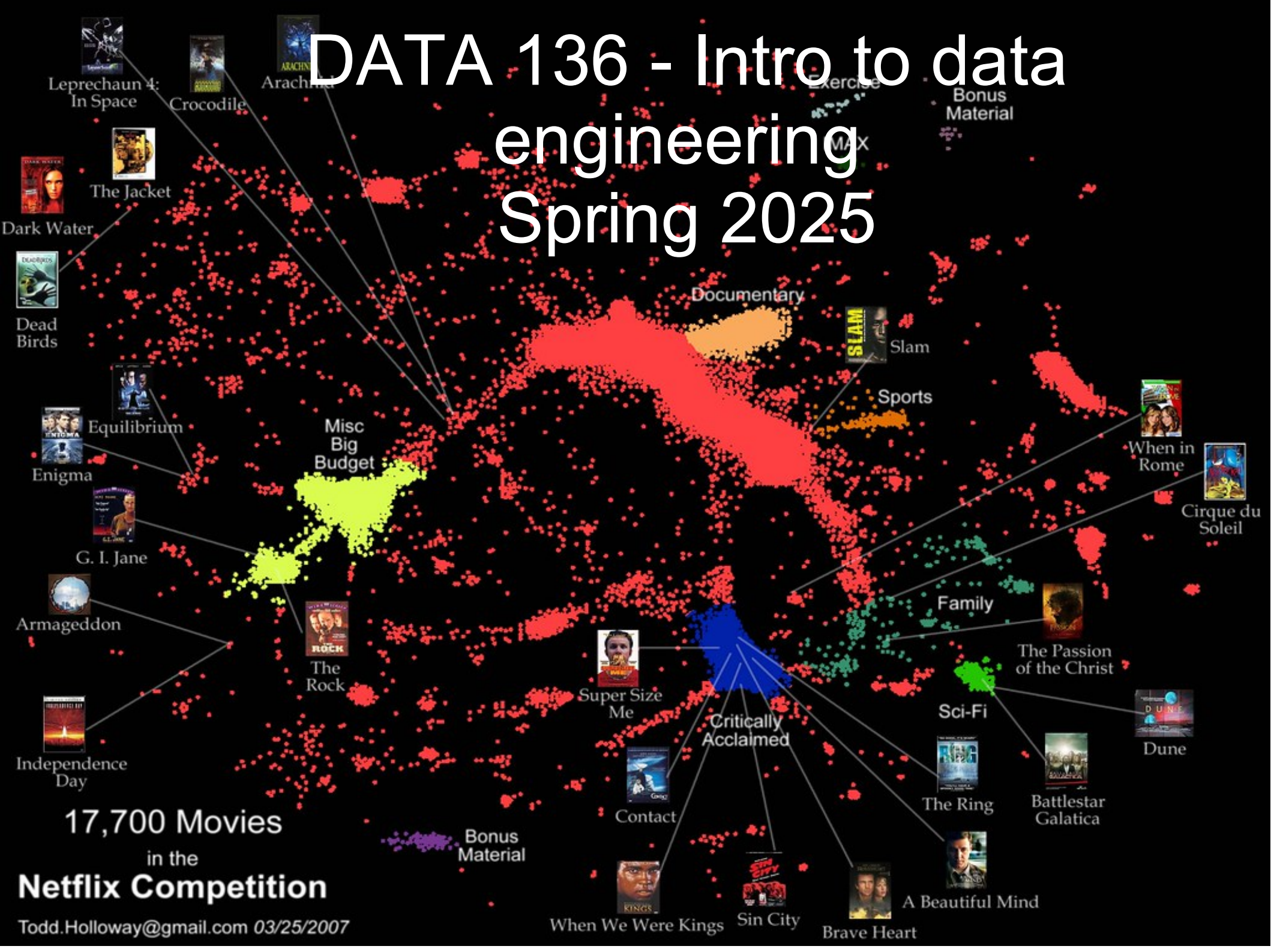
DATA 136 - Intro to data engineering Spring 2025



Teaching Staff

Teaching Assistants

Isaac Harlem (he/him)

Gio Maya (he/him)

Professor

Will Trimble (he/him)

Data Engineering course content

~~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

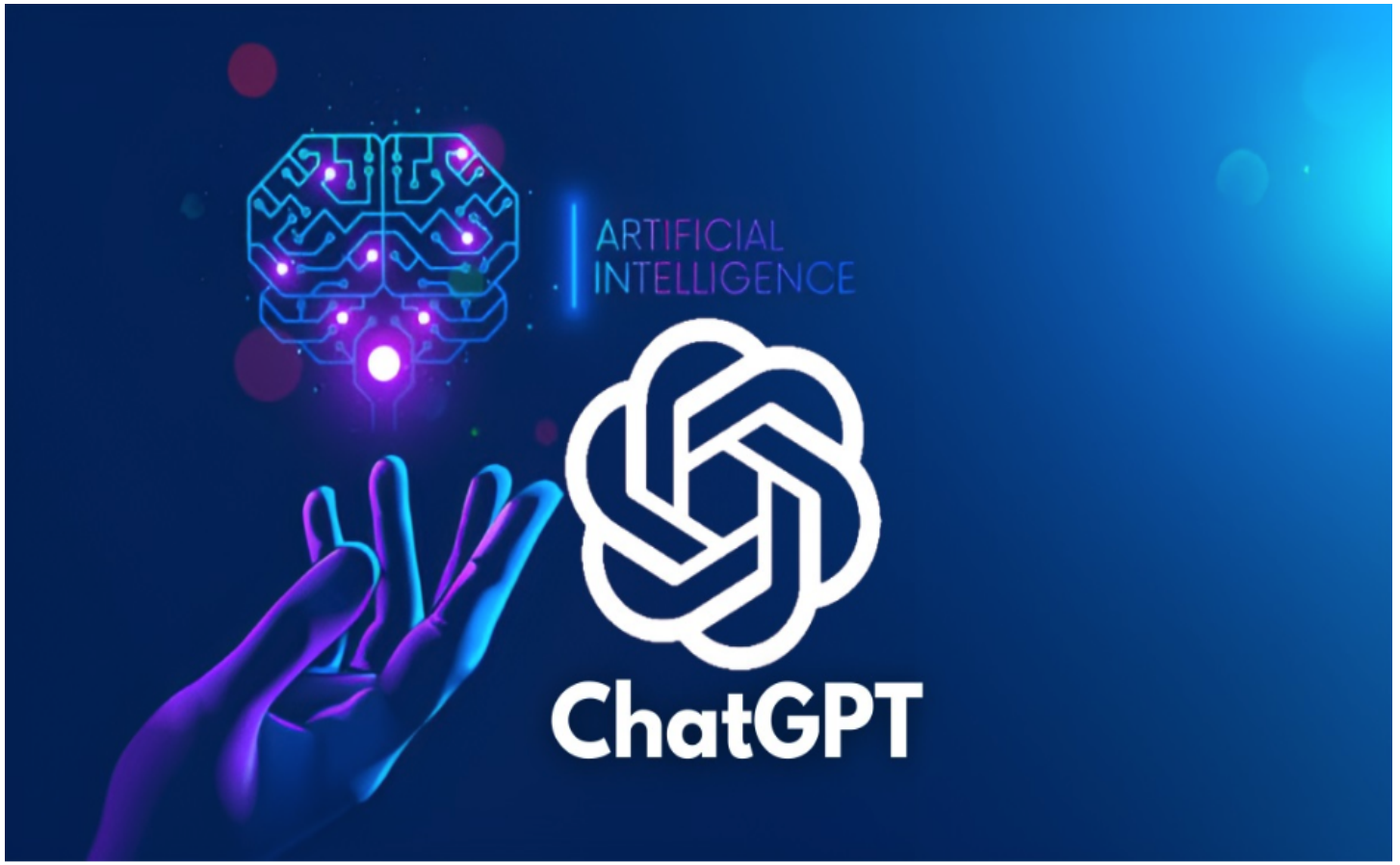
- quarter-long class project: develop an app to spec 60%
- 3 exams 40%
- The project requires substantial programming and para-programming hacking effort, and contains open-ended, poorly-specified parts that you will have to fill in with some reasonable behavior.
- "Code to specification"

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.

The plan

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



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 integration: not great

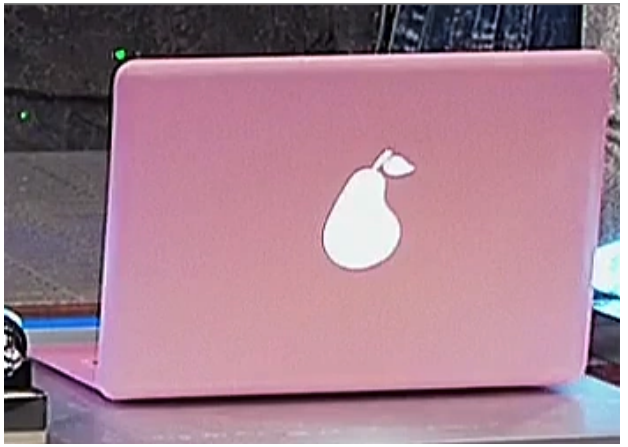
debugging: good explaining: pretty good

THE SIMPLE ANSWERS

TO THE QUESTIONS THAT GET ASKED ABOUT EVERY NEW TECHNOLOGY:

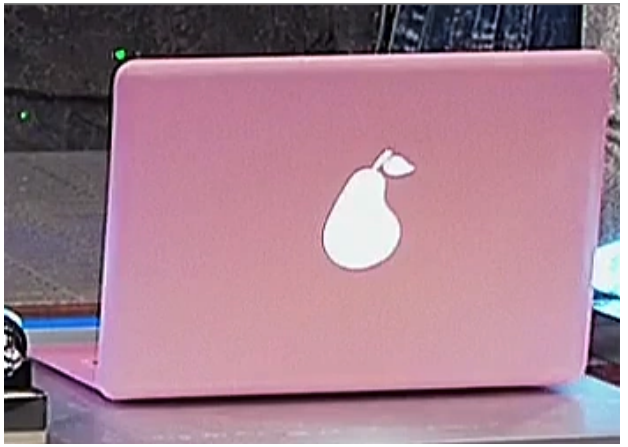
WILL [ ] MAKE US ALL GENIUSES?	NO
WILL [ ] MAKE US ALL MORONS?	NO
WILL [ ] DESTROY WHOLE INDUSTRIES?	YES
WILL [ ] MAKE US MORE EMPATHETIC?	NO
WILL [ ] MAKE US LESS CARING?	NO
WILL TEENS USE [ ] FOR SEX?	YES
WERE THEY GOING TO HAVE SEX ANYWAY?	YES
WILL [ ] DESTROY MUSIC?	NO
WILL [ ] DESTROY ART?	NO
BUT 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

The three resources



The three resources

$45/month
for
~30Mbits/s



$1500 /
3 years
$30/month

$0.02 / Gb
/ 5 years?

The cloud permits you to rent the three resources

Data Transfer OUT form Amazon EC2 to Internet

$0.09 / Gb out

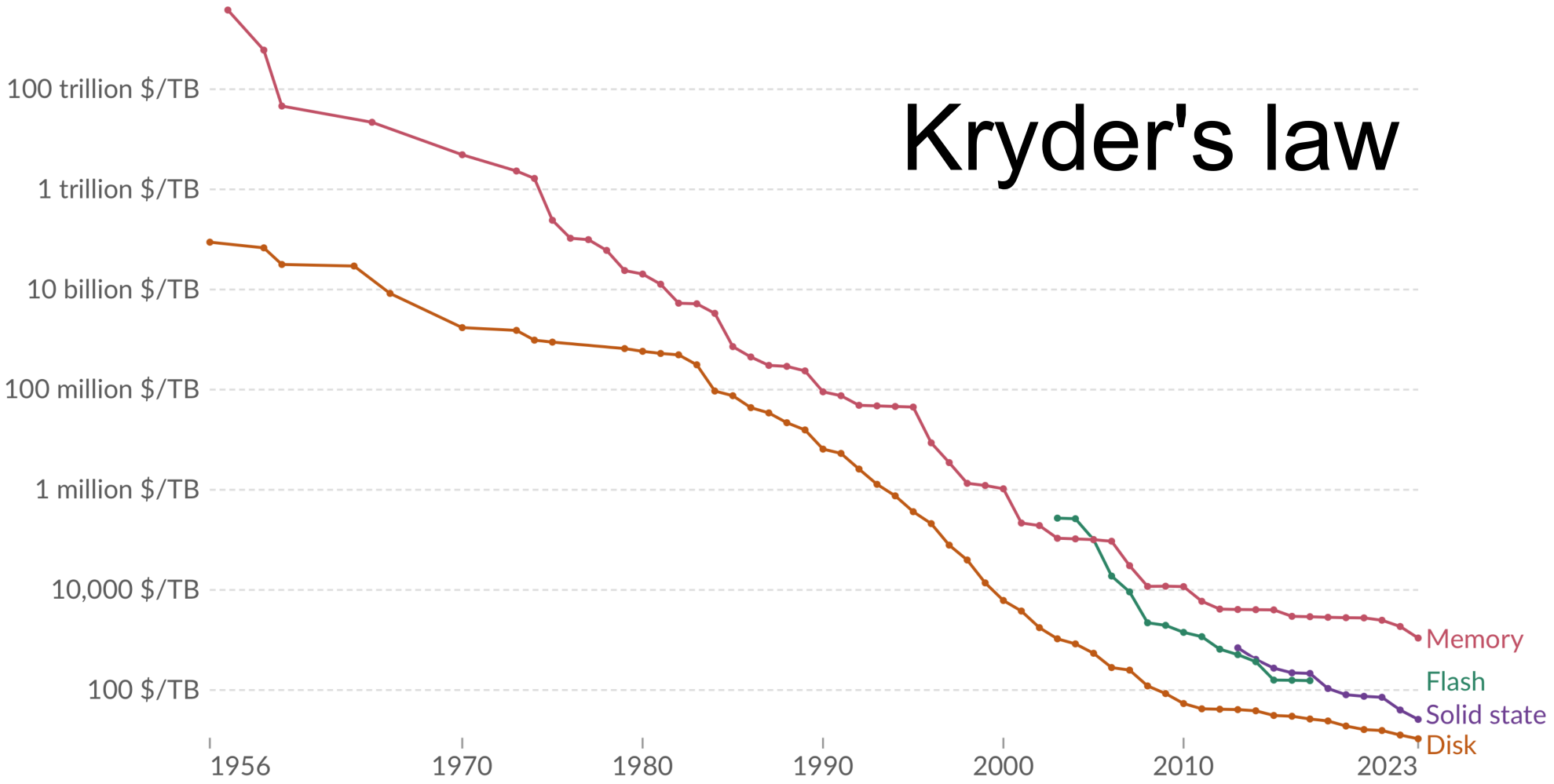
t4g.medium
$0.0336 /h
$0.81 / day

$0.023 / Gb / month

Historical price of computer memory and storage

Our World in Data

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)

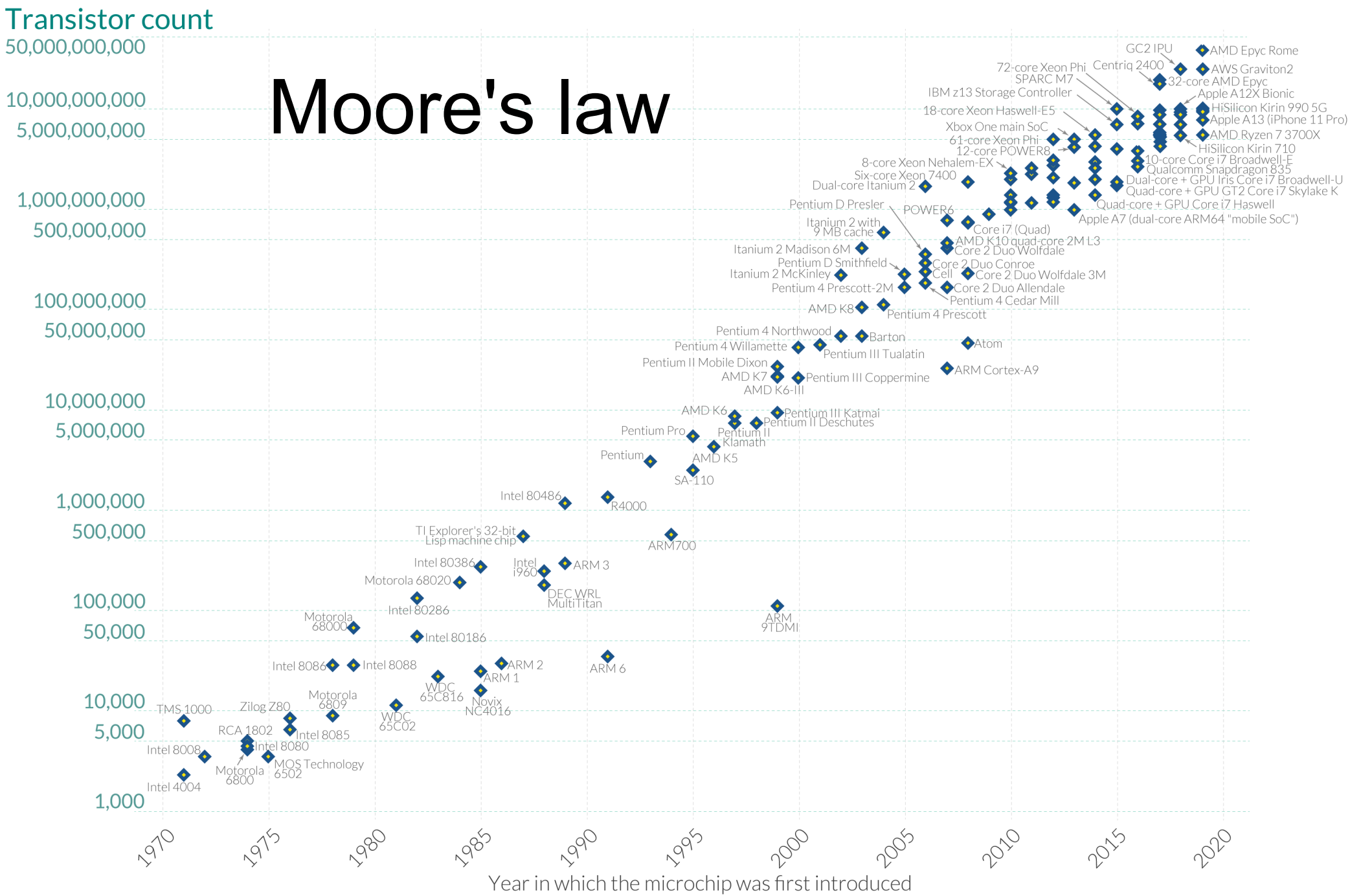
OurWorldInData.org/technological-change | CC BY

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

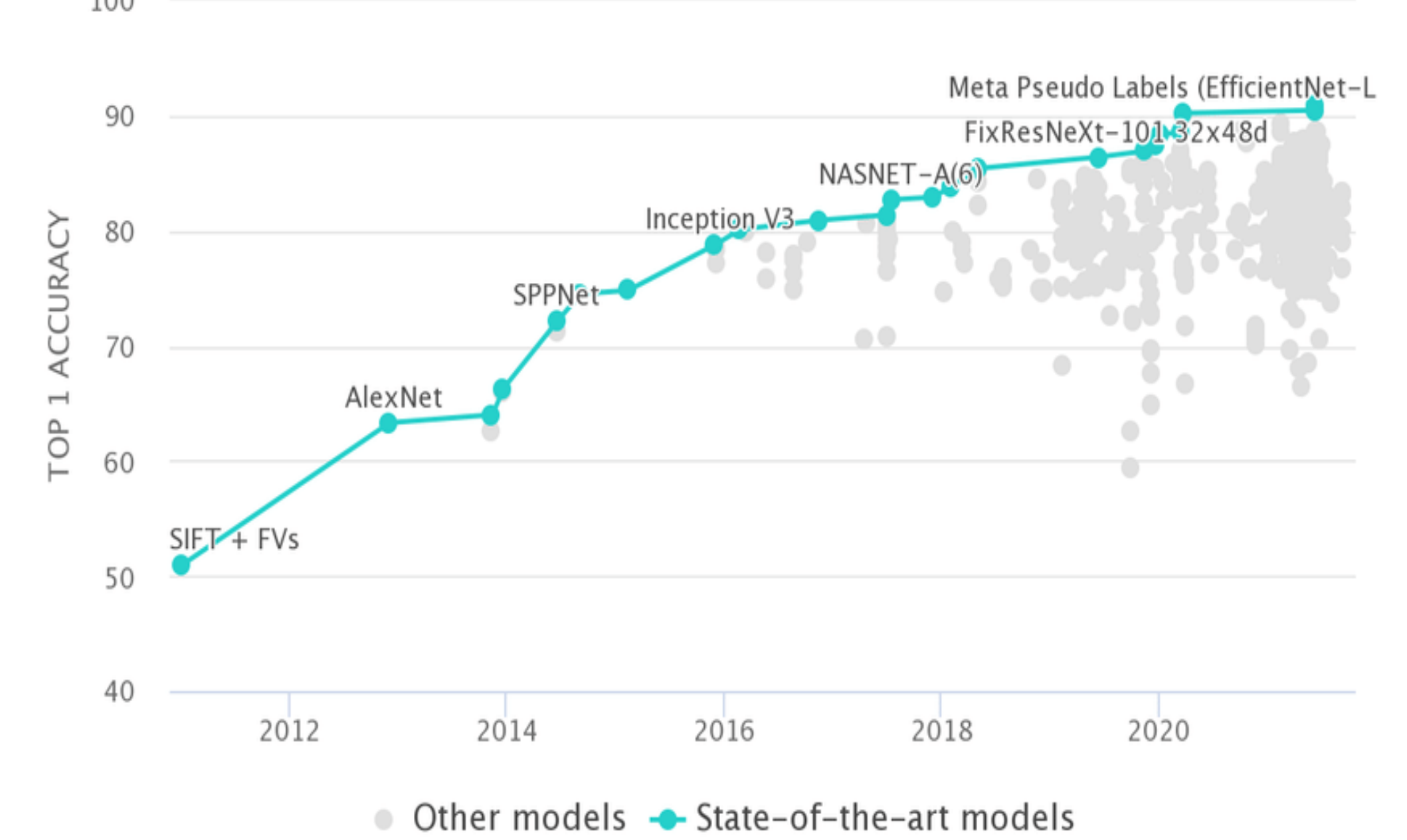
Moore's Law: The number of transistors on microchips has doubled every two years

Our World in Data

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.



Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)



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.

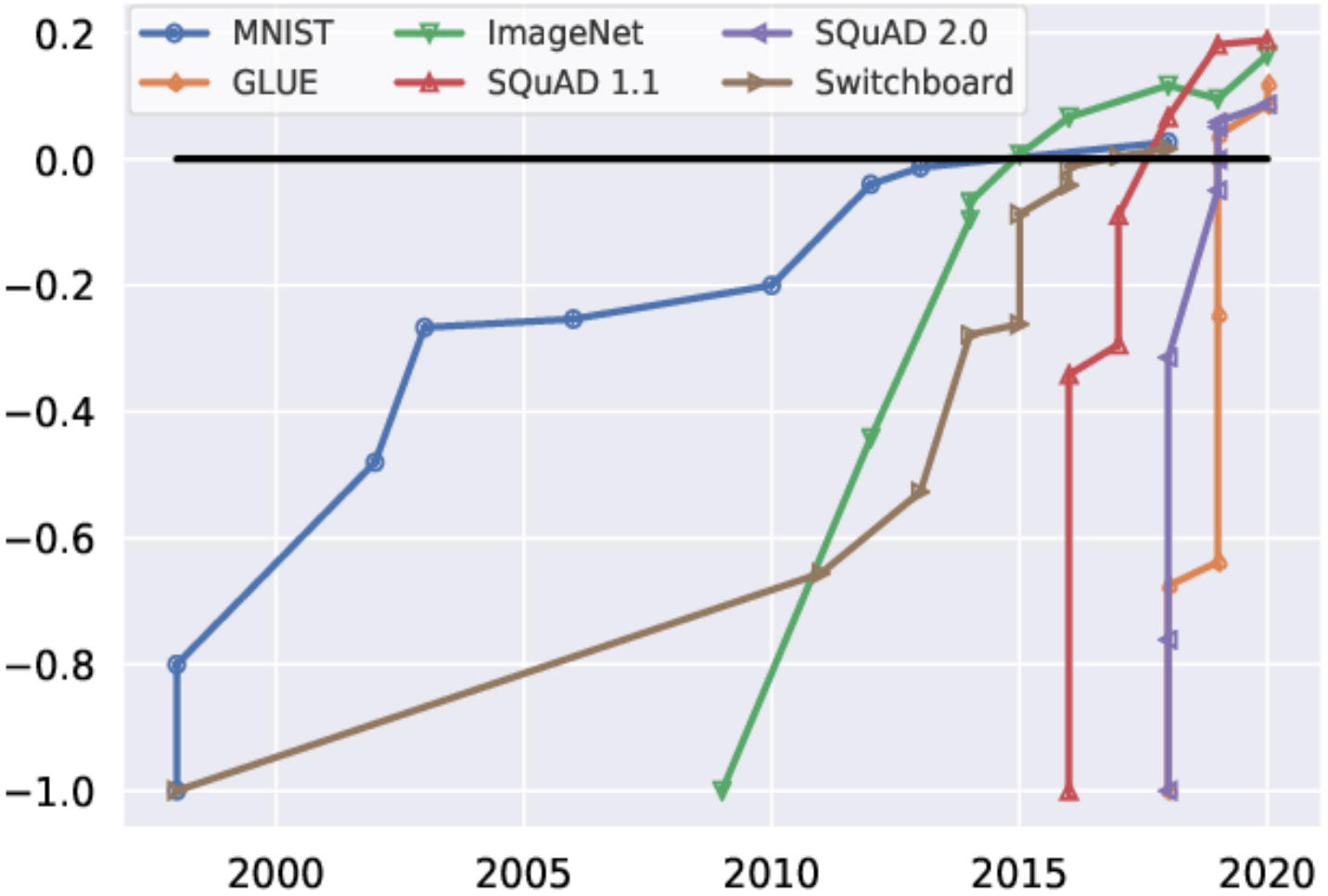
So the cost of everything data-related is going down, so what?

- 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!

Randall Munroe, xkcd, September 24, 2014

Which tasks are easy and which are hard?

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



MNIST: handwriting recognition

GLUE: natural language understanding

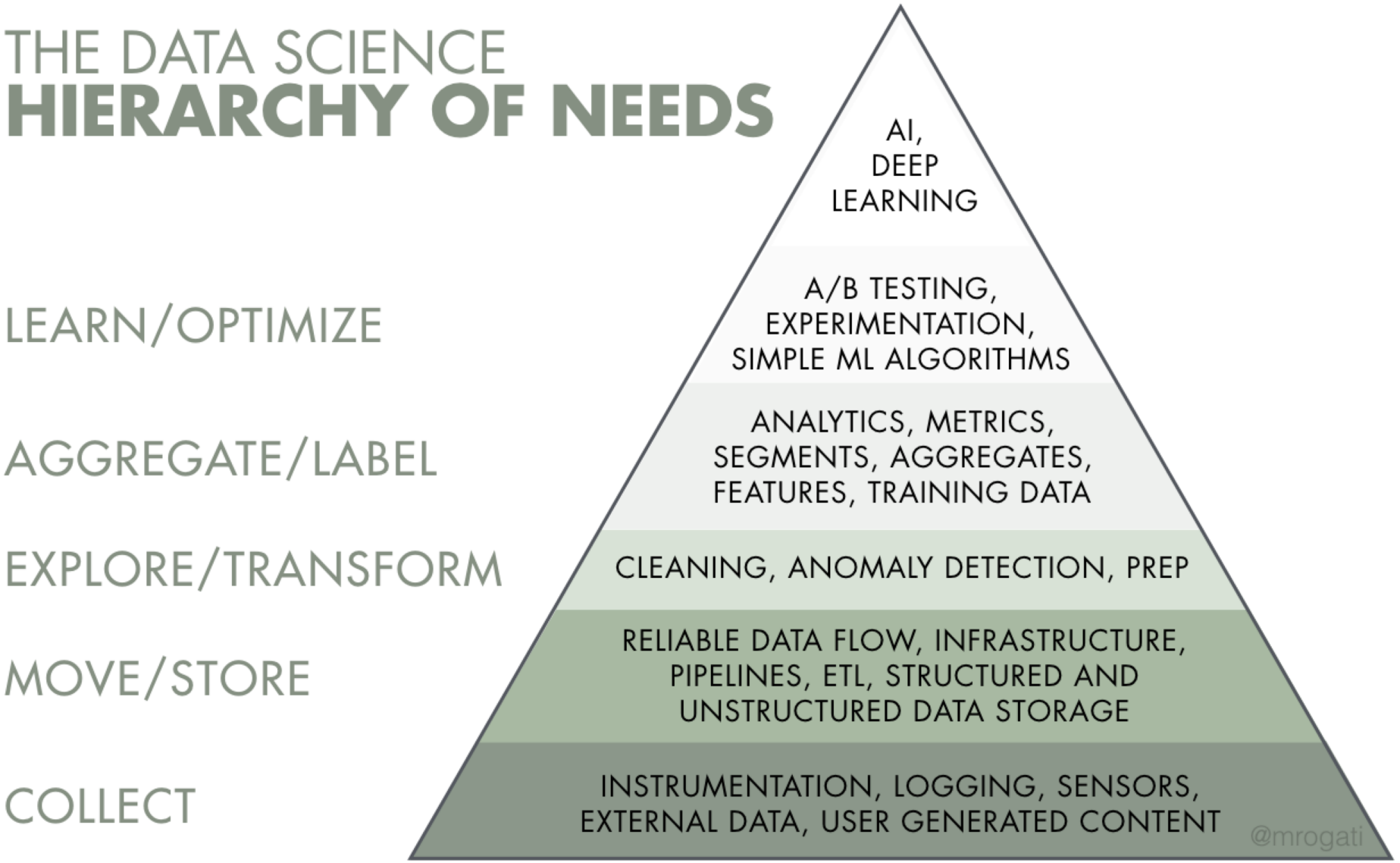
ImageNet: image recognition

SQuAD: reading comprehension

Switchboard: English transcription

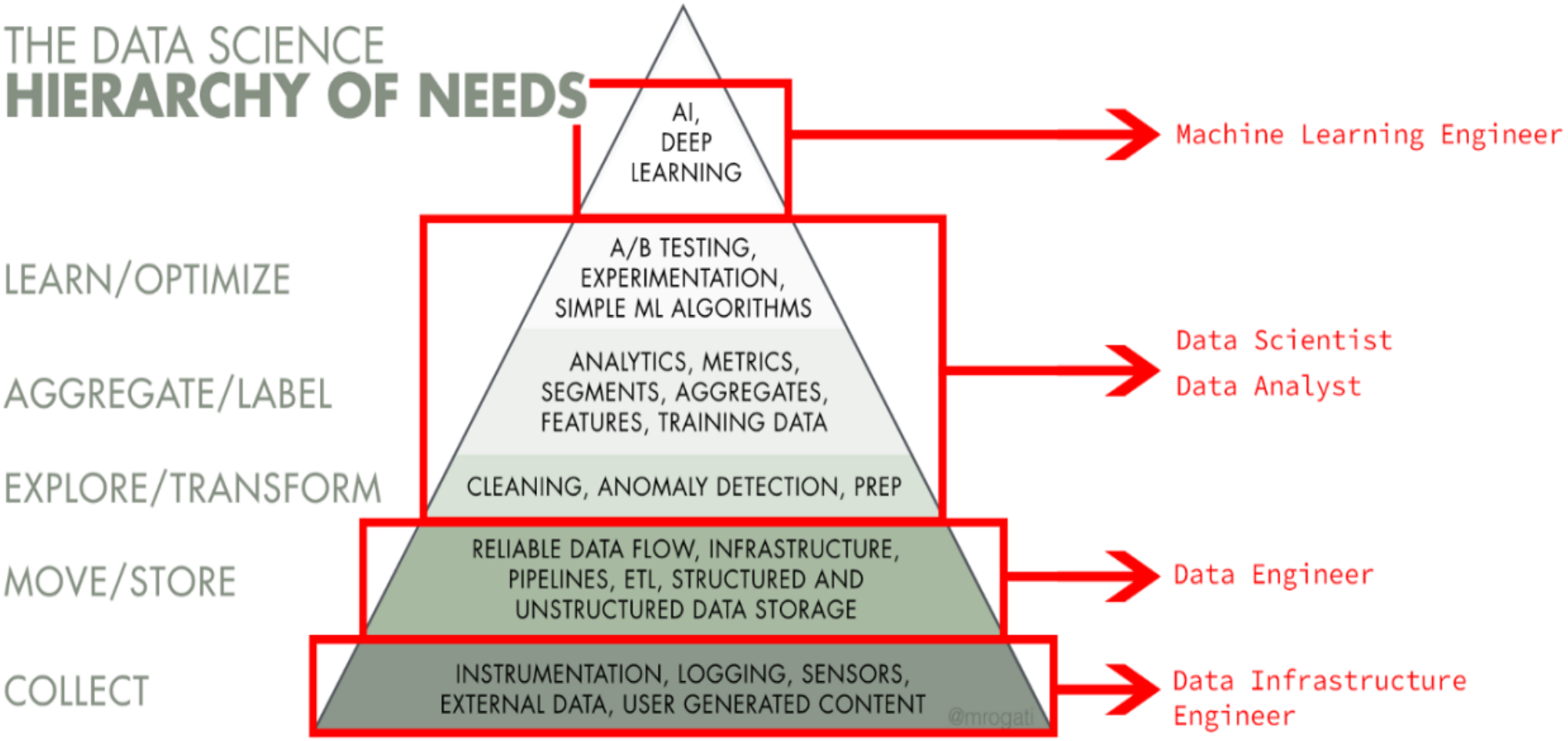
10.48550/arXiv.2104.14337 Kiela et al. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2021
Dynabench: Rethinking Benchmarking in NLP

The data pyriamid...



https://medium.com/@hugh_data_science/the-pyramid-of-data-needs-and-why-it-matters-for-your-career-b0f695c13f11

...is populated with data workers...



https://www.linkedin.com/pulse/data-science-hierarchy-needs-emmanuel-ogungbemi-phd

...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).

Michael Hochster, on Quora, 2015

Here's the plan...

number	topic	date
HW 0	setup git / public key authentication	Mar 28
HW 1	setup git collaboration	Apr 4
HW 2	setup/install django / ORM	Apr 11
Exam 1	databases & ORM	Apr 18
HW 3	define DDL	Apr 25
HW 4	front end: login page	May 2
HW 5	implementing the API	May 9
HW 6	Testing the API, creating fixtures	May 16
Exam 2	architecture, hashing	May 14
HW 7	Testing the app, writing tests	May 23
Exam 3	pulling it all together	TBA May 27-29

Office Hours

Office hours:

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

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

IH:

GM: