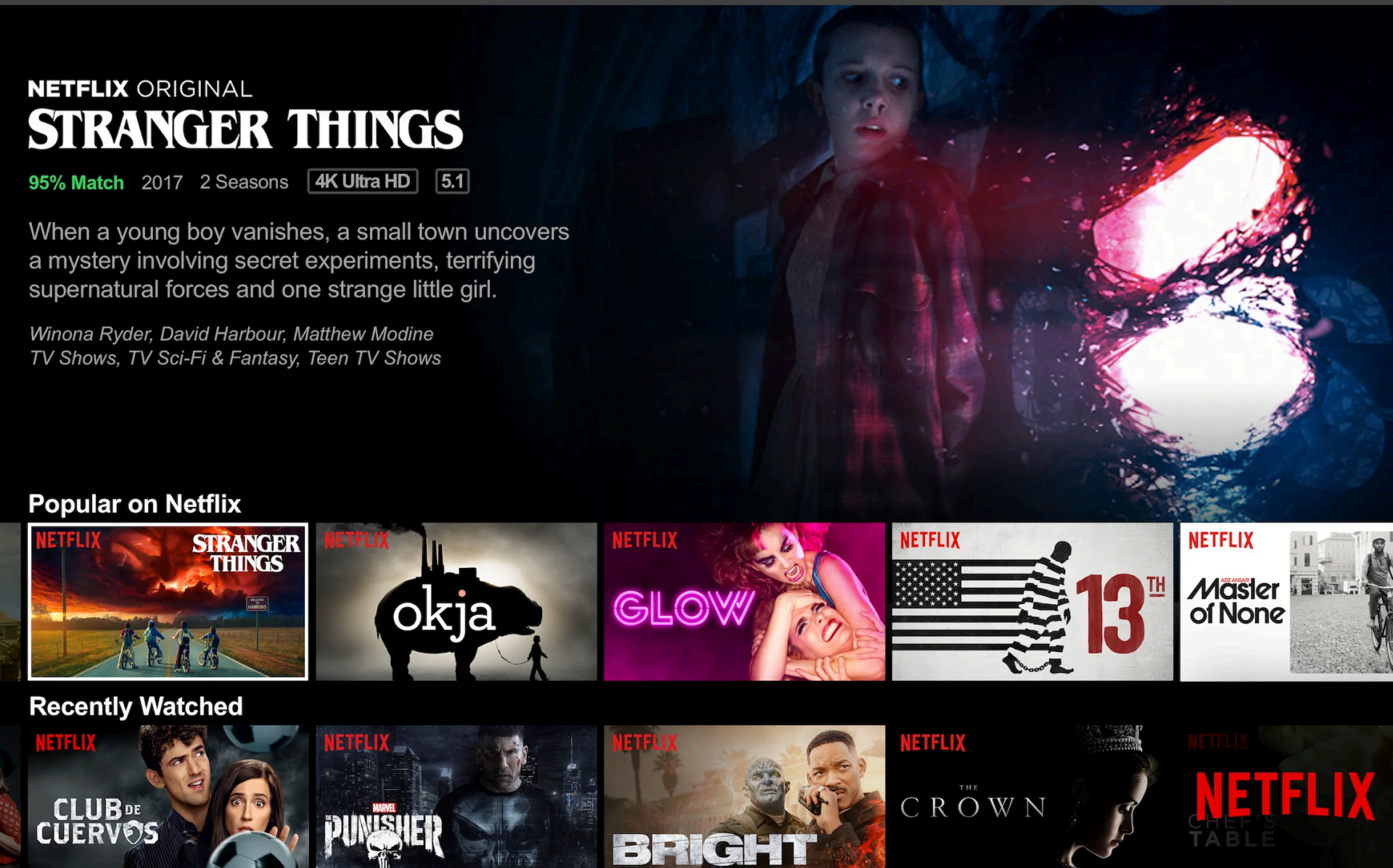


Research Approach: ***Abstract*** key steps + ***Formalize*** computation + ***Automate*** grunt work + ***Optimize*** execution

https://adalabucsd.github.io/

What is this course about? Why take it?

1. Netflix's "spot-on" recommendations

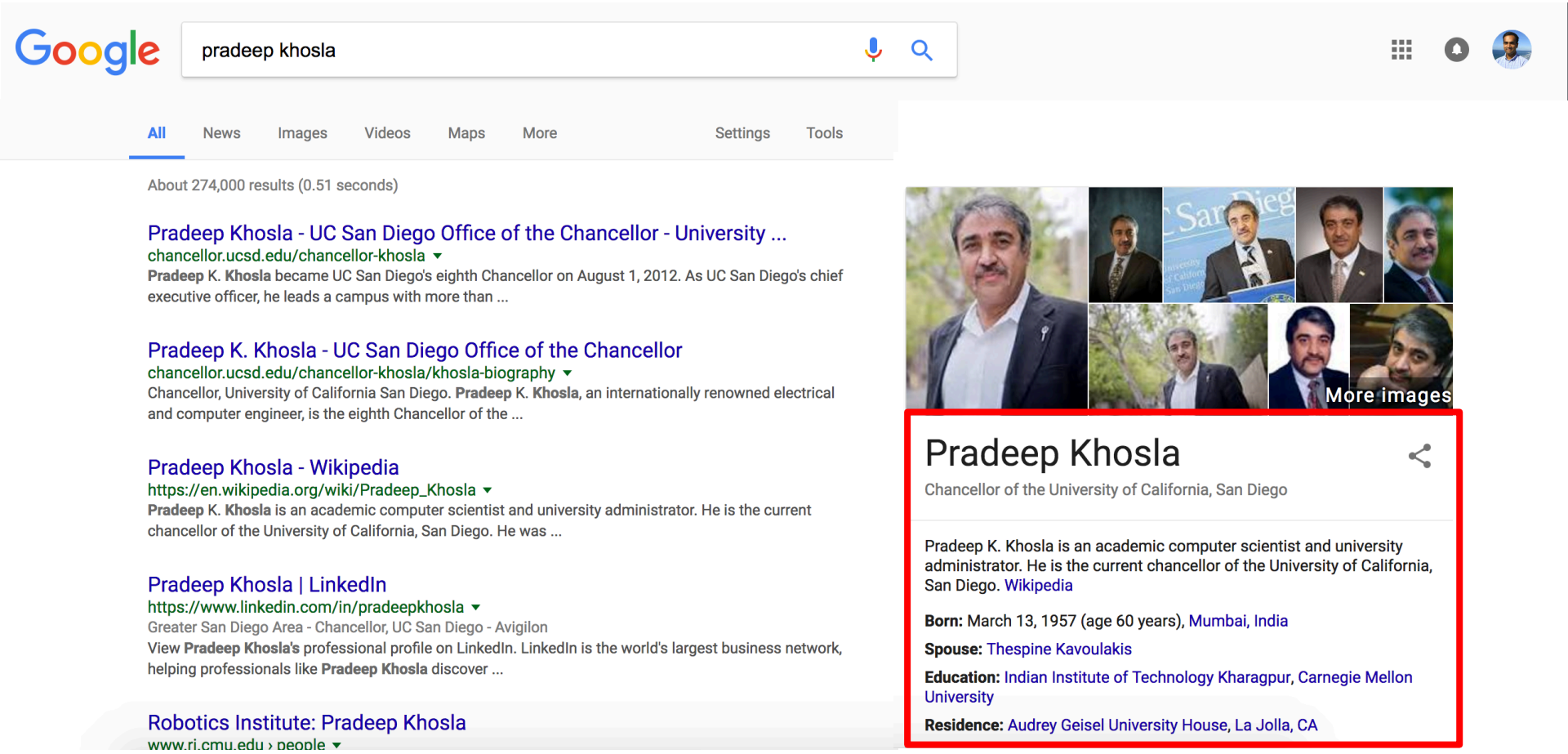


How does Netflix know that?

Large datasets + Machine learning!

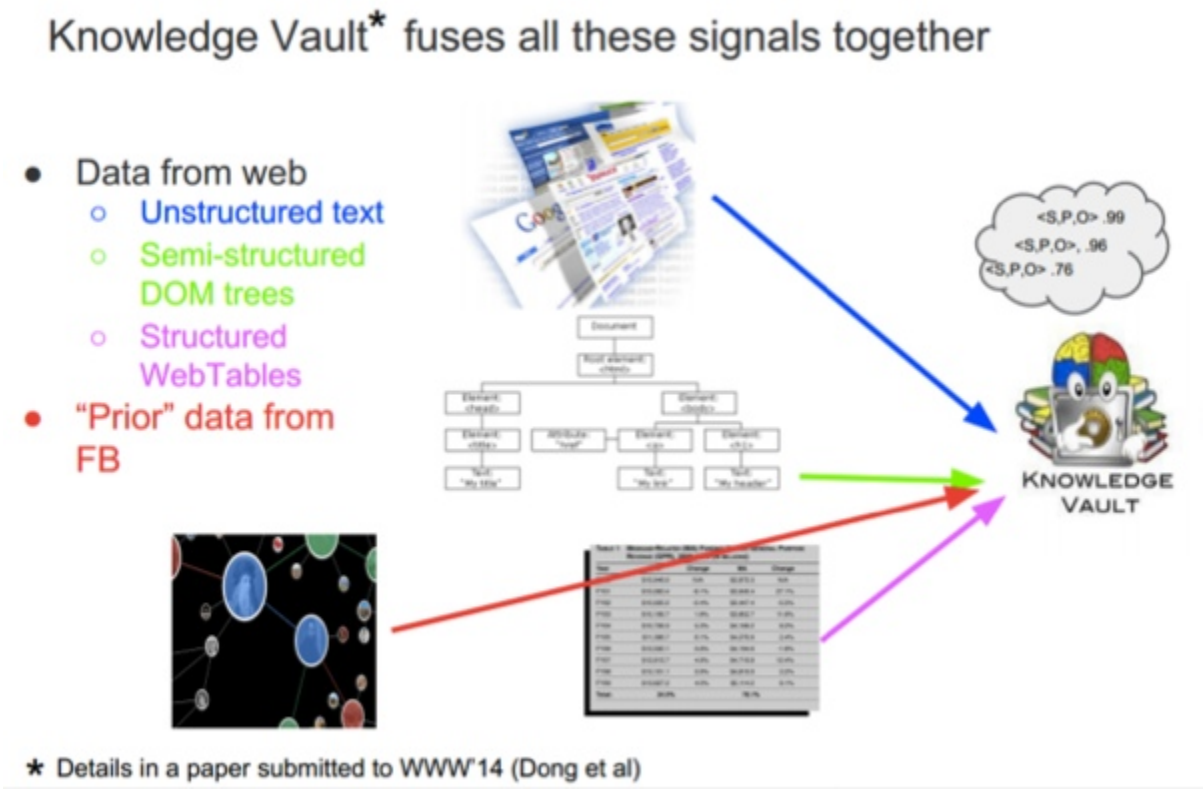
Log all user behavior (views, clicks, pauses, searches, etc.)
Recommender systems apply ML to TBs of data from all users and movies to deliver a tailored experience

2. Structured data with search results



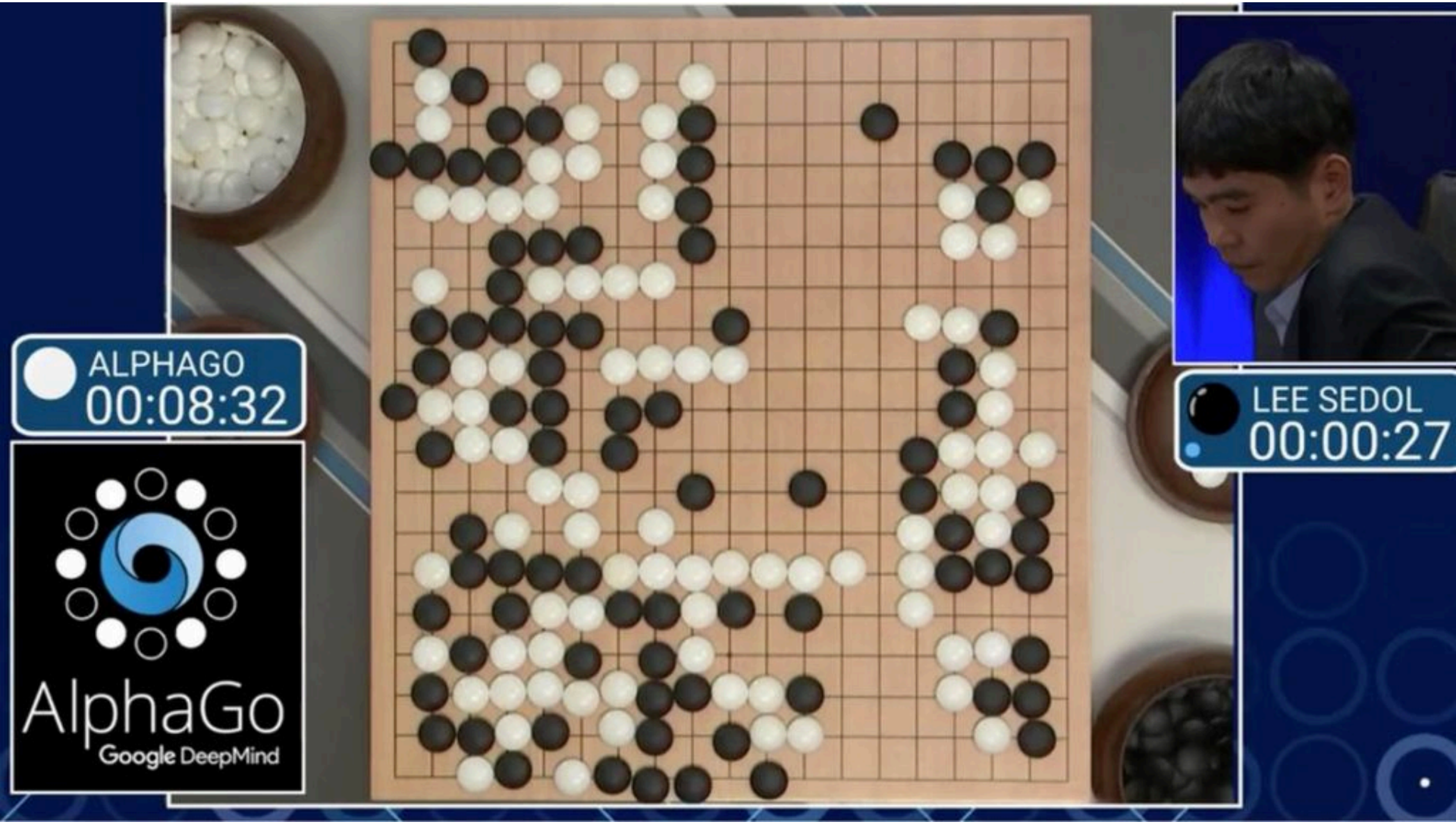
How does Google know that?

Large datasets + Machine learning!



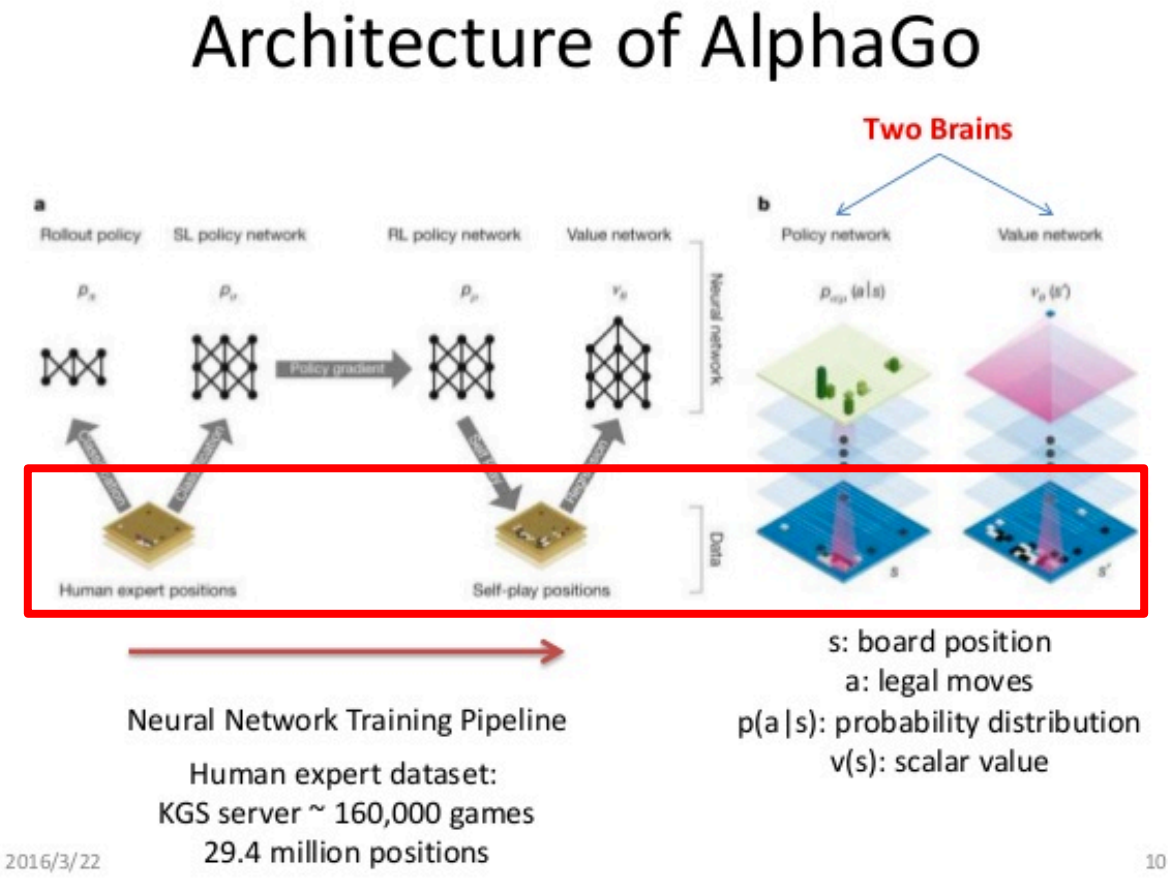
Knowledge Base Construction (KBC) process extracts tabular/relational data from large amounts of text data

3. AlphaGo defeats human champion!



How did AlphaGo achieve that?

Breakthrough powered by deep learning!



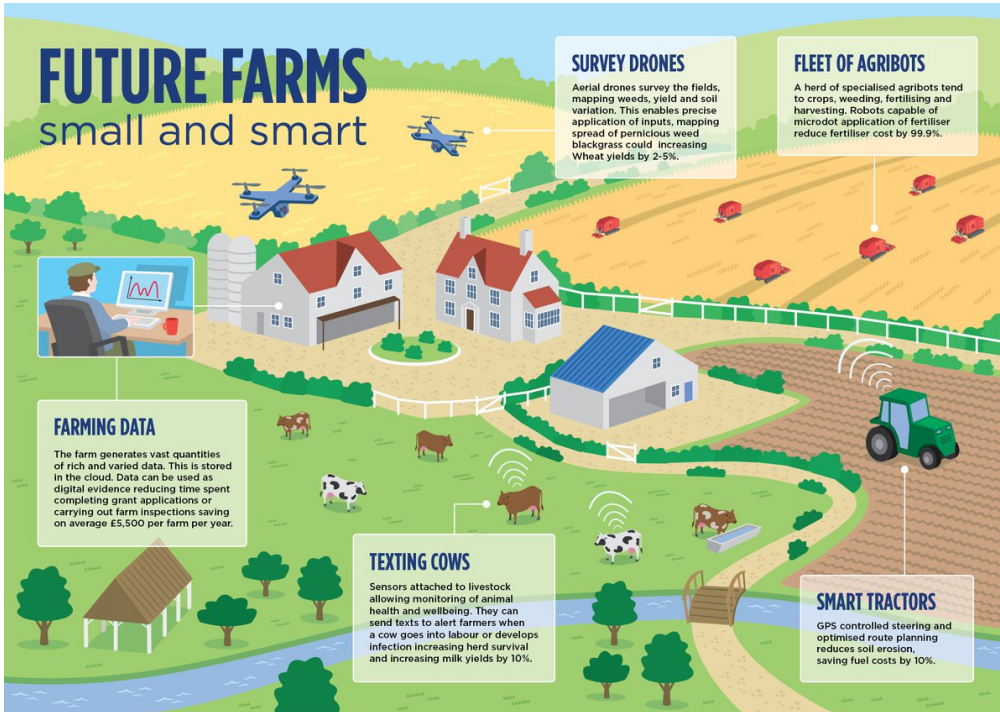
Deep CNNs to visually process board status in plays

https://www.slideshare.net/SanFengChang/mastering-the-game-of-go-with-deep-neural-networks-and-tree-search

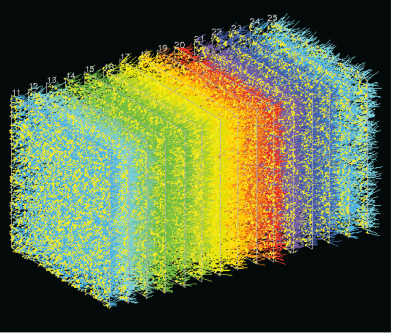
Innumerable “enterprise” applications

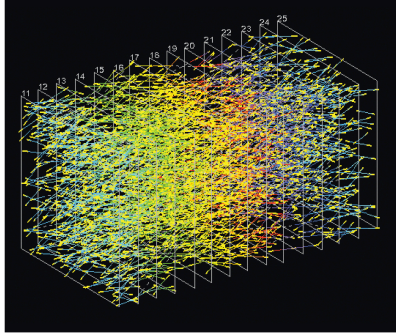

Enter

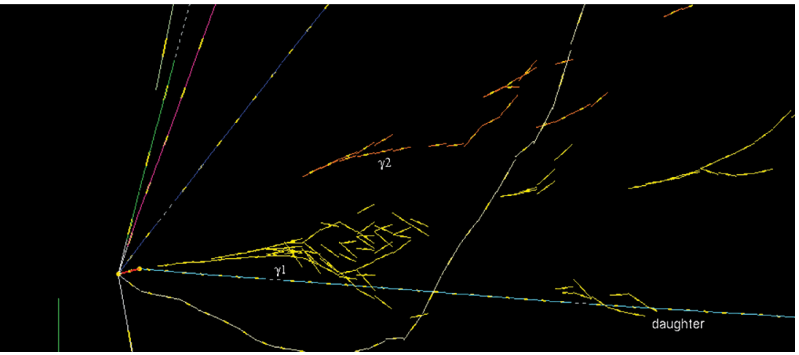

I ♥ INSURANCE
FLO

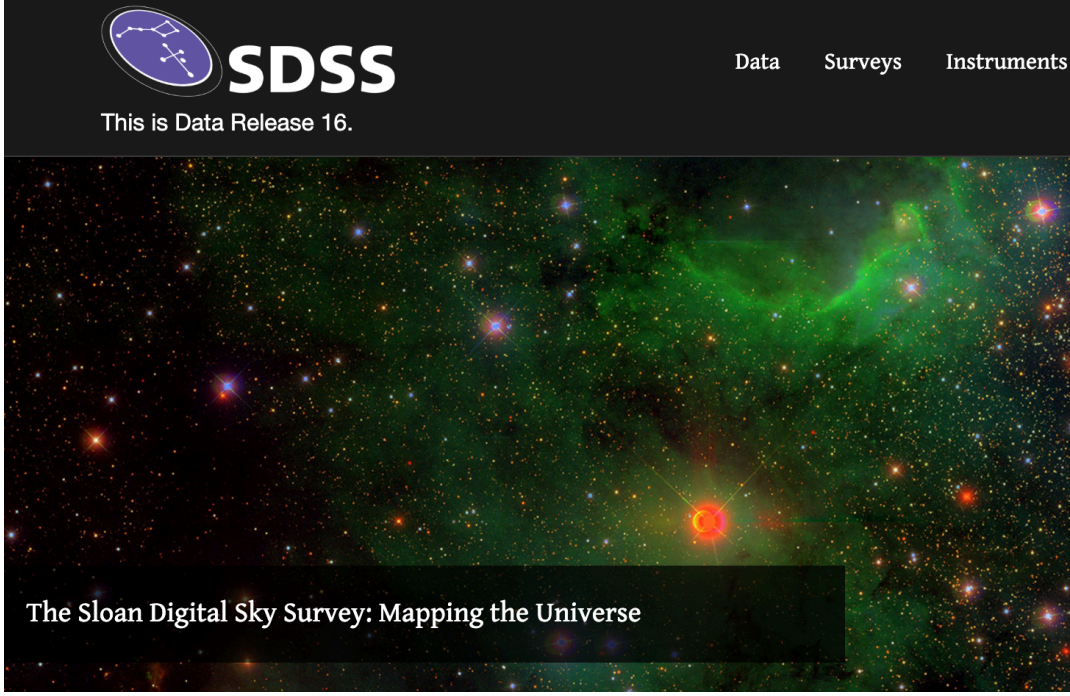

FUTURE FARMS
small and smart
SURVEY DRONES
Aerial drones survey the fields, mapping weeds, yield and soil variation. This enables precise application of inputs, mapping spread of pernicious weed blackgrass could increasing Wheat yields by 2-5%.
FLEET OF AGRIBOTS
A herd of specialised agribots tend to 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

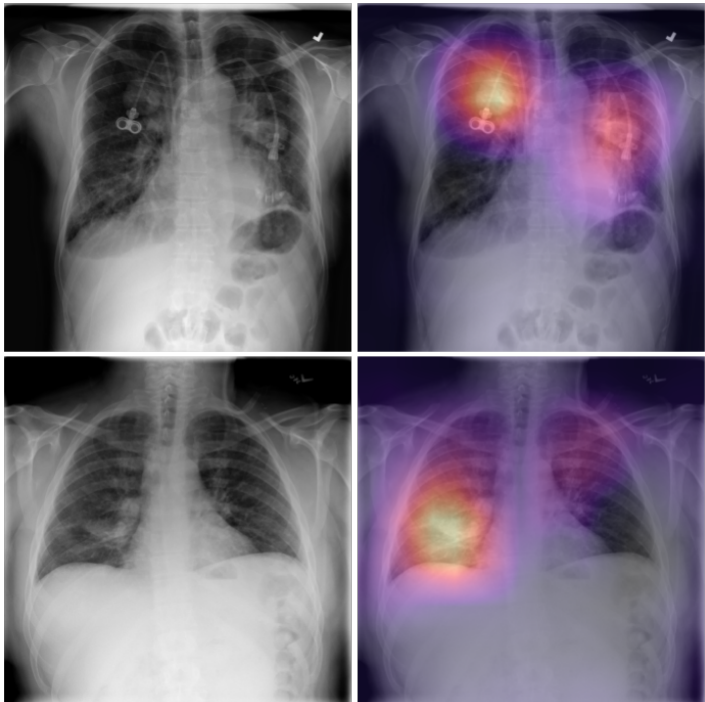





γ2
γ1
daughter


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
2361
2295
2230
2165
2095
2030
1965
1895



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—

Data data everywhere,

All the wallets did shrink!

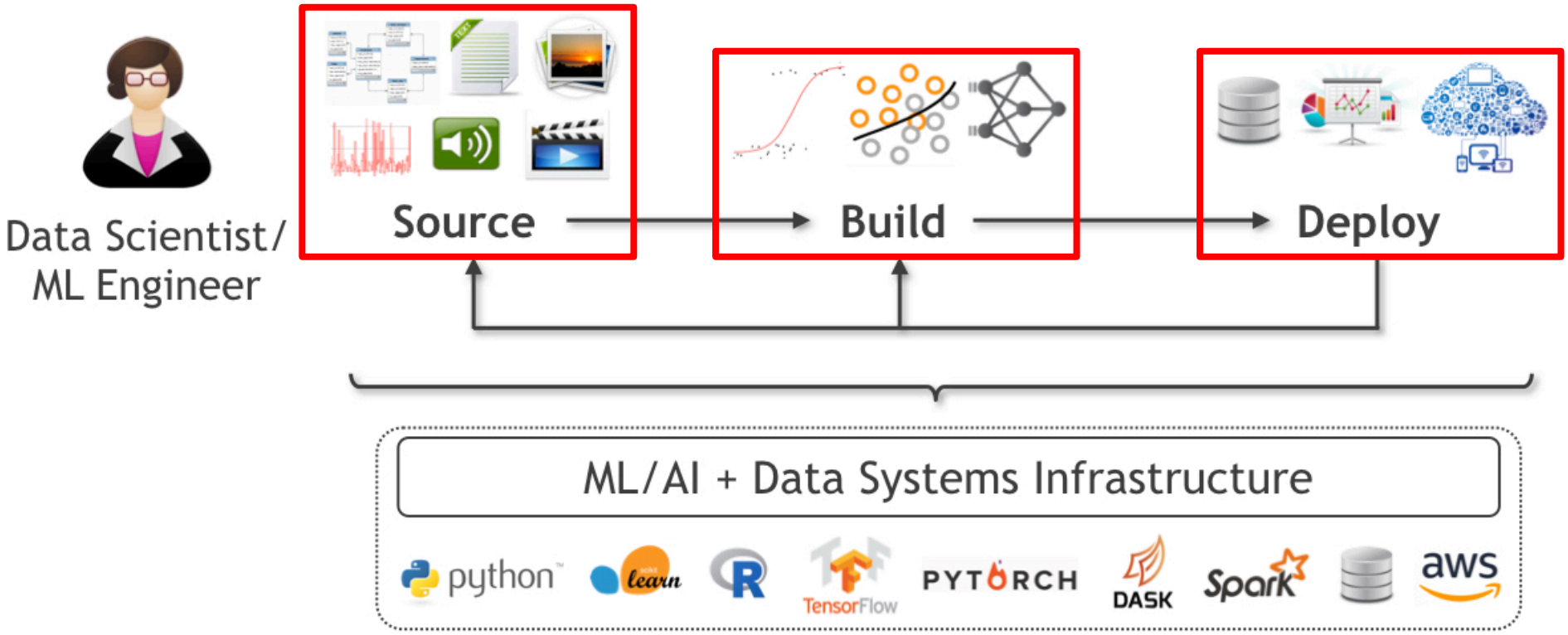
Data data everywhere,

Nor any moment to think?

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

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"**:
   - 3.1. **Source**: Data acquisition & preparation for ML
   - 3.2. **Build**: Model selection & deep learning systems
   - 3.3. **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 vs Scalable ML system (works on larger-than-memory datasets)
  2. **Target Workloads:** General ML library vs Decision tree-oriented vs Deep learning, etc.
  3. **Implementation Reuse:** Layered on top of scalable data system vs Custom from-scratch framework

Major Existing ML Systems

General ML libraries:

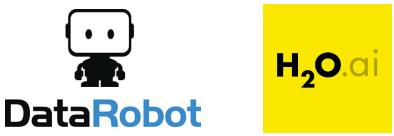
In-memory:

Disk-based files:

Layered on RDBMS/Spark:

Cloud-native:

"AutoML" platforms:



Decision tree-oriented:

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?

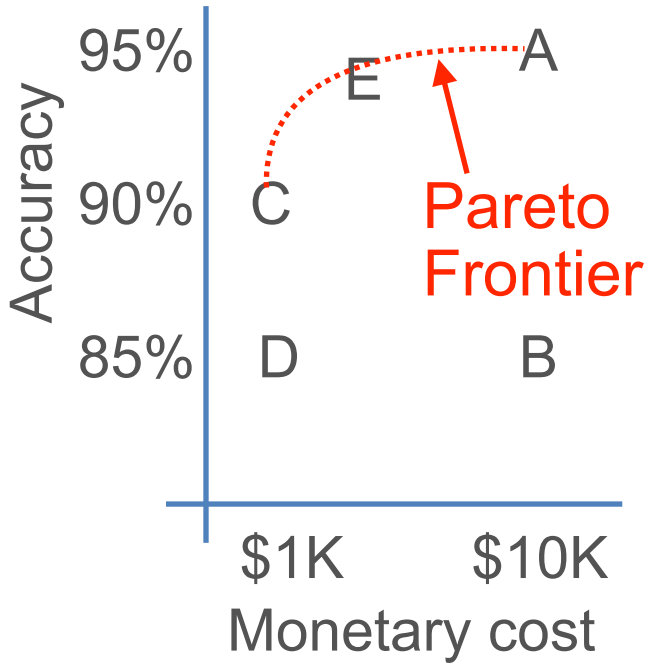
Conceptual System Stack Analogy

	Relational DB Systems	ML Systems
Theory	First-Order Logic Complexity Theory	Learning Theory Optimization Theory
Program Formalism	Relational Algebra	Matrix 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, cloud computing, and containerization.
- *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 deep learning inference with TensorFlow and Keras.
- *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, efficiency, 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*

Now for the (boring) logistics …

Prerequisites

- **DSC 100** (or equivalent) is necessary
- Transitively **DSC 80**; a mainstream ML algorithmics course is necessary
- Proficiency in Python programming
- For all other cases, email the instructor with proper justification; a waiver can be considered

http://cseweb.ucsd.edu/~arunkk/dsc102_winter22

Components and Grading

- **3 Programming Assignments: 35%** (7% + 14% + 14%)
  - No late days! Plan your work well ahead.
- **Best 5 of 6 Surprise Quizzes: 15%**; in-class using iClicker
- **Midterm Exam: 15%**
  - **Wed, Feb 9**; in-class (50min)
- **Cumulative Final Exam**: **35%**
  - **Mon, Mar 14**; 180min
- All quizzes and exams are *in-person* only; plan accordingly
- LMK ahead of time if you need makeup exam slot

http://cseweb.ucsd.edu/~arunkk/dsc102_winter22

Grading Scheme

Hybrid of relative and absolute; grade is better of the two

Grade	Relative Bin (Use strictest)	Absolute Cutoff (>=)
A+	Highest 5%	95
A	Next 10% (5-15)	90
A-	Next 15% (15-30)	85
B+	Next 15% (30-45)	80
B	Next 15% (45-60)	75
B-	Next 15% (60-75)	70
C+	Next 5% (75-80)	65
C	Next 5% (80-85)	60
C-	Next 5% (85-90)	55
D	Next 5% (90-95)	50
F	Lowest 5%	< 50

Example: Score 82 but 33%le; Rel.: B-; Abs.: B+; so, B+

Tentative Course Schedule

Week	Topic
1-3	Basics of Computer Org. and Operating Systems
4	Basics of Cloud Computing
5-6	Parallel and Scalable Data Processing: Parallelism
	Midterm Exam on Wed, Feb 9
7-8	Parallel and Scalable Data Processing: Scalability
8	Dataflow Systems
9-10	ML Data Sourcing
10	ML Model Building Systems
10	ML Deployment
11	**Final Exam on Wed, Mar 17**

Systems Principles

Scalability Principles

Scalable Analytics Systems

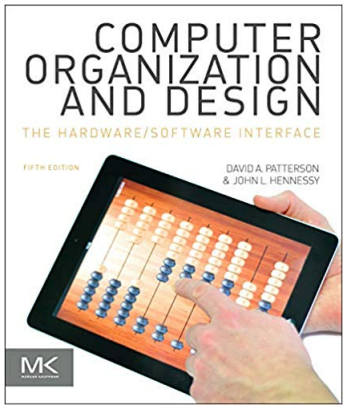
There will likely be 2 industry guest lectures; speakers TBD

Tentative Schedule for Prog. Assignments

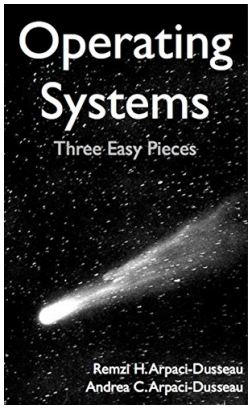
Date	Agenda
Jan 12	PA 0 released
Jan 14	Discussion on PA 0 by TA
Jan 25	**PA 0 due**
Jan 26	PA 1 released
Jan 28	Discussion on PA 1 by TA
Feb 16	**PA 1 due**
Feb 17	PA 2 released
Feb 18	Discussion on PA 2 by TA
Mar 8	**PA 2 due**

No late days! Plan your work upfront!
Try to join the Discussion slot talks by the TAs.

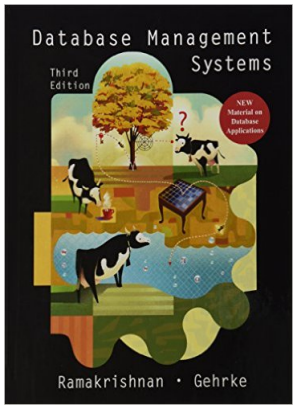
Suggested Textbooks



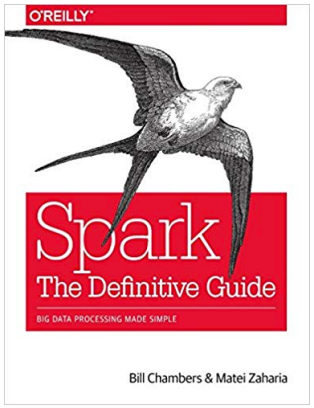
Aka “CompOrg Book”



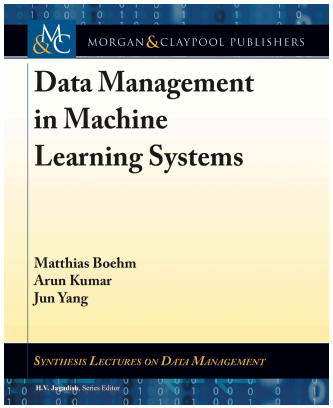
Aka “Comet Book”



Aka “Cow Book”



Aka “Spark Book”

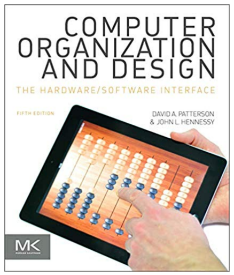


Aka “MLSys Book”

(Free PDFs available online; also check out our library)

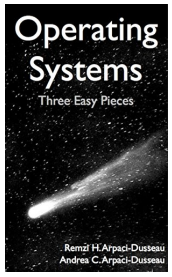
Why so many textbooks?!

1. Computer systems are about carefully layering *levels of abstraction*.

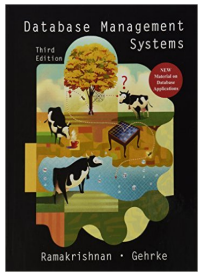


Hardware

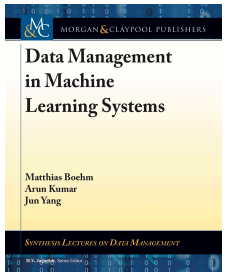
Low-level systems software

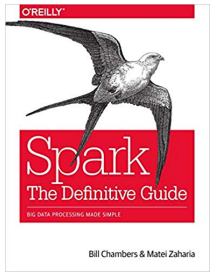


Higher-level relational dataflows



ML-oriented dataflows & lifecycle





More general scalable dataflows

2. Analytics/ML Systems is a recent/emerging area of research.
3. Also, DSC 102 is the first UG course of its kind in the world!

Course Administrivia

- **Lectures**: **MWF 1:00pm-1:50pm PT**
  - Virtual-only for weeks 1 and 2; in-person only afterward
  - Will take live Q&A throughout
- **Instructor**: Arun Kumar; arunkk [at] eng.ucsd.edu
  - Office hours: **Wed 2:00-3:00pm PT**
- See **Piazza** (or Canvas) for all Zoom meeting links, logistical announcements; see Canvas for gradebook
- **TAs**: Umesh Singla, Vignesh Nandakumar, and Pradyumna Sridhara (see webpage for details)

http://cseweb.ucsd.edu/~arunkk/dsc102_winter22

General Dos and Do NOTs

Do:

- Follow all announcements on Piazza/Canvas
- Try to join the lectures/discussions live
- View/review videos asynchronously by yourself
- Participate in discussions in class / on Piazza
- Raise your hand before speaking
- Use “DSC102:” as subject prefix for all emails to me/TA

Do NOT:

- Record lectures on your side without permission of instructor and other students
- Harass, intimidate, or intentionally talk over others
- Violate academic integrity on the graded quizzes, exams, or programming assignments; I am *very strict* on this matter!