UC San Diego

DSC 102
Systems for Scalable Analytics

Spring 2021

Rod Albuyeh

About Me

2016: PhD Political Science at USC

- emphasis in econometrics and experimental methods

2016-2019: Senior Data Scientist at Intuit

2019-2020: Senior Manager, Data Science at Oportun

2020-Present: Principal Data Scientist at Figure

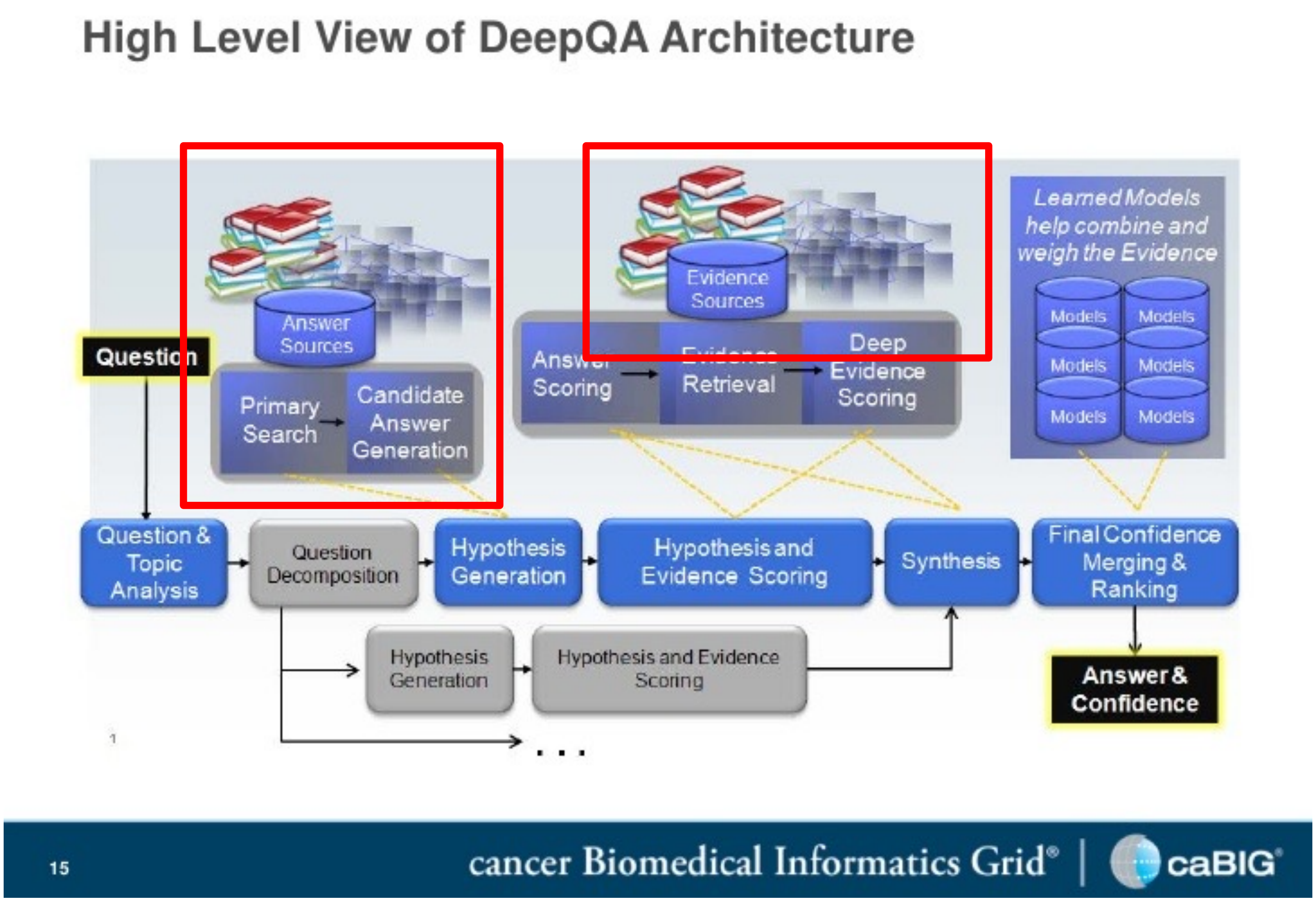
Specialties: structured time-series data, big data, anomaly detection, fraud, credit risk, direct marketing, cloud infrastructure (“for a data scientist”)

What is this course about? Why take it?

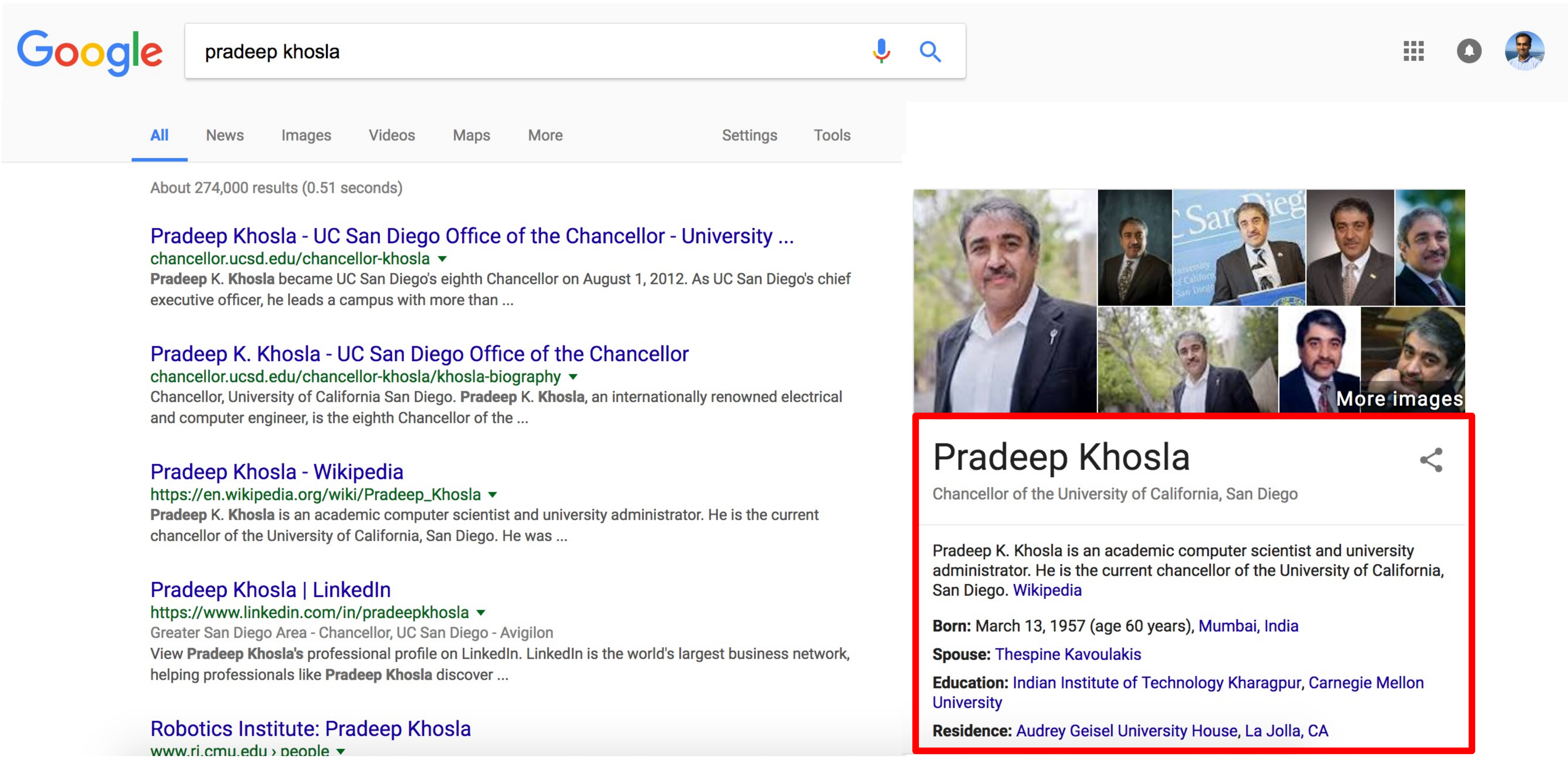
1. IBM's Watson wins Jeapordy!

How did Watson achieve that?

Watson devoured LOTS of data!

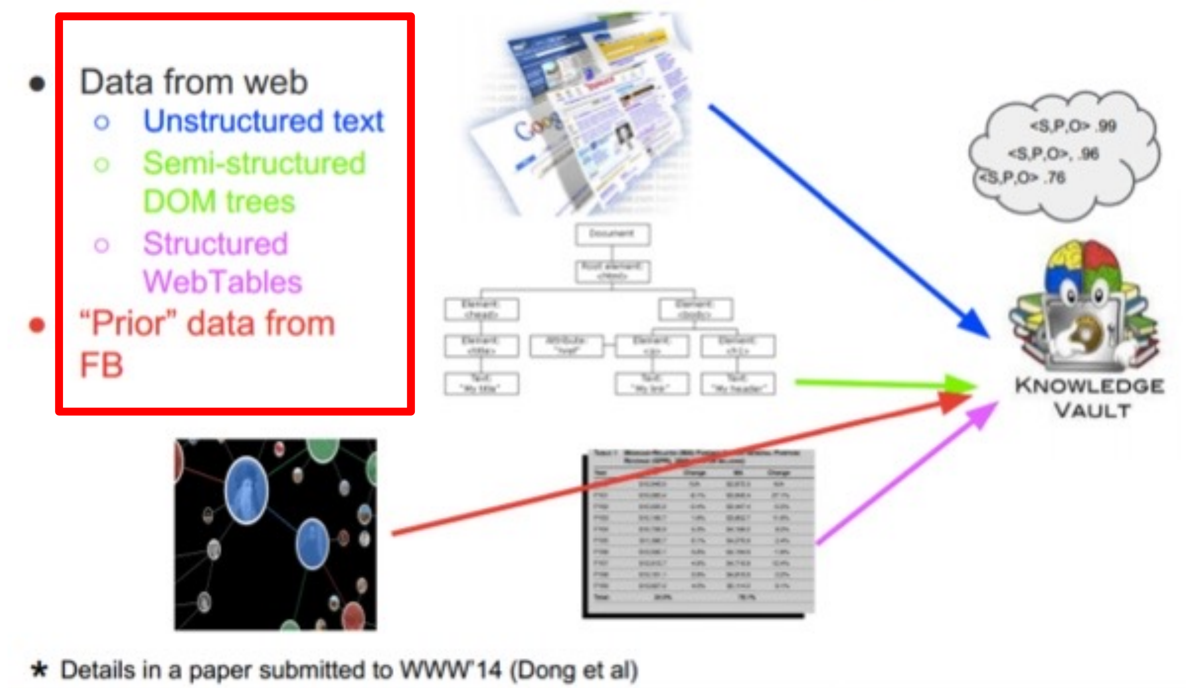


2. "Structured" data with search results

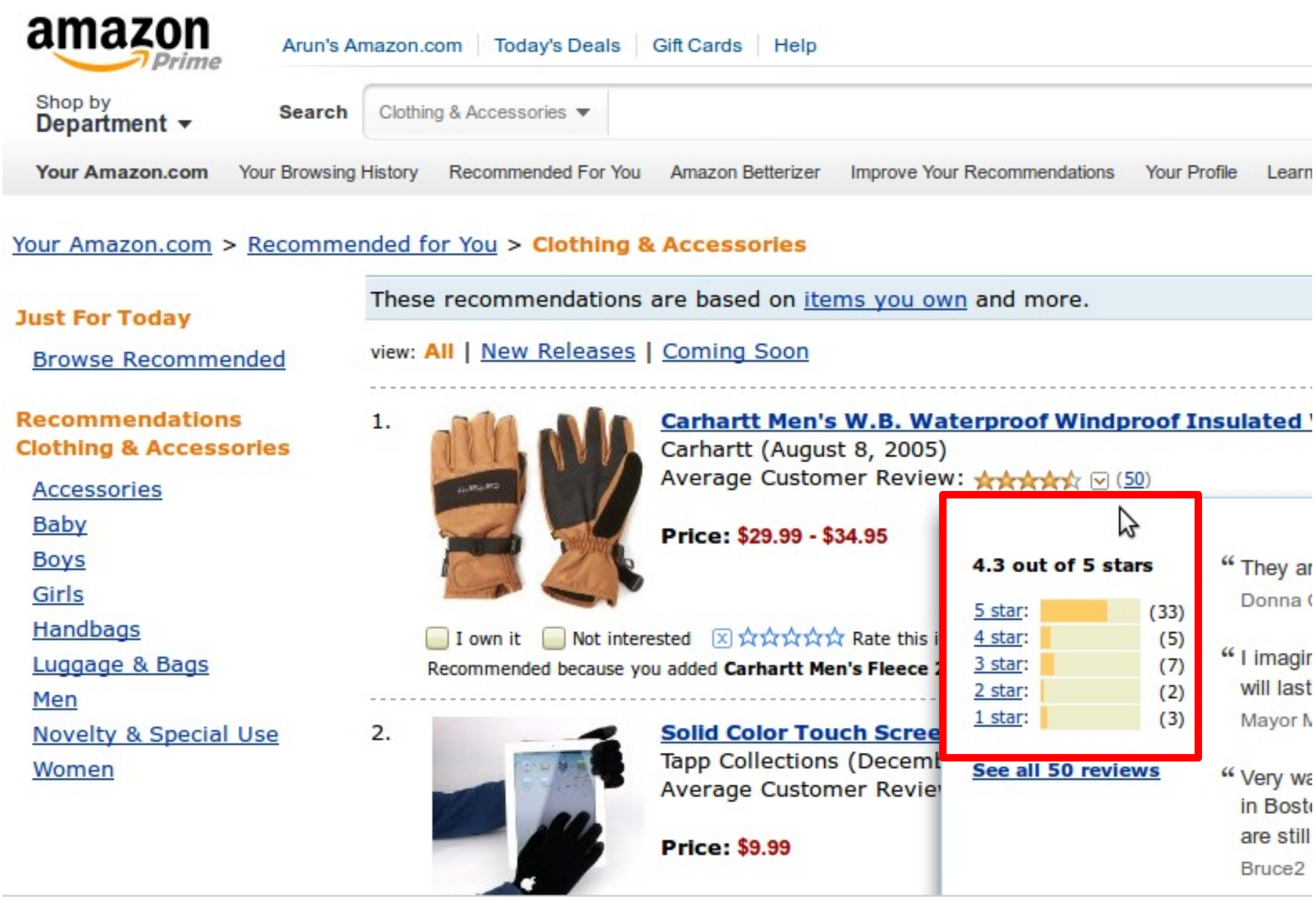


How does Google know that?

Google also devours LOTS of data!

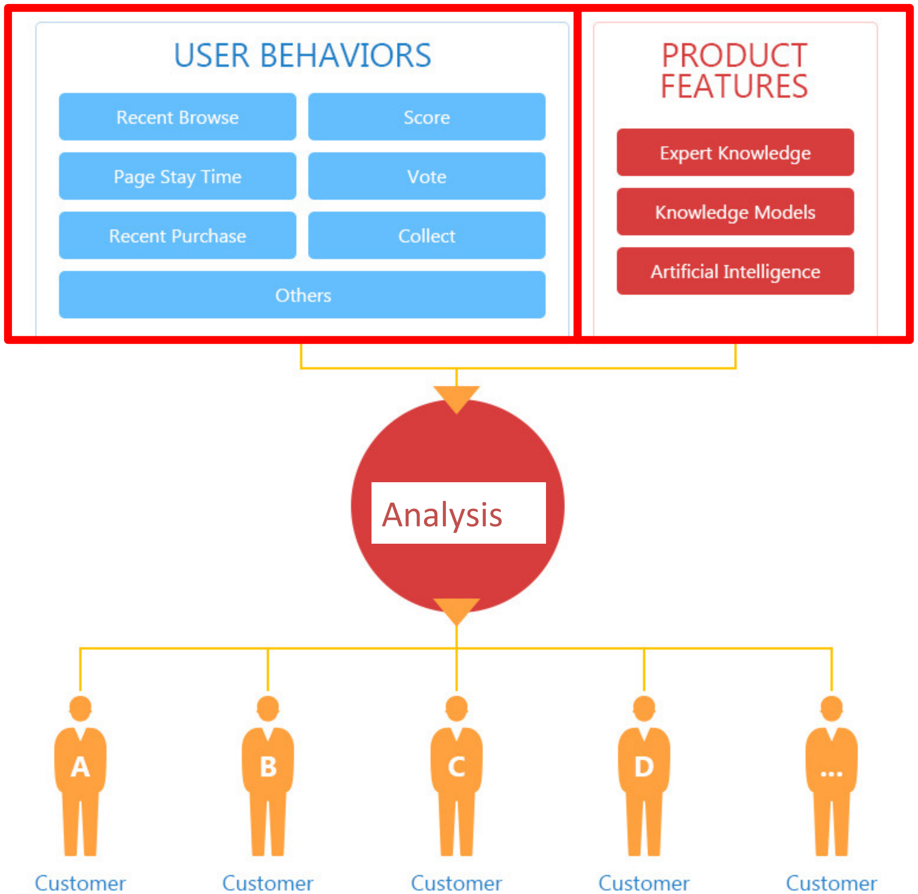


3. Amazon's "spot-on" recommendations



How does Amazon know that?

You guessed it! LOTS and LOTS of data!



And innumerable “traditional” applications


FEDERAL RESERVE


I ♥ INSURANCE
FLO

Scalable software systems for data management and analytics are the cornerstone of many digital applications, both modern and traditional

The Age of “Big Data”/“Data Science”

The New York Times

SundayReview | NEWS ANALYSIS

The Age

By STEVE LOHR F

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, CON

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

SUMMARY SAVE SHARE COMMENT TEXT SIZE PRINT $8.95 BUY COPIES

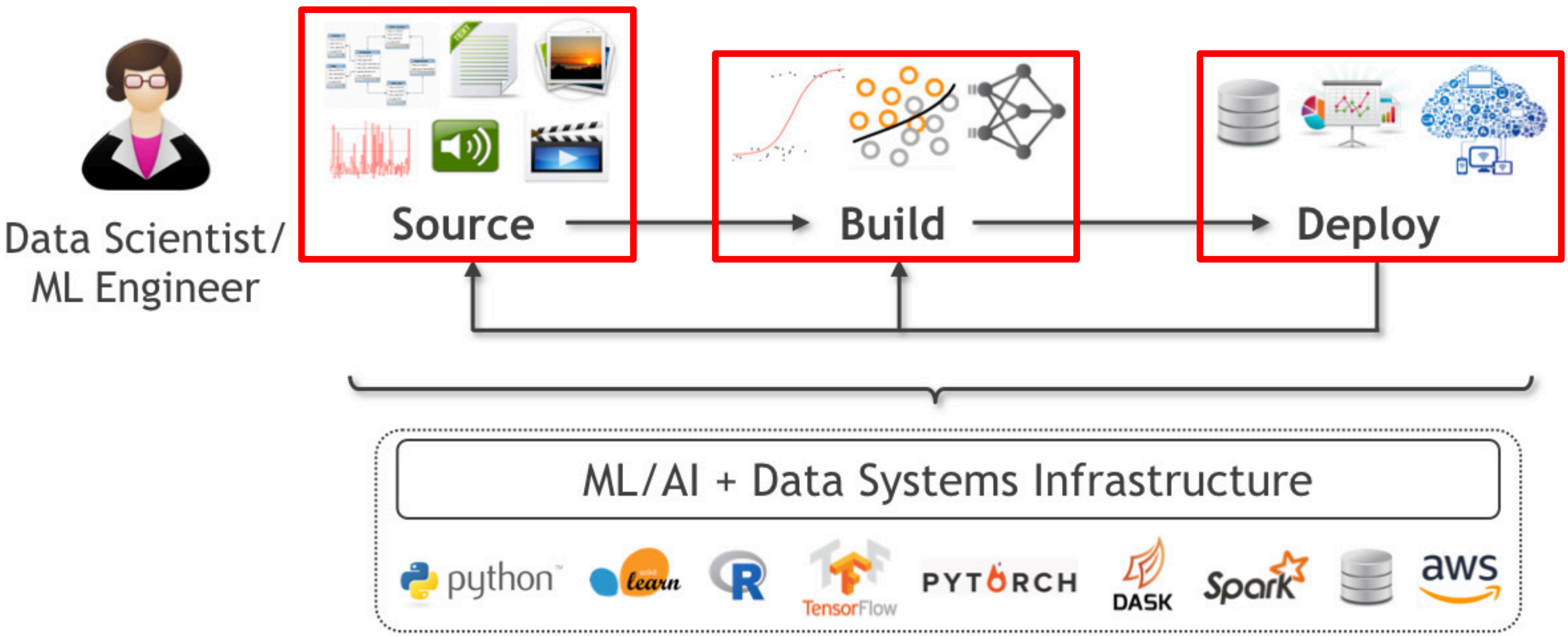
VE RI TAS Harvard Business Review

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–

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 compute efficiently over large data? What is cloud computing?
2. "**Scalability**": How to scale and parallelize data-intensive computations?
3. **Scalable Systems for "Analytics"**:
   1. **Source**: Data acquisition & preparation for ML
   2. **Build**: Dataflow & Deep Learning systems
   3. **Deploying** ML models
4. Hands-on experience with tools for scalable analytics

The Lifecycle of ML-based Analytics



Data acquisition
Data preparation

Feature Engineering
Training & Inference
Model Selection

Model Serving
Monitoring

Learning Outcomes of this course

- Understand the basic systems principles of the memory hierarchy, scalable data access, parallelism paradigms, cloud computing, and containerization.
- Identify the abstract data access patterns of, and opportunities for parallelism in, data processing and ML algorithms.
- Reason critically about practical tradeoffs between accuracy, efficiency, scalability, usability, and total cost.
- Learn the basics of dataflow ("Big Data") programming with HDFS, MapReduce, and Spark.
- Gain exposure to deep learning inference on unstructured data with TensorFlow and Keras.
- Apply SQL, dataflow programming, and DL inference for end-to-end pipelines for data preparation, feature engineering, and model selection on large-scale heterogeneous datasets.

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 how to use DBMSs or SQL for DB-backed applications (indexing, JDBC, triggers, etc.)
  - Take CSE 132B instead
- NOT a training module for how to use Spark
- NOT a course on internal details of RDBMSs/Spark
  - Take CSE 132C instead
- NOT a course on ML or data mining *algorithmics*; instead, we focus on ML *systems*

Advanced Analytics/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), i.e., beyond just SQL aggregates
  - Statistical analysis; ML, deep learning (DL); data mining (domain-specific applied ML + feature eng.)
  - *High-level APIs* for expressing statistical/ML/DL computations over large datasets

Background: ML 101

Generalized Linear Models (GLMs); from statistics

Bayesian Networks; inspired by causal reasoning

Decision Tree-based: CART, Random Forest, Gradient-Boosted Trees (GBT), etc.; inspired by symbolic logic

Support Vector Machines (SVMs); inspired by psychology

Artificial Neural Networks (ANNs): Multi-Layer Perceptrons (MLPs), Convolutional NNs (CNNs), Recurrent NNs (RNNs), Transformers, etc.; inspired by brain neuroscience

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

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

Can often trade off accuracy a bit to gain on the rest!

***Q:** How does it fit within production systems and workflows?*

***Q:** How are the features and models configured?*

***Q:** What if the dataset is larger than single-node RAM?*

***Q:** How to simplify the implementation of such 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	Declarative Query Language	TensorFlow? R? Scikit-learn?
Program Modification	Query Optimization	???
Execution Primitives	Parallel Relational Operator Dataflows	Depends on ML Algorithm
Hardware	CPU, GPU, FPGA, NVM, RDMA, etc.	

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: scikit learn, R

Disk-based files: SAS, DASK

Layered on RDBMS/Spark: MADlib, Apache Spark MLlib

Cloud-native: Azure Machine Learning, Amazon SageMaker

"AutoML" platforms: DataRobot, H_2O.ai

Decision tree-oriented:

dmlc XGBoost, Microsoft LightGBM

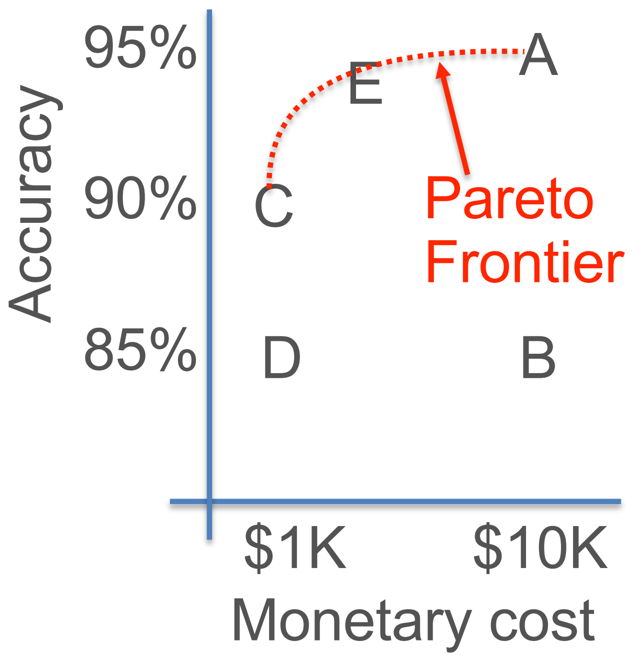
Deep learning-oriented:

TensorFlow, PYTORCH

Pareto Surfaces in Real-World ML

***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 data scientists 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.

And now for the (boring) logistics …

Prerequisites

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

https://albuyeh.github.io/dsc102-spring-2021/

Course Administrivia

- **Lectures**: MonWedFri 11-11:50am, PCYNH 106
- **Instructor**: Rod Albuyeh; **ralbuyeh@ucsd.edu**

  Office hours: Mon 8-9am

  TA: Taruj Goyal; tgoyal@eng.ucsd.edu

  Discussions: Fri 1:00-1:50pm

https://albuyeh.github.io/dsc102-spring-2021/

Grading

- **Midterm Exam**: **30%**

  Date: **Fri, April 30**; in-class (11:00-11:50am)

- **Programming Assignment**: **25%**
  - **I may adjust due to cold start.**
- **”Many” Surprise Quizzes: 5%**
- **Final Exam**: **40%** (cumulative)

  Date: **Fri, June 11**; 11:30am-2:30pm

Grading Scheme

Hybrid of relative and absolute; grade is <u>better</u> of the two

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

Tentative Course Schedule

Week	Topic	References
1-4	Basics of Computer Organization and Operating Systems	Ch. 1, 2.1-2.3, 2.12, 4.1, and 5.1-5.5 of CompOrg Book; Ch. 2, 4.1-4.2, 6, 7, 13, 14.1, 18.1, 21, 22, 26, 36, 37, 39, and 40.1-40.2 of Comet Book
4	Basics of Cloud Computing	-
5-6	Parallel and Scalable Data Processing: Parallelism Basics	Ch. 9.4, 12.2, 14.1.1, 14.6, 22.1-22.3, 22.4.1, 22.8 of Cow Book; Ch. 5, 6.1, 6.3, 6.4 of MLSys Book
7	Parallel and Scalable Data Processing: Scalable Data Access	-
8	Parallel and Scalable Data Processing: Data Parallelism	-
9	Dataflow Systems	Ch. 2.2 of MLSys Book
10	ML Data Sourcing	Ch. 8.1, 8.3 of MLSys Book
10	ML Model Building Systems	Ch. 8-8.4 of MLSys Book
11	Review for Final	-

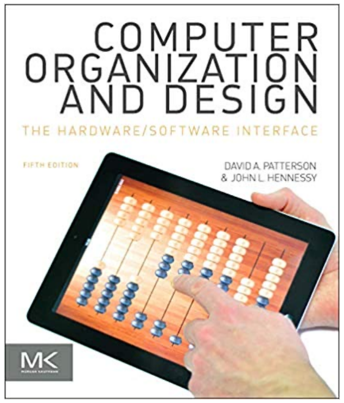
Tentative Schedule for Prog. Assignments

Date	Agenda
Fri, Apr 30	Midterm Exam
Fri, Apr 30	PA released
Fri, May 28	PA due
Fri, June 11	Final Exam

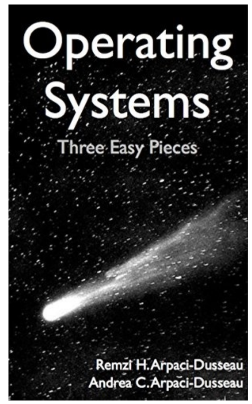
Guest Lectures

We have a slate of guest lectures from industry mixed in. Material from their talks will be fair game for the midterm final. We will have review sessions on this. Keep an eye on the website for updates. In many cases, guest lectures will be async and I will release and hold “watch parties” during class.

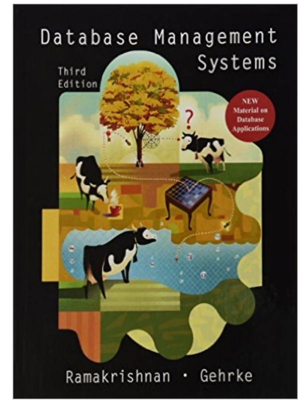
Suggested Textbooks



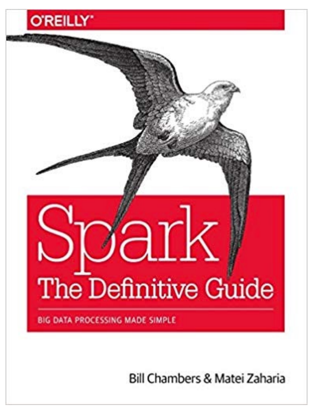
Aka “CompOrg Book”



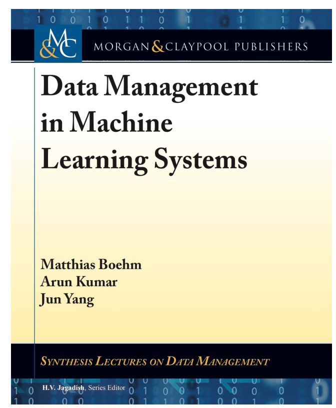
Aka “Comet Book”



Aka “Cow Book”



Aka “Spark Book”



Aka “MLSys Book”

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