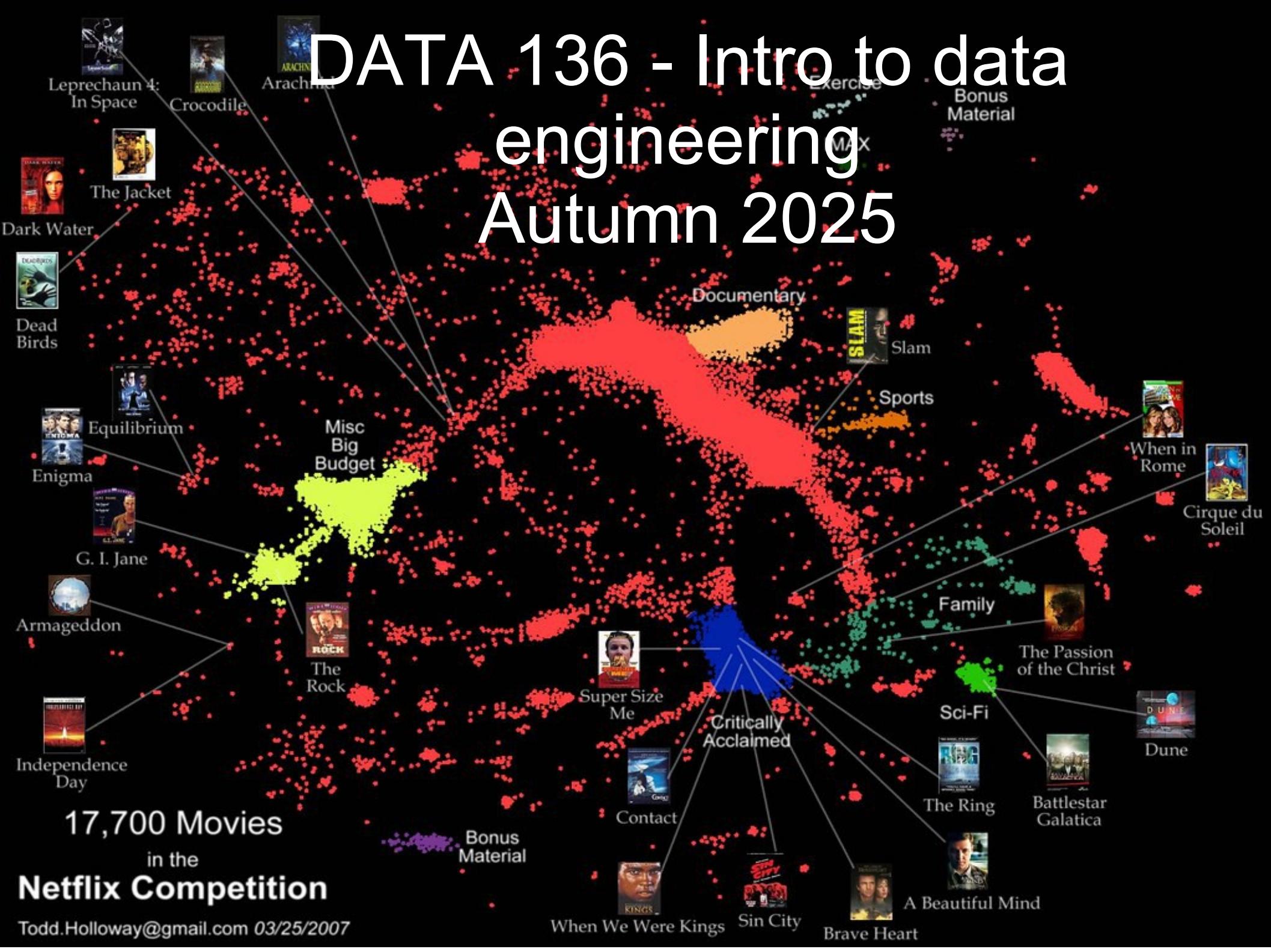


DATA 136 - Intro to data engineering

Autumn 2025



Teaching Staff

Professor

Will Trimble (he/him)



Teaching Assistants

Harleen Buttar

Cielo Martinez Flores

Data Engineering course content

How do we handle the firehose?

~~Algorithms, Optimization, and Statistics~~

Data collection and processing infrastructure

Software development: version control, testing

Scalability (bag of tricks)

Security, privacy, and governance

Q: What is Data ?

Q: What is "Big Data" ?

What is Data Engineering?

"Data engineering is a set of operations aimed at creating interfaces and mechanisms for the flow and access of information. It takes dedicated specialists—data engineers—to maintain data so that it remains available and usable by others. In short, data engineers set up and operate the organization's data infrastructure, preparing it for further analysis by data analysts and scientists."

AlexSoft <https://www.altexsoft.com/blog/what-is-data-engineering-explaining-data-pipeline-data-warehouse-and-data-engineer-role/>

DE is a box of tricks to allow you to perform tasks with data that would otherwise run you out of resources.

Structure / grading

- quarter-long class project: develop an app to spec 60%
- 3 exams 40%
- The project requires substantial programming and para-programming hacking effort, and contains open-ended, poorly-specified parts that you will have to fill in with some reasonable behavior.
- "Code to specification"
- Autograder, anyone?

HW0 Friday Mar 28

- Make a copy of a github repository, add a file, remove a file, check in your changes, push these changes to the cloud and share the branch in your repository with gradescope.
- Some people can finish HW0 in 15-20 minutes.
- Some people can spend 3 hours on it.
- Maybe find out who the 15 minute people are and find out how they like to be bribed?

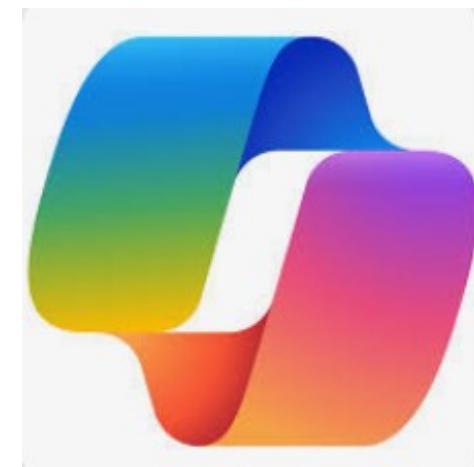
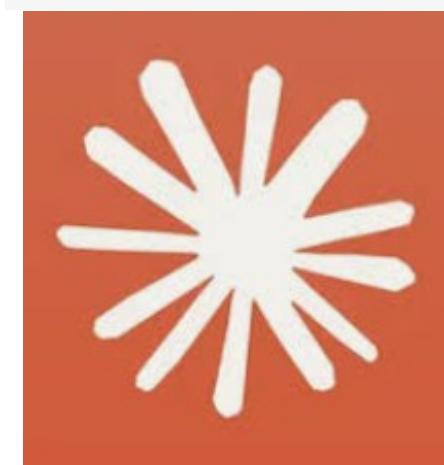
The plan

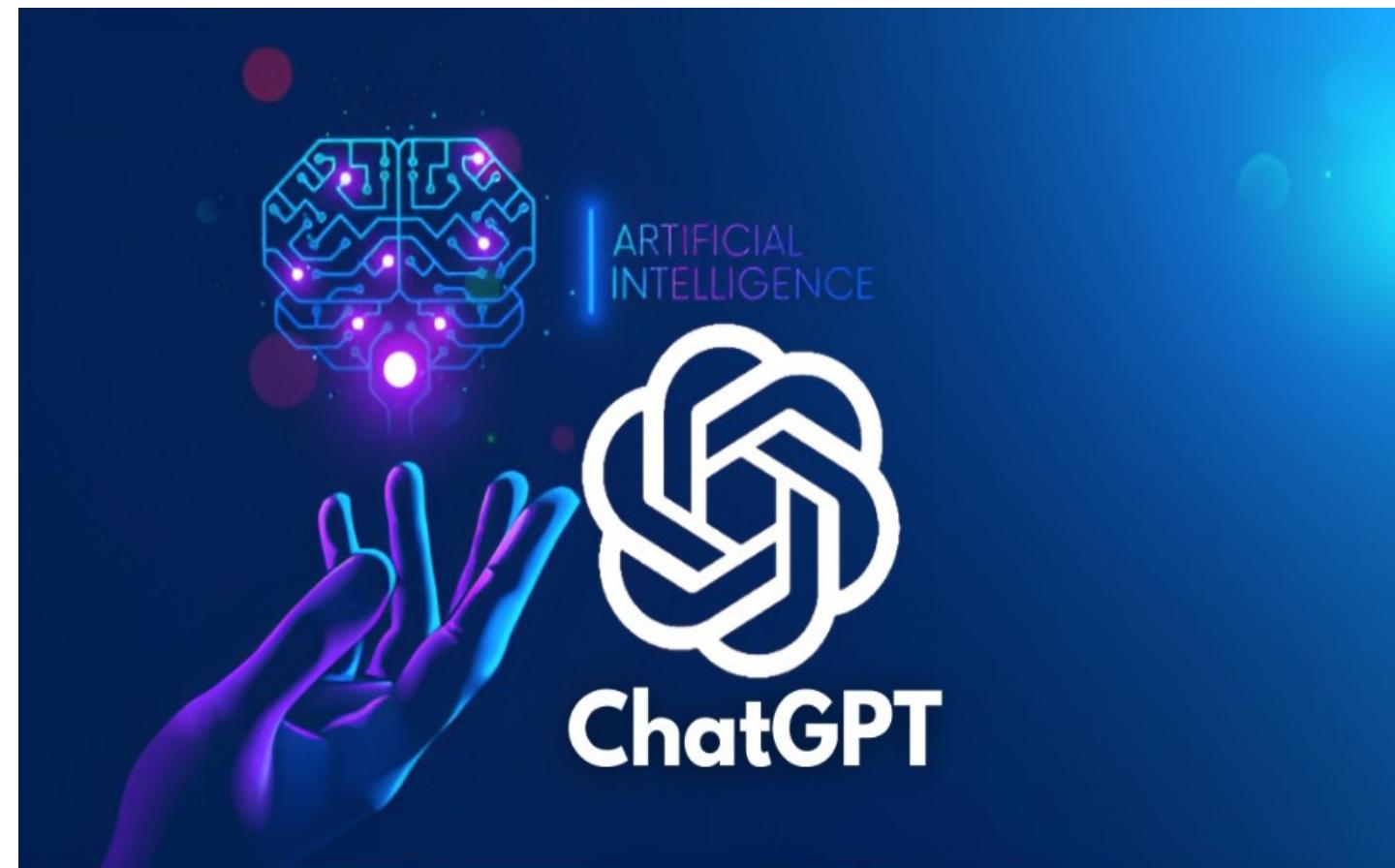
- Weeks 1-2 Shell/ git / Relational databases
- Week 3 architecture / data implementation
- Weeks 4-5 Indexing tricks
- Weeks 6-7 Tradeoffs / parallelization/ containerization / cloud deployment
- Weeks 8-9 Security, privacy, embeddings

THE SIMPLE ANSWERS

TO THE QUESTIONS THAT GET ASKED
ABOUT EVERY NEW TECHNOLOGY:

| | |
|--|---------------------------|
| WILL [] MAKE US ALL GENIUSES? | NO |
| WILL [] MAKE US ALL MORONS? | NO |
| WILL [] DESTROY WHOLE INDUSTRIES? | YES |
| WILL [] MAKE US MORE EMPATHETIC? | NO |
| WILL [] MAKE US LESS CARING? | NO |
| WILL TEENS USE [] FOR SEX? | YES |
| WERE THEY GOING TO HAVE SEX ANYWAY? | YES |
| WILL [] DESTROY MUSIC? | NO |
| WILL [] DESTROY ART? | NO |
| BUT CAN'T WE GO BACK TO A TIME WHEN- | NO |
| WILL [] BRING ABOUT WORLD PEACE? | NO |
| WILL [] CAUSE WIDESPREAD ALIENATION BY CREATING A WORLD OF EMPTY EXPERIENCES? | WE WERE ALREADY ALIENATED |





Having memorized wikipedia and ALL the documentation, LLMs are pretty good at a lot of tasks. My suggestion is that you use it similarly to the way you use a search engine, to get a feel for which problems are better for search engine and which are better for LLM.

stubbing : excellent

debugging: good

integration: not great

explaining: pretty good

AI strategy ...

For a given (huge) corpus of training data, AI gives answers with accuracies that depend on the problem

Literacy - agreed-upon facts - expert questions

Search engine skill: finding the answer among the (increasingly crummy) list of hits quickly

LLM skill: determining whether the LLM is helping you or wasting your time quickly

AI goals

- Some AI use solves the problem but does not program the primate to solve the problem.
(stackoverflow?)
- Don't embarrass yourself !
- Copying from documentation (and class materials) is explicitly permitted, but the IP culture in CS forbids copying code without licensing.

too hard

The three resources



The three resources



\$45/month
for
~30Mbits/s



\$1500 /
3 years
\$30/month



1Tb is about
\$100...
\$0.02 / Gb
/ year?

The cloud permits you to rent the three resources

Data Transfer
OUT form
Amazon EC2 to
Internet



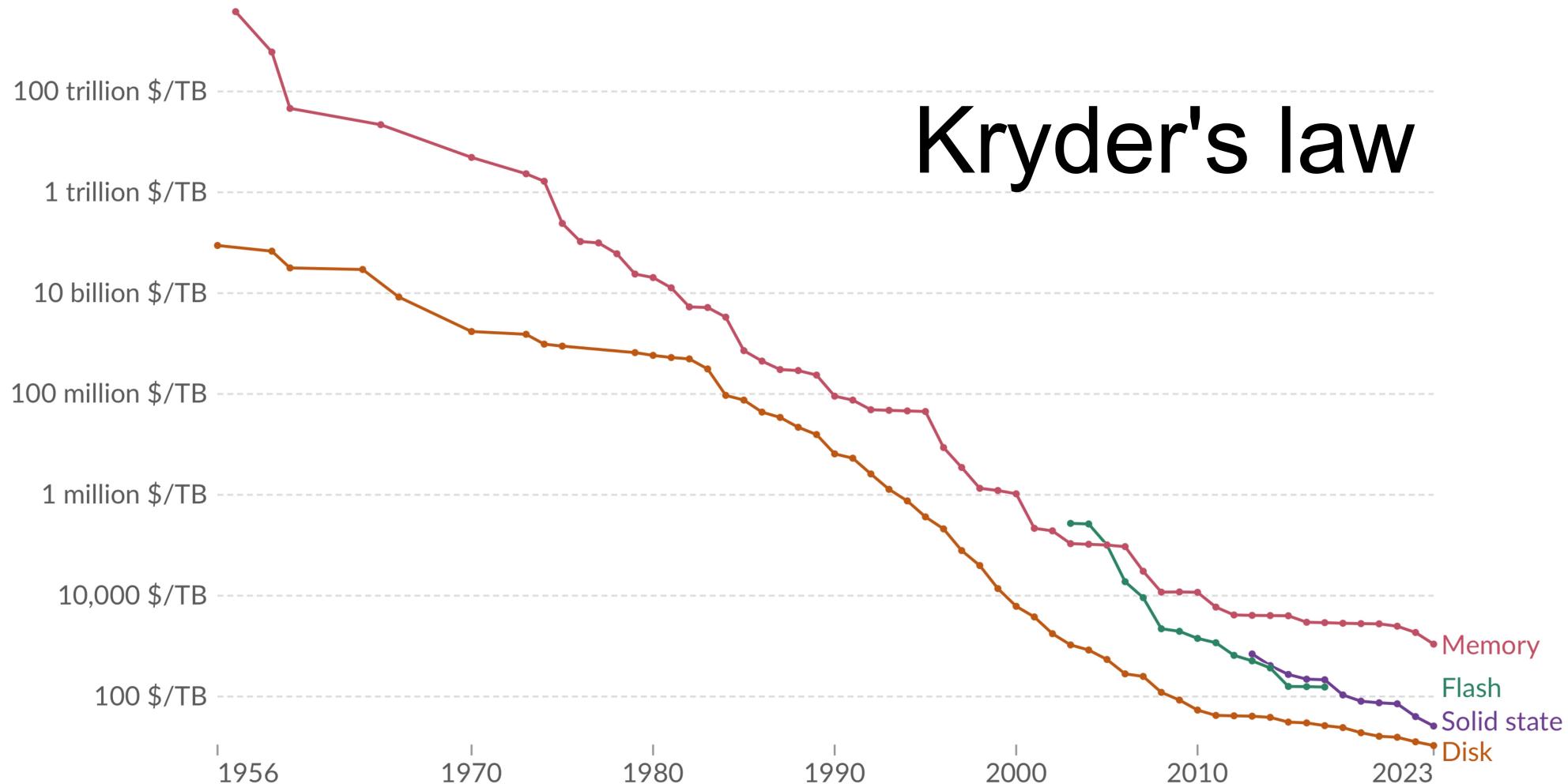
\$0.09 / Gb
out

t4g.medium
\$0.0336 /h
\$0.81 / day

S3 standard:
\$0.023 /
Gb / month

Historical price of computer memory and storage

This data is expressed in US dollars per terabyte (TB), adjusted for inflation. "Memory" refers to random access memory (RAM), "disk" to magnetic storage, "flash" to special memory used for rapid data access and rewriting, and "solid state" to solid-state drives (SSDs).



Data source: John C. McCallum (2023); U.S. Bureau of Labor Statistics (2024)

OurWorldInData.org/technological-change | CC BY

Note: For each year, the time series shows the cheapest historical price recorded until that year. This data is expressed in constant 2020 US\$.

Moore's Law: The number of transistors on microchips has doubled every two years

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.

Transistor count

50,000,000,000

10,000,000,000

5,000,000,000

1,000,000,000

500,000,000

100,000,000

50,000,000

10,000,000

5,000,000

1,000,000

500,000

100,000

50,000

10,000

5,000

1,000

Moore's law

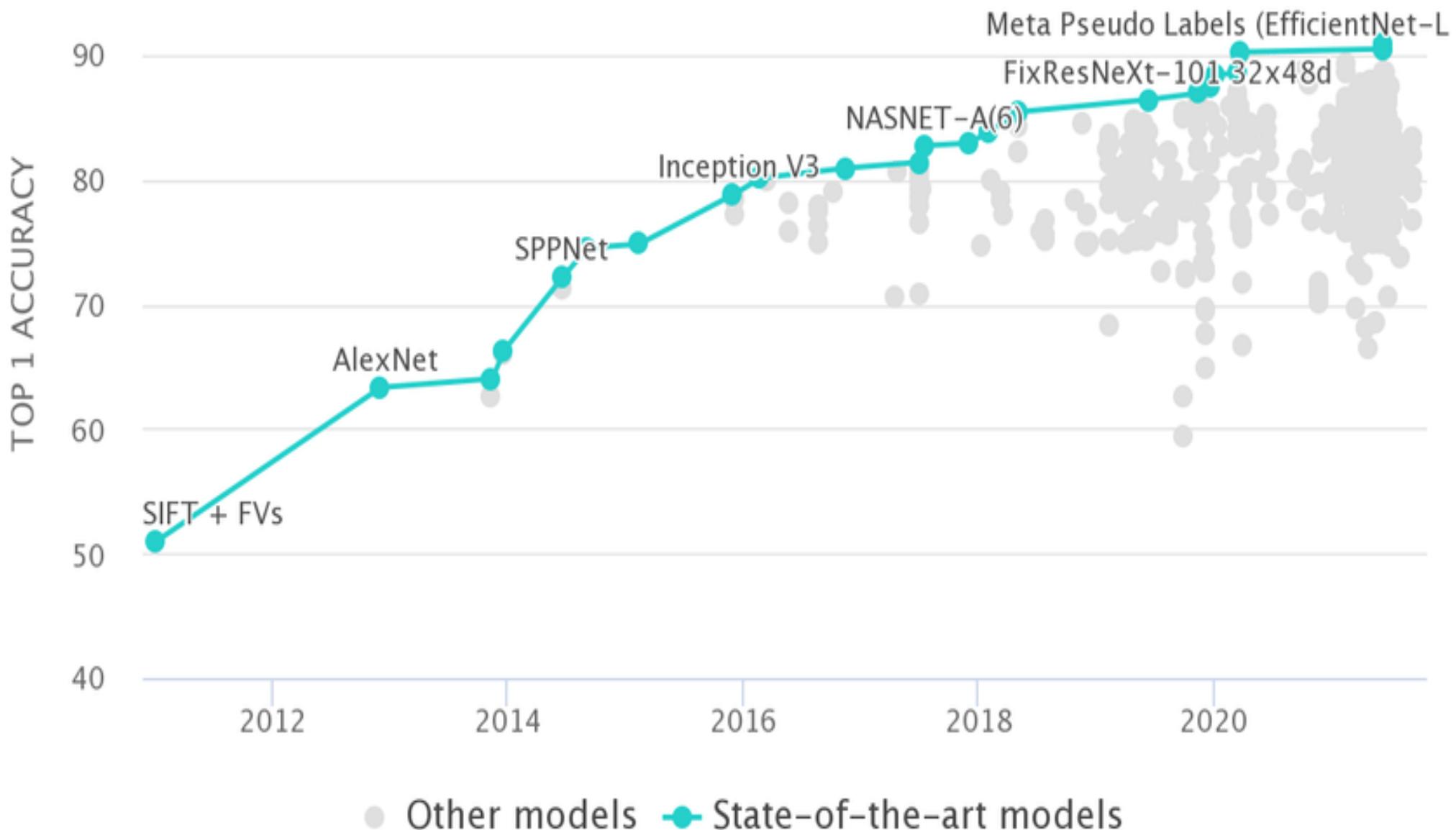
1970 1975 1980 1985 1990 1995 2000 2005 2010 2015 2020

Year in which the microchip was first introduced

Data source: Wikipedia ([wikipedia.org/wiki/Transistor_Count](https://en.wikipedia.org/w/index.php?title=Transistor_Count&oldid=910000000))

OurWorldInData.org – Research and data to make progress against the world's largest problems.

Licensed under CC-BY by the authors Hannah Ritchie and Max Roser.



The state of the art in image classification-in context of the Top-1 accuracy benchmark on the ImageNet dataset-has improved almost every year since 2012. The presence of a consistent numeric benchmark spurred the development of better systems for image classification. Image source : <https://paperswithcode.com/sota/image-classification-on-imagenet>.

So the cost of everything data-related is going down, so what?

- Data volume. Term "big data" was popular for a decade to describe the disquiet
- "Big data" -- data not suitable to analyze on your laptop/desktop
- Huge data volume... tempts many agents into data hoarding.
- Huge data volume inspires many miners.
There must be gold in that pile of (Uber, Netflix, Instagram, Whatsapp...) data somewhere!

WHEN A USER TAKES A PHOTO,
THE APP SHOULD CHECK WHETHER
THEY'RE IN A NATIONAL PARK...

SURE, EASY GIS LOOKUP.
GIMME A FEW HOURS.

... AND CHECK WHETHER
THE PHOTO IS OF A BIRD.

I'LL NEED A RESEARCH
TEAM AND FIVE YEARS.



IN CS, IT CAN BE HARD TO EXPLAIN
THE DIFFERENCE BETWEEN THE EASY
AND THE VIRTUALLY IMPOSSIBLE.

**Which tasks are
easy and which
are hard?**

Computers have recently solved the 20th century's problems.



MNIST: handwriting recognition

GLUE: natural language understanding

ImageNet: image recognition

SQuAD: reading comprehension

Switchboard: English transcription

The data pyramid...

THE DATA SCIENCE **HIERARCHY OF NEEDS**

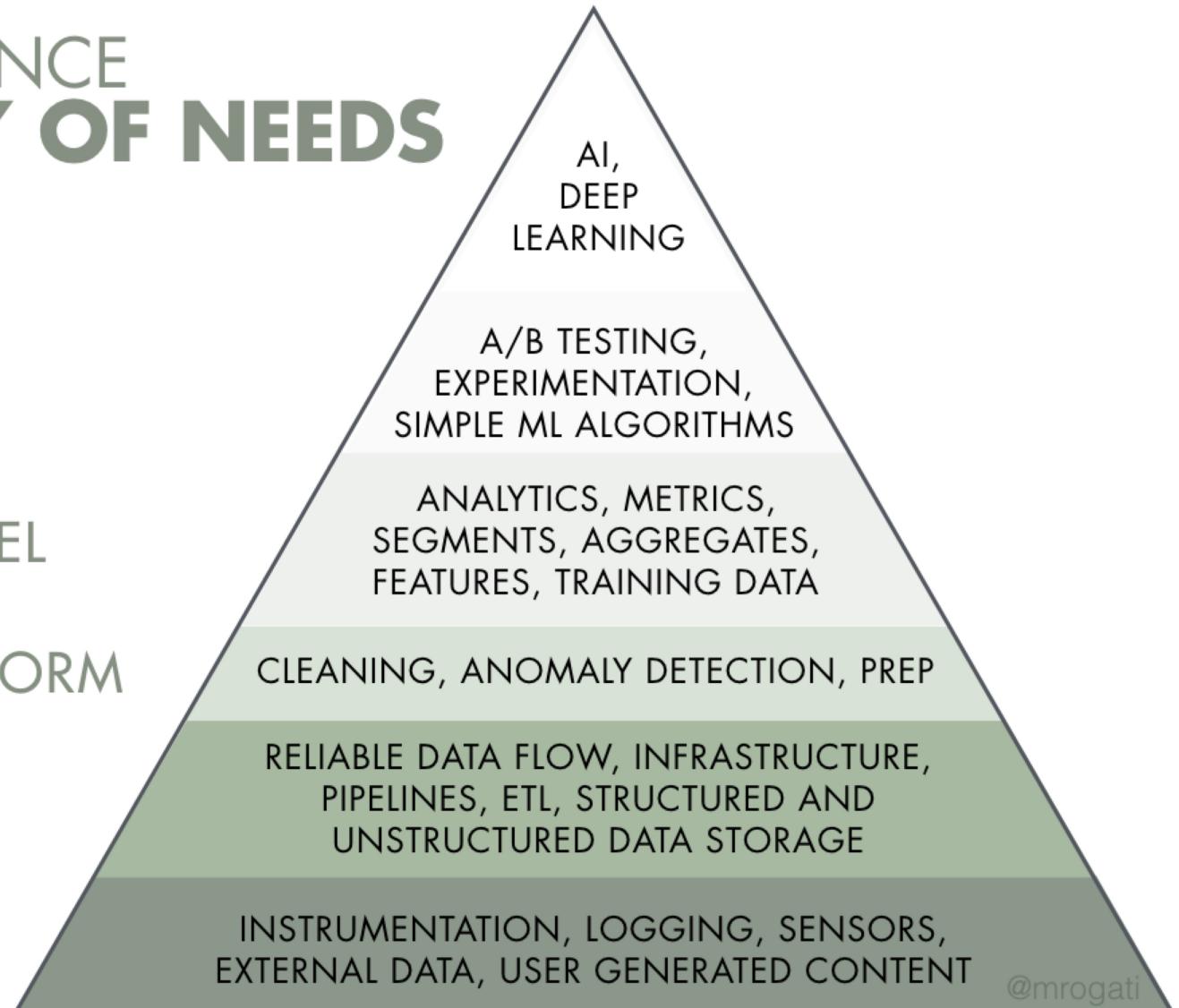
LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT



...is populated with data workers...

THE DATA SCIENCE **HIERARCHY OF NEEDS**

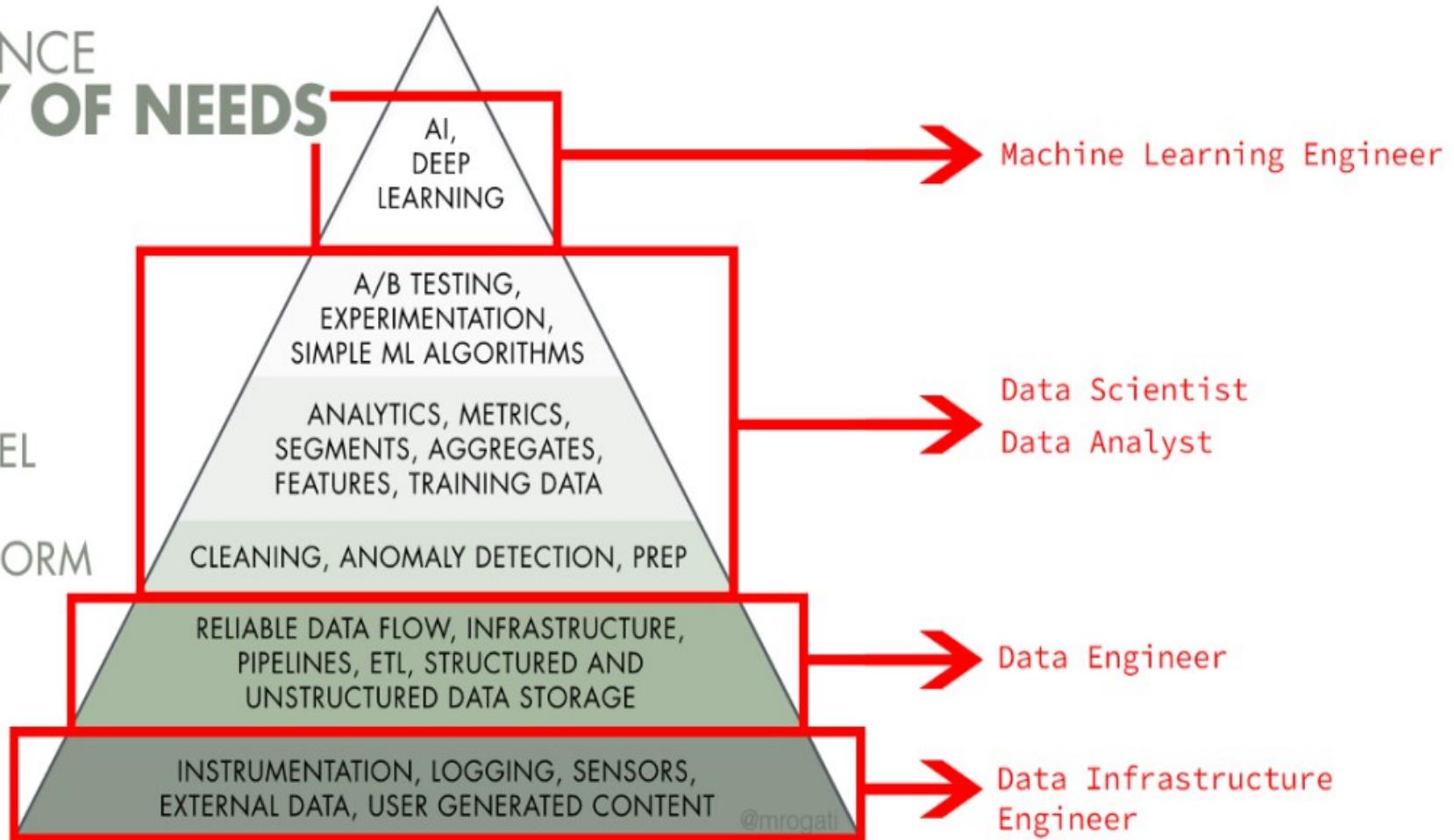
LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT



...yeah, there's a "technical" hierarchy...

- Type A Data Scientist: The A is for Analysis. This type is primarily concerned with making sense of data or working with it in a fairly static way. The Type A Data Scientist is very similar to a statistician (and may be one) but knows all the practical details of working with data that aren't taught in the statistics curriculum: data cleaning, methods for dealing with very large data sets, visualization, deep knowledge of a particular domain, writing well about data, and so on.
- Type B Data Scientist: The B is for Building. Type B Data Scientists share some statistical background with Type A, but they are also very strong coders and may be trained software engineers. The Type B Data Scientist is mainly interested in using data “in production.” They build models which interact with users, often serving recommendations (products, people you may know, ads, movies, search results).

Here's the plan...

| number | topic | date |
|--------|---------------------------------------|--------------|
| HW 0 | setup git / public key authentication | Oct 3 (wk1) |
| HW 1 | setup git collaboration | Oct 10 (wk2) |
| HW 2 | setup/install django / ORM | Oct 17 (wk3) |
| Exam 1 | databases & ORM | Oct 24 (wk4) |
| HW 3 | define DDL | Oct 31 (wk5) |
| HW 4 | front end: login page | Nov 7 (wk6) |
| HW 5 | implementing the API | Nov 14 (wk7) |
| Exam 2 | architecture, hashing | Nov 19 (wk8) |
| HW 6 | Proof of work, front end feedkkkkjj | Nov 21 (wk8) |
| | Thanksgiving break | |
| HW 7 | Writing tests for the app | Dec 5 (wk9) |
| Exam 3 | pulling it all together | TBA Dec 9-11 |

Office Hours

Office hours:

WT: Wednesday 12:30-1:30 (Ryerson 257B)

WT: Thursday 12:30-1:30 (Ryerson 257B)

HB:

CMF: