

DSC 102

Systems for Scalable Analytics

Fall 2023

Rod Albuyeh

About Me



2016: PhD in Political Science from USC

Intuit **Oportun**

FIGURE  **ResMed**

2016-2019: Senior Data Scientist at Intuit

2019-2020: Senior Manager, Data Science at Oportun

2020-2022: Principal Data Scientist at Figure

2022-2023: Machine Learning Architect at Resmed



2021-Present: Part-Time Lecturer at UCSD

2023-Present: Founder, Chief AI Scientist at Albell



My Current Academic Work

Deep learning for tabular data, the “last unconquered castle” of deep learning.

Tabular Data

columns = attributes for those observations

Rows = observations

Player	Minutes	Points	Rebounds	Assists
A	41	20	6	5
B	30	29	7	6
C	22	7	7	2
D	26	3	3	9
E	20	19	8	0
F	9	6	14	14
G	14	22	8	3
I	22	36	0	9
J	34	8	1	3

Reproducible enterprise-grade infrastructure for machine learning research.

Machine learning for political science research.

What is this course about? Why take it?

1. Netflix's “spot-on” recommendations

NETFLIX ORIGINAL **STRANGER THINGS**

95% Match 2017 2 Seasons 4K Ultra HD 5.1

When a young boy vanishes, a small town uncovers a mystery involving secret experiments, terrifying supernatural forces and one strange little girl.

Winona Ryder, David Harbour, Matthew Modine
TV Shows, TV Sci-Fi & Fantasy, Teen TV Shows



Popular on Netflix



Recently Watched



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

Google X 🔍 ⚙️

All Images Books News Videos More Tools

About 13,000,000 results (0.52 seconds)

 **Alan Turing**
Mathematician ⋮

Overview Education Books Videos

     More images

[https://en.wikipedia.org › wiki › Alan_Turing](https://en.wikipedia.org/wiki/Alan_Turing) ⋮

Alan Turing - Wikipedia

Alan Mathison Turing OBE FRS (/'tjøərɪŋ/; 23 June 1912 – 7 June 1954) was an English mathematician, computer scientist, logician, cryptanalyst, philosopher, ...

Partner(s): [Joan Clarke](#); (engaged in 194... Known for: [Cryptanalysis of the Enigm...](#)

Awards: Smith's Prize (1936) Resting place: Ashes scattered in gard...

[The Enigma](#) · [Alan Turing law](#) · [Legacy of Alan Turing](#) · [Alan Turing Year](#)

About

Alan Mathison Turing OBE FRS was an English mathematician, computer scientist, logician, cryptanalyst, philosopher, and theoretical biologist.

[Wikipedia](#)

Born: June 23, 1912, [Maida Vale, London, United Kingdom](#)

Died: June 7, 1954, [Wilmslow, United Kingdom](#)

Academic advisor: [Alonzo Church](#)

Education: Princeton University (1936–1938), [MORE](#)

Influenced by: [Alonzo Church](#), [Kurt Gödel](#), [Ludwig Wittgenstein](#), [Max Newman](#)

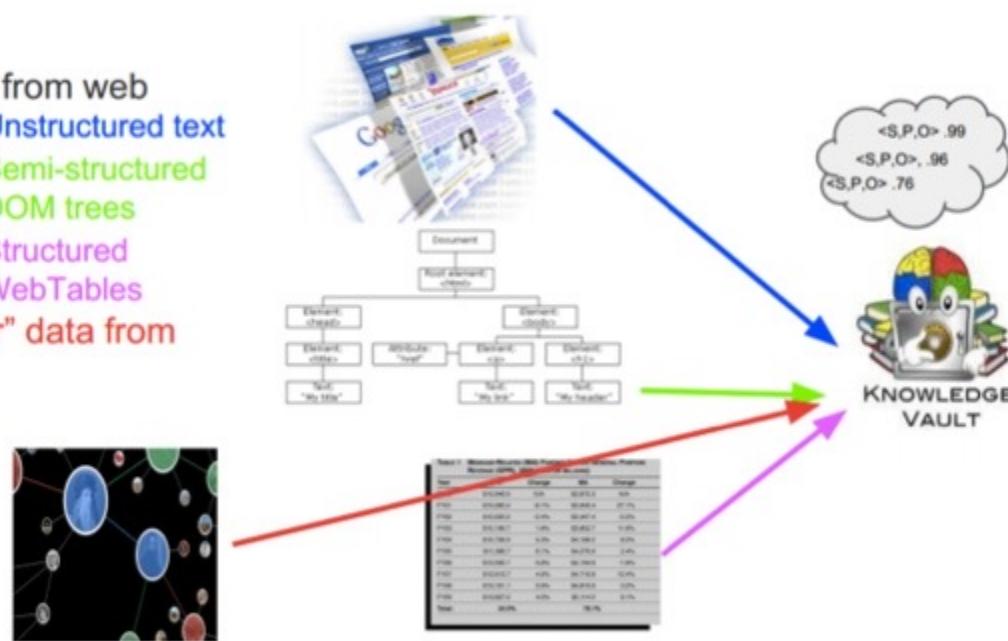
Notable students: [Robin Gandy](#), [Beatrice Worsley](#)

How does Google know that?

Large datasets + Machine learning!

Knowledge Vault* fuses all these signals together

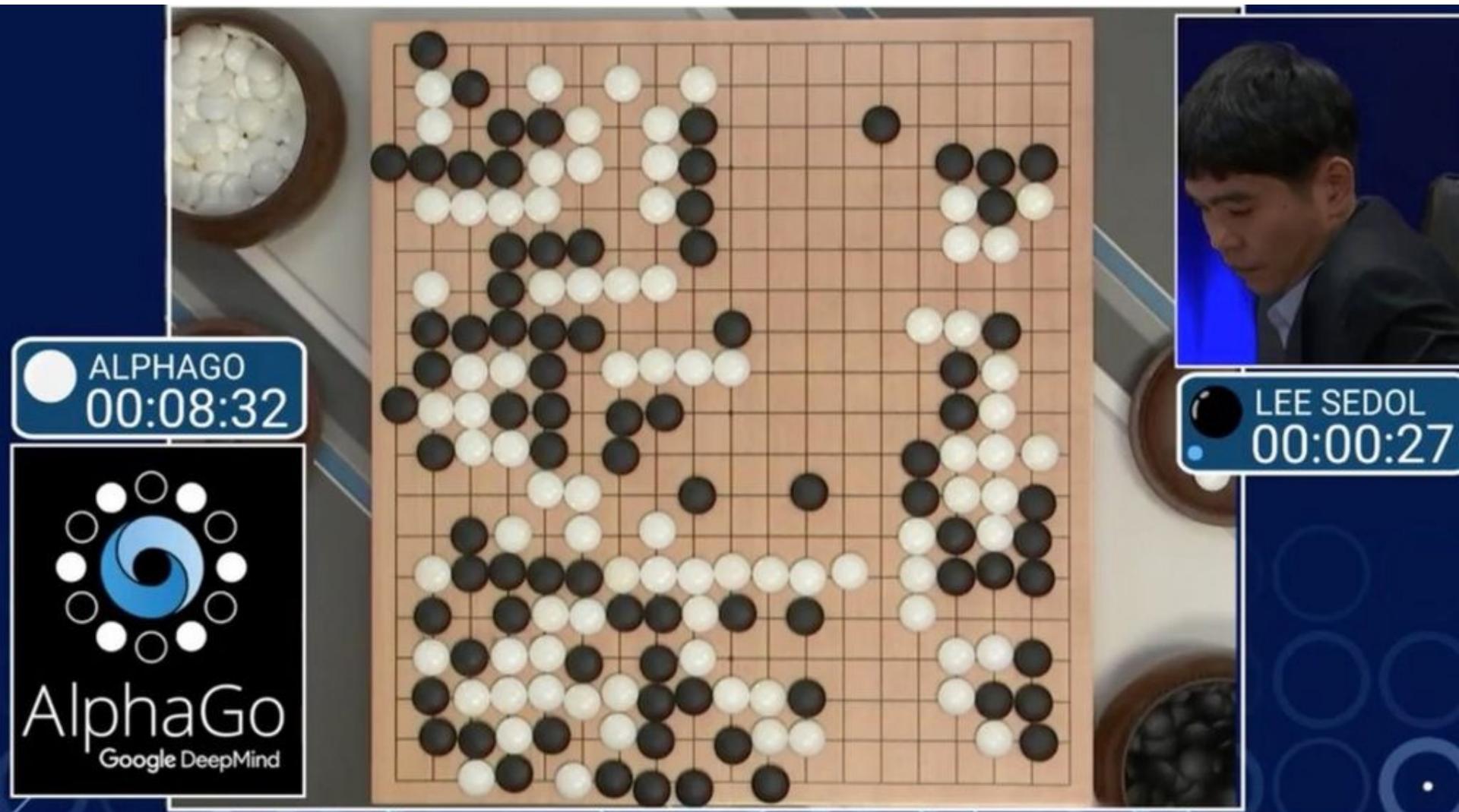
- Data from web
 - Unstructured text
 - Semi-structured DOM trees
 - Structured WebTables
- “Prior” data from FB



* Details in a paper submitted to WWW'14 (Dong et al)

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

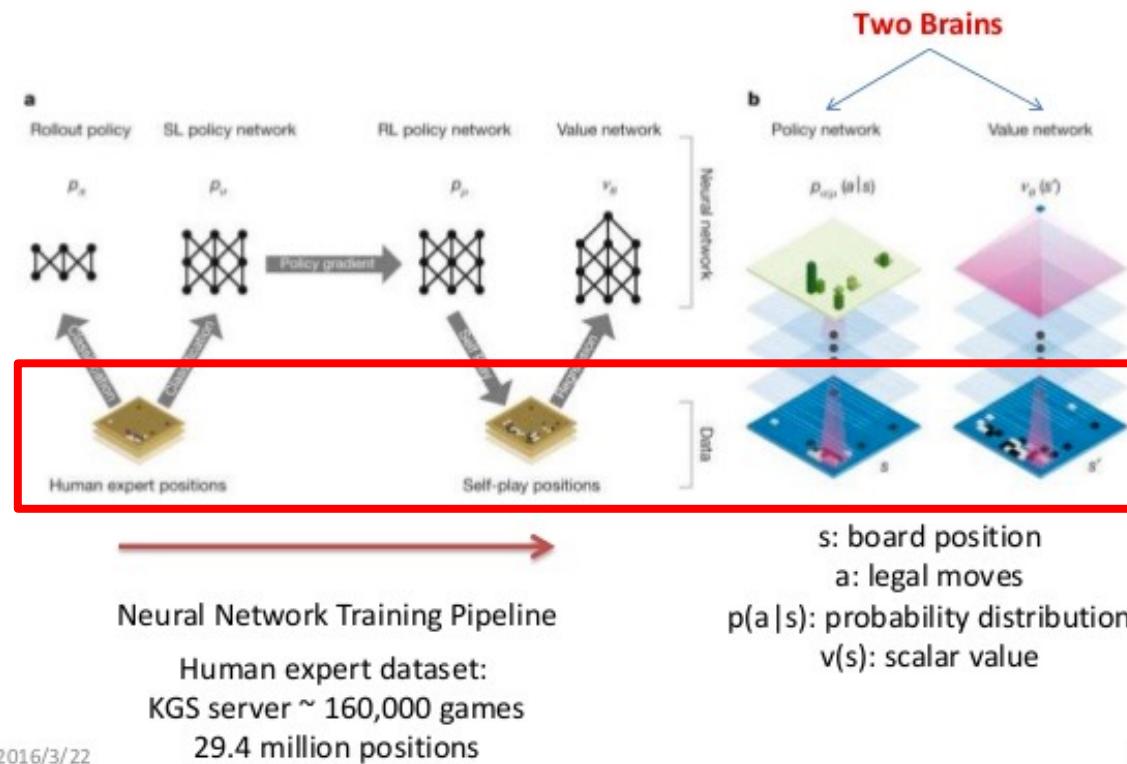
3. AlphaGo defeats human champion!



How did AlphaGo achieve that?

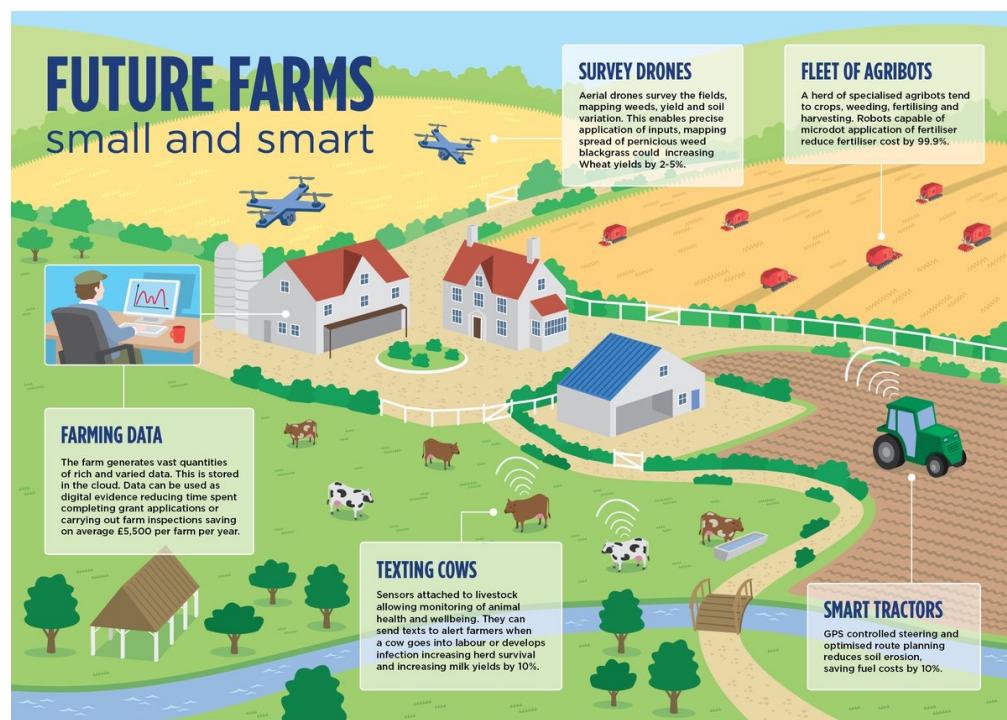
Breakthrough powered by deep learning!

Architecture of AlphaGo

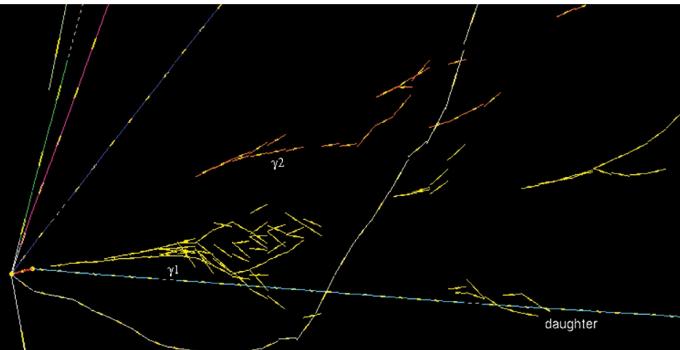
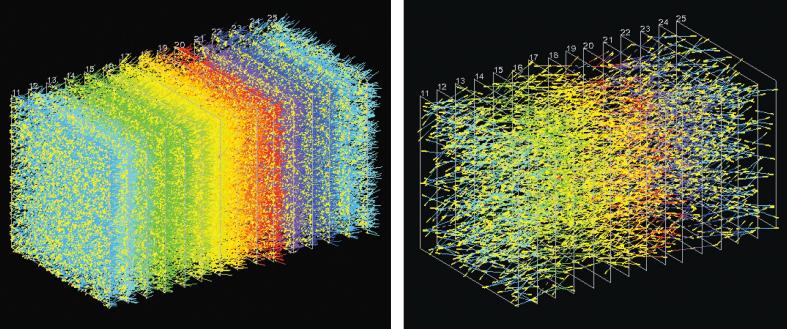


Deep CNNs to visually process board status in plays

Innumerable “enterprise” applications



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

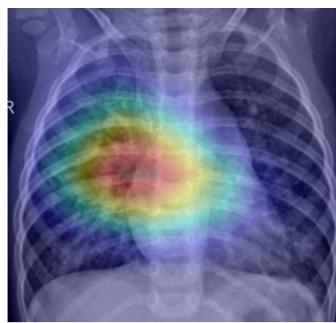


This is Data Release 16.

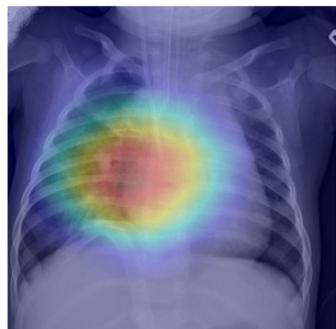
Data Surveys Instruments



(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 **Forbes** / Entrepreneurs

By STEVE LOHR F

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

Email

Share

Tweet

Save

Forbes

Drowning In Big Data - Finding Insight In A Digital DATA Josh Steimle, CON **Data Scientist: The Sexiest Job of the 21st Century**

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

For roughly a decade, information about Big Data. The IDC industry will experience by 2018. What this



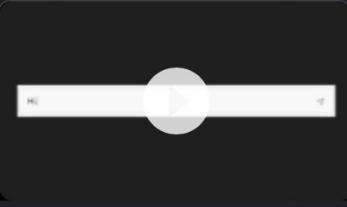
**Harvard
Business
Review**

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—

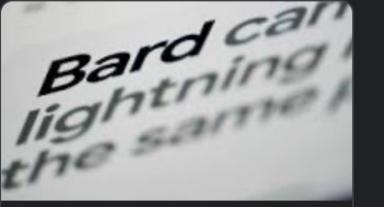
Emerging Age of LLMs



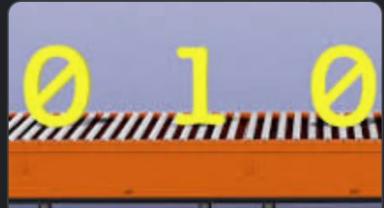
Fox Business
ChatGPT: Who and what is behind the artificial intelligence tool changing the tech...
9 hours ago



The Verge
ChatGPT started a new kind of AI race — and made text boxes cool again
1 day ago



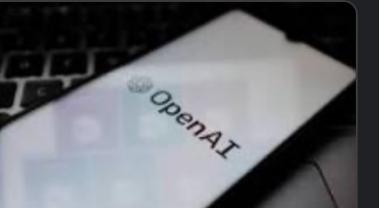
Yahoo Finance
What can Google's AI-powered Bard do? We tested it for you
29 mins ago



Axios
ChatGPT and generative AI are changing the software-making game
7 hours ago

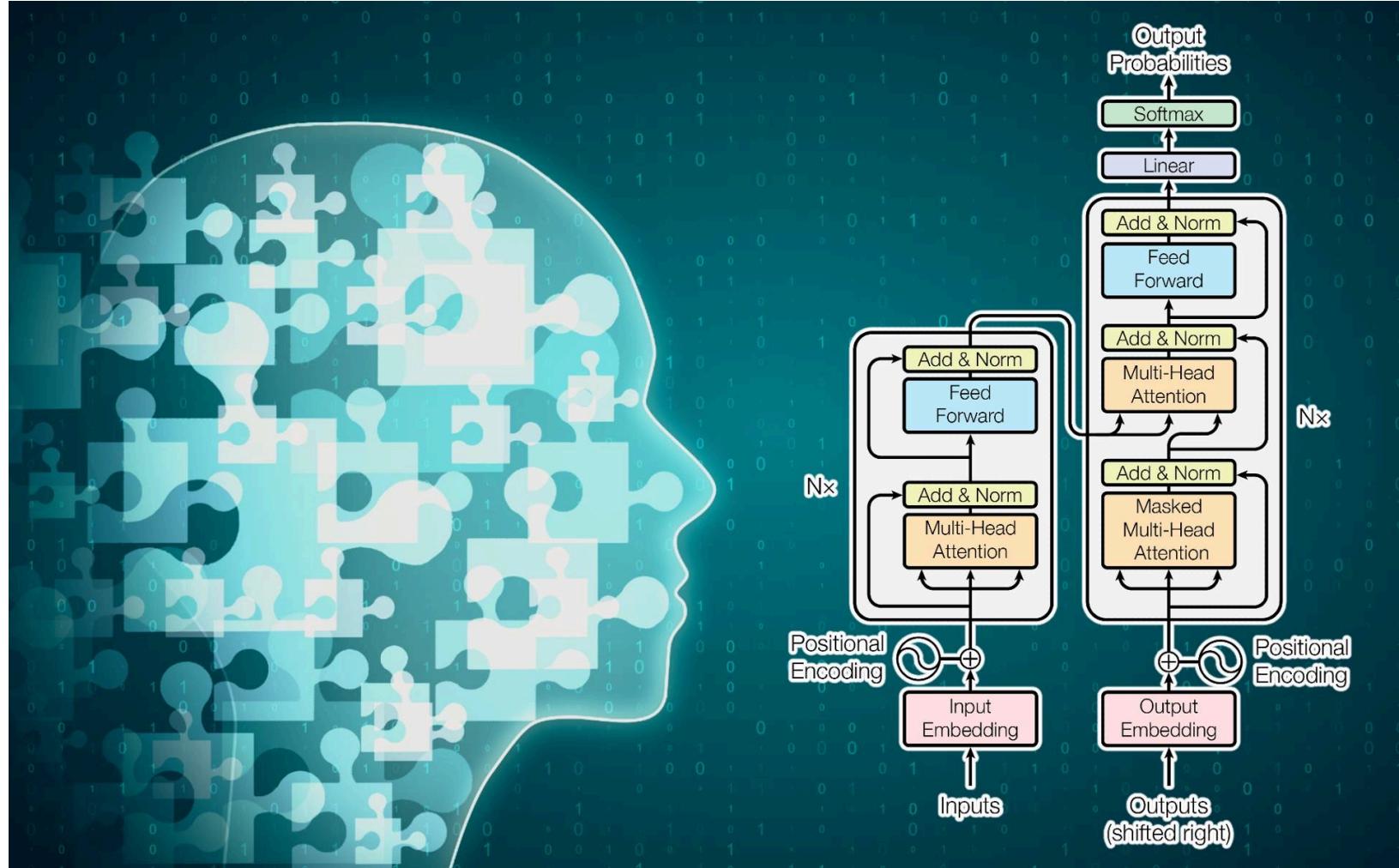


Search Engine Journal
New Open Source ChatGPT Clone - Called Dolly
7 hours ago



Engadget
ChatGPT's new plugins will deliver real-time stats
4 days ago

LLMs and Transformers

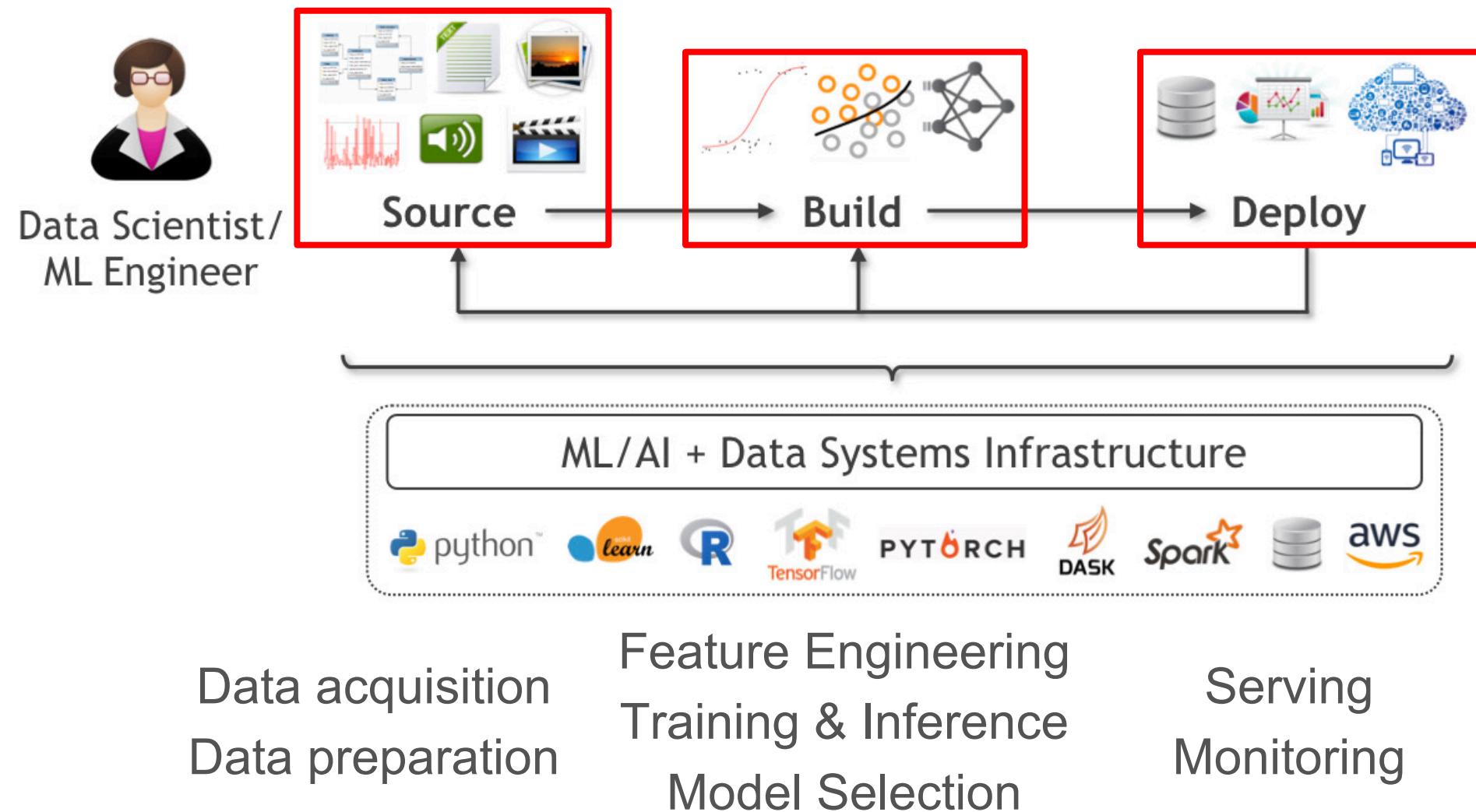


Many researchers find that they run out of hardware before they manage to overfit. 21

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



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:



TensorFlow



Data Systems Concerns in ML

Key concerns in ML:

Q: How do “ML Systems” relate to ML?

Runtime efficiency (sometimes)

Additional key *practical* concerns in ML Systems:
ML Systems : ML :: Computer Systems : TCS

Scalability (and **efficiency** at scale)

Usability

Manageability

Developability

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

Q: Q: What if I didn't have the discipline to take my ideas from the PPT to code?

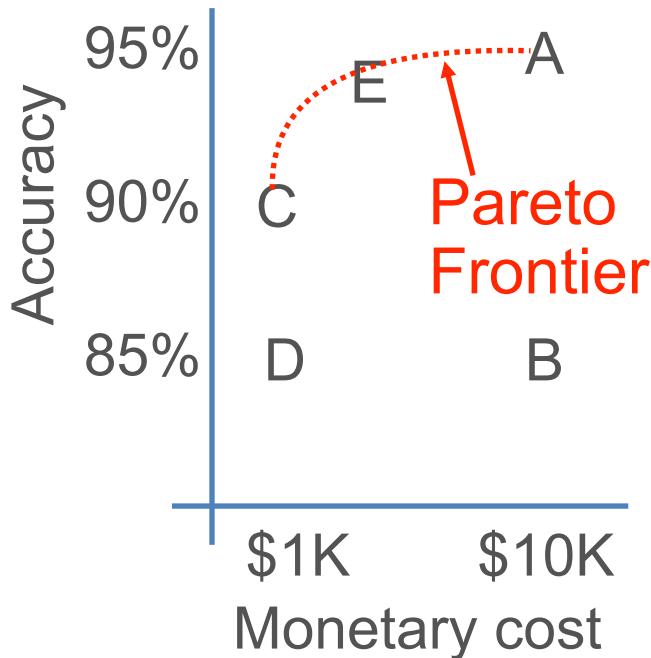
Conceptual System Stack Analogy

	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, others...
Program Modification	Query Optimization	Model optimization, tuning, regularization
Execution Primitives	Parallel Relational Operator Dataflows	Parallel computing primitives
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.

After this course, you'll be able to:

- ❖ **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 (but our suggested Spark textbook is excellent for that)
- ❖ NOT a course on ML or data mining *algorithmics*; instead, we focus on ML *systems*

Now for the course logistics ...

Prerequisites

- ❖ **DSC 100** (or equivalent) is necessary
- ❖ Transitively **DSC 80**; a mainstream ML algorithmics course is necessary
- ❖ Proficiency in Python programming – some familiarity with Linux command line is also helpful, but not required.
- ❖ For all other cases, email me with proper justification; a waiver can be considered

Course website is listed in Canvas.

Components and Grading

- ❖ **3 Programming Assignments: 40% (8% + 16% + 16%)**
 - ❖ No late days! Plan your work well ahead.
- ❖ **Midterm Exam: 15%**
 - ❖ **Wed, Nov 2**; in-class only (50min)
- ❖ **Cumulative Final Exam: 35%**
 - ❖ **Fri, Dec 9**; Canvas Quiz only (3hrs long but 4hrs limit)
- ❖ **10 (of 12) Peer Instruction Activities: 10%**
 - ❖ Bring iClicker to every lecture
- ❖ **Extra Credit Activities: 4% (likely)**
- ❖ LMK ahead of time if you need makeup exam slot

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

Tentative Course Schedule

Week	Topic
Week 4	Basics of Machine Resources: Computer Organization Basics of Machine Resources: Operating Systems Basics of Cloud Computing
Week 4-5	Parallel and Scalable Data Processing: Parallelism Basics Midterm Exam on Thursday, May 11 – in class
Week 6-7	Parallel and Scalable Data Processing: Scalable Data Access
Week 7-8	Parallel and Scalable Data Processing: Data Parallelism
Week 9	Dataflow Systems
Week 10	ML Model Building Systems
Week 11	Final Exam on Fri, June 15, remote

There will be 2 industry guest lectures (maybe 3)

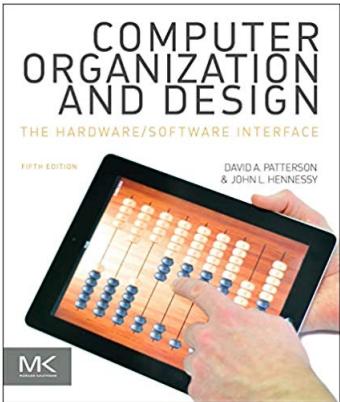
Programming Assignments

- ❖ **PA0: Setting up AWS and Dask**
 - ❖ April 10 to April 25
- ❖ **PA1: Data Exploration with Dask**
 - ❖ April 25 to May 16
- ❖ **PA2: Feature Eng. and Model Selection with Spark**
 - ❖ May 16 to June 9
- ❖ **Expectations on the PAs:**
 - ❖ Teams of 2 or 1 (individual); see webpage on academic integrity
 - ❖ I will cover the concepts and tools' tradeoffs in the lectures
 - ❖ TAs will explain and demo the tools; handle all Q&A
 - ❖ You are expected to put in the effort to learn the details of the tools' APIs using their documentation on your own!

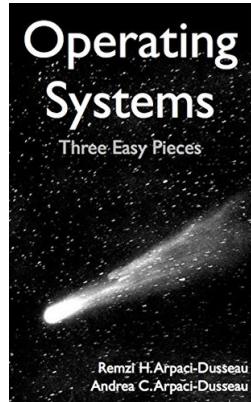
Course Administrivia

- ❖ **Lectures: TuTh 8-9:20am PT in MANDE B-210**
 - ❖ Attendance optional but encouraged; podcast available
 - ❖ Bring iClicker to class for PI activities; app is OK too
- ❖ **Discussions: Tu 7-7:50pm PT in MANDE B-210**
 - ❖ Only for talks on PAs and exams by TAs
- ❖ **Instructor:** Rod Albuyeh; ralbuyeh@ucsd.edu
- ❖ **OHs: Thu 9:30-10:30am PT at SDSC 2nd Floor**
- ❖ **TAs:** Golokesh Patra, Trevor Tuttle; see webpage for details on TA OHs
- ❖ **Course Website** for all announcements
- ❖ **Campuswire for async discussion**
- ❖ **Canvas** for PA submission, Final Exam, Extra Credits

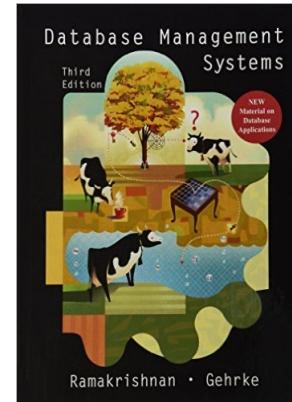
Suggested Textbooks



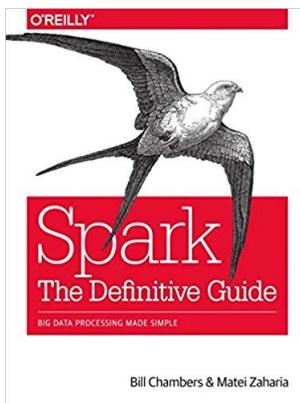
Aka “CompOrg Book”



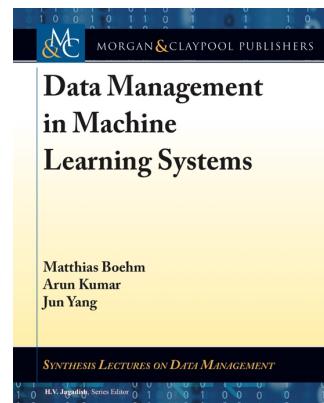
Aka “Comet Book”



Aka “Cow Book”



Aka “Spark Book”

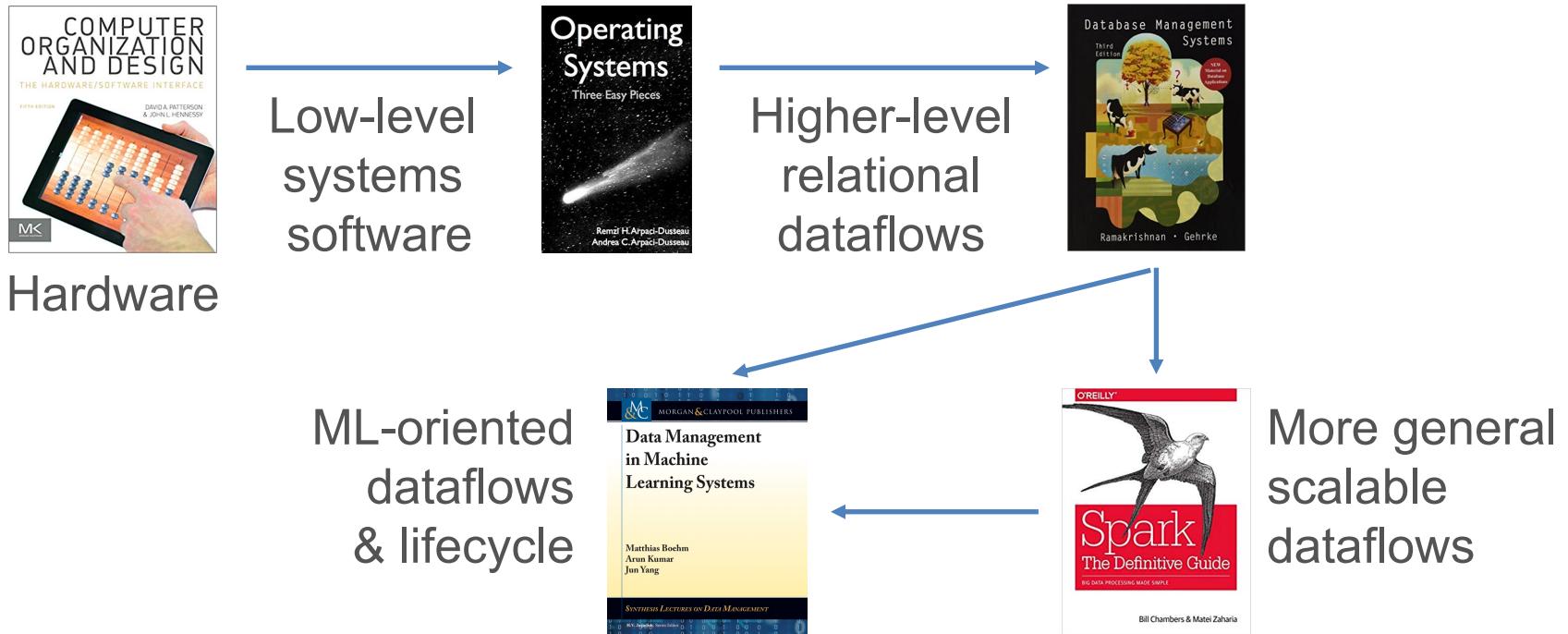


Aka “MLSys Book”

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

Why so many textbooks?!

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



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!

General Dos and Do NOTs

Do:

- ❖ Follow all announcements on course website
- ❖ Try to join the lectures/discussions live
- ❖ Participate in discussions in class / on Campuswire
- ❖ Raise your hand before speaking
- ❖ View/review podcast videos asynchronously by yourself

Do NOT:

- ❖ Harass, intimidate, or intentionally talk over others
- ❖ Violate academic integrity on the PAs, exams, or other components; I am *very strict* on this matter!