Dataset Challenges and Hardware Accelerators

Danna Gurari
University of Colorado Boulder
Spring 2025

https://dannagurari.colorado.edu/course/neural-networks-and-deep-learning-spring-2025/

Review

- Last lecture:
  - Key challenge: training large capacity, deep models
  - AlexNet: key tricks for going 8 layers deep
  - ResNet: key tricks for extending to 152 layers deep
  - Programming tutorial

- Assignments (Canvas)
  - Problem set 2 grades are out
    - Review session will be held at 2pm today
    - Email all regrade requests to our TA, Nick Cooper (a comment in Canvas is not sufficient)
  - Lab assignment 1 due Thursday

- Questions?

Today’s Topics

- Dataset challenges: before versus after 2012
- Hardware: before versus after 2012
- Programming tutorial

Today's Topics

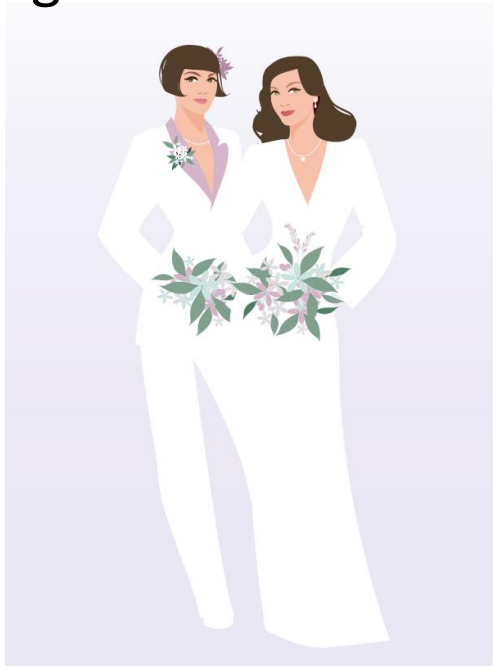
- Dataset challenges: before versus after 2012
- Hardware: before versus after 2012
- Programming tutorial

Status Quo for Advancing AI Until 2012

Algorithm Dataset

Algorithm Dataset

Algorithm Dataset



Algorithm Dataset

Datasets tended to be relatively small (e.g., 10s or 100s of examples)

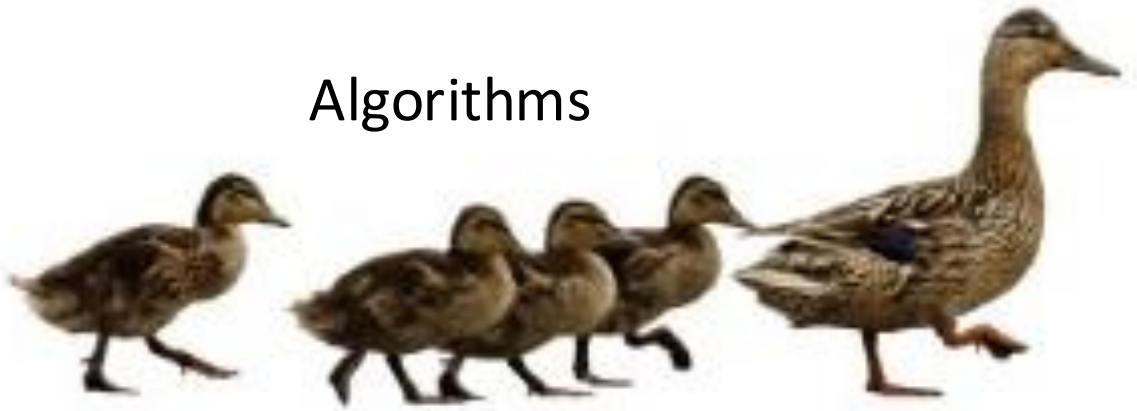
Status Quo for Advancing AI Until 2012

- Authors created their own datasets (e.g., using their cameras, purchasing data from companies, or downloading Internet images)
- What's wrong with this approach?
  - Unable to perform "fair" comparison between algorithms
  - Lacks a community around a shared goal

Status Quo for Advancing AI Since 2012

Dataset

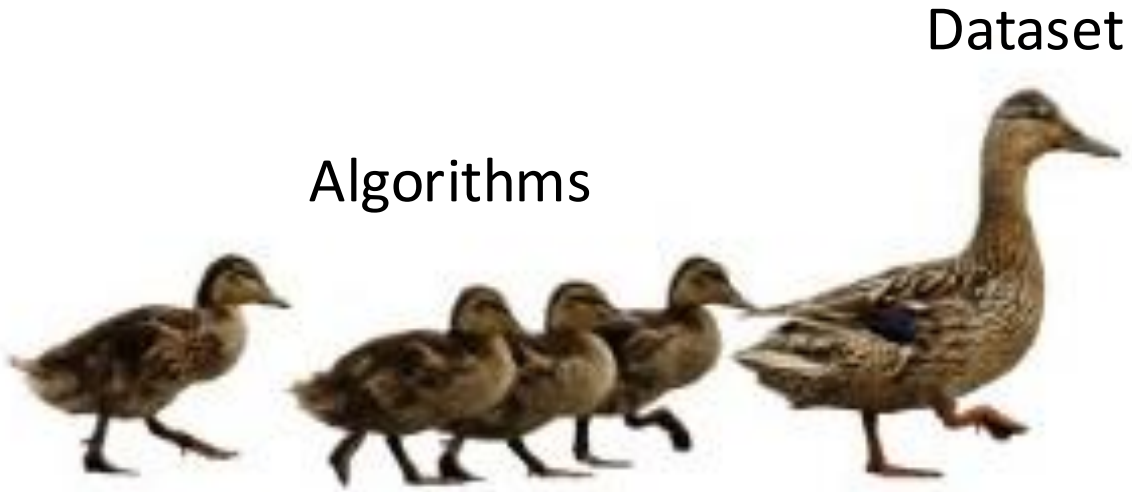
Algorithms



Datasets tend to be large (e.g., thousands to billions of examples)

Status Quo for Advancing AI Since 2012

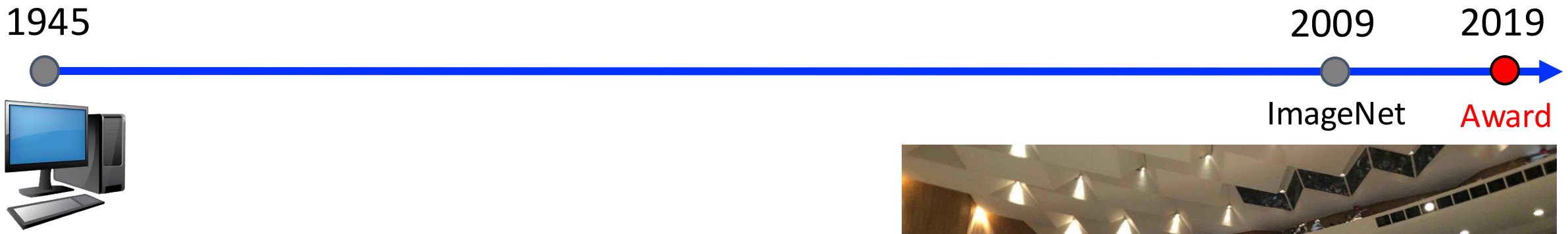
What do you think prompted this shift to large-scale datasets?



Datasets tend to be large (e.g., thousands to billions of examples)

http://larryzitnick.org/Talks/CVPR15_Dataset.pptx

ImageNet: Inspired Era of Dataset Challenges



PAMI Longuet-Higgins Prize

Retrospective Most Impactful Paper from CVPR 2009

ImageNet: A large-scale hierarchical image database

Jia Deng, Wei Dong, Richard Socher,
Li-Jia Li, Kai Li, and Li Fei-Fei

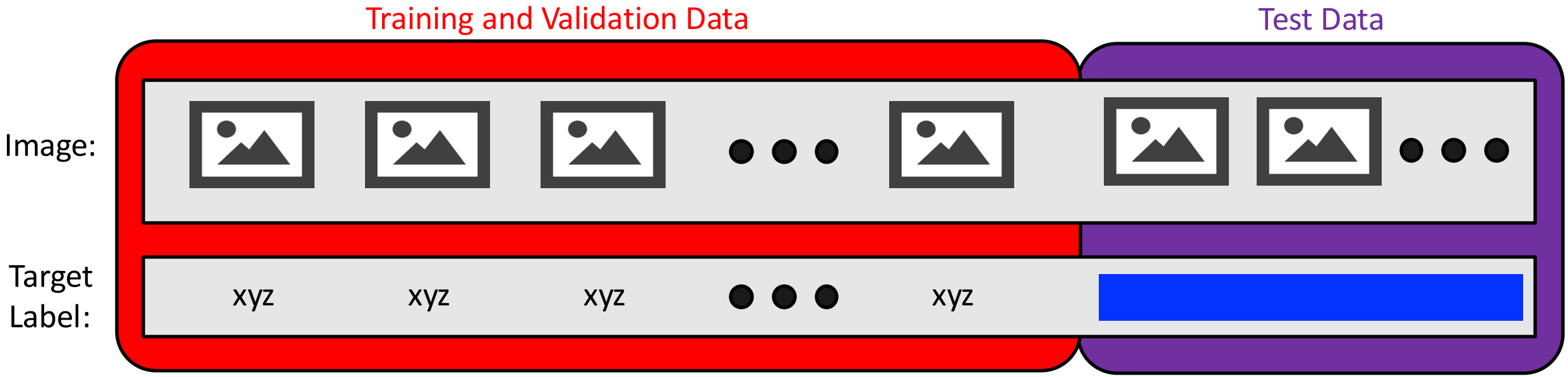
https://syncedreview.com/2019/06/18/cvpr-2019-attracts-9k-attendees-best-papers-announced-imagenet-honoured-10-years-later/

ImageNet: Inspired Era of Dataset Challenges

Hear from Fei-Fei Li how she began her career as an Assistant Professor creating ImageNet:

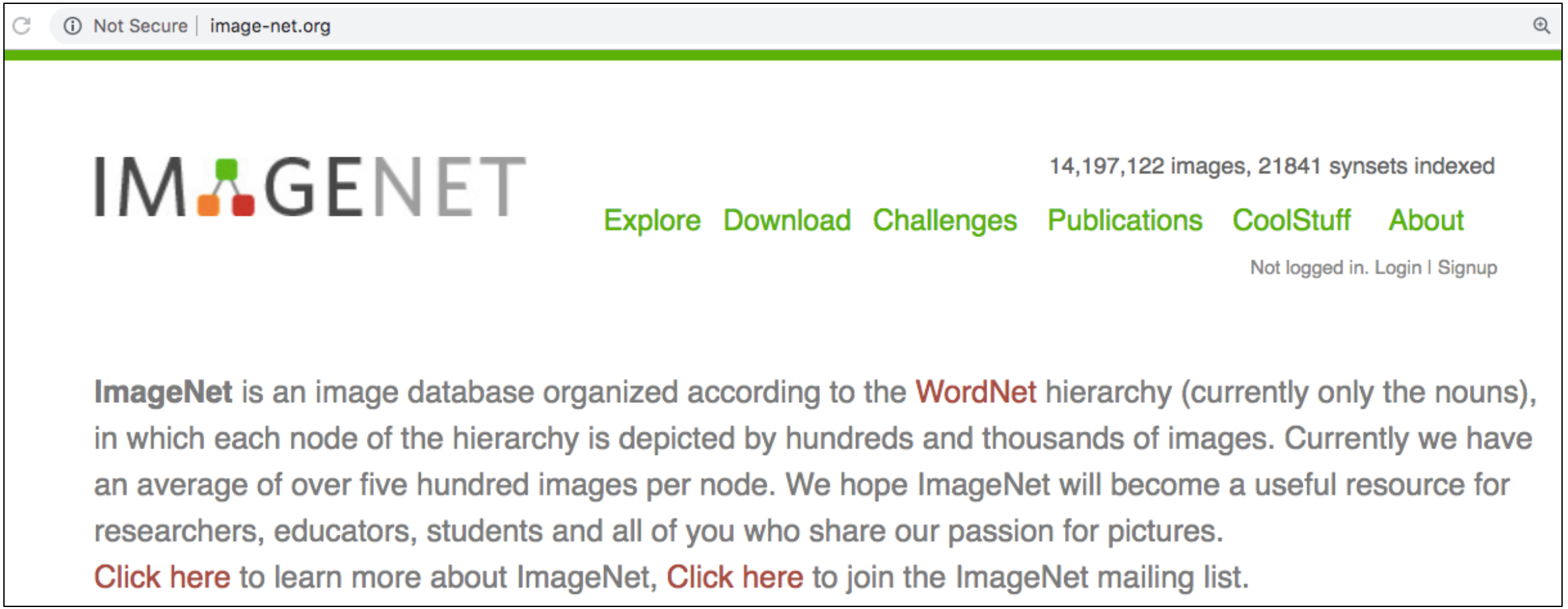
Video (5:44 – 9:35): https://youtu.be/40riCqvRoMs?feature=shared&t=343

ImageNet: Challenge



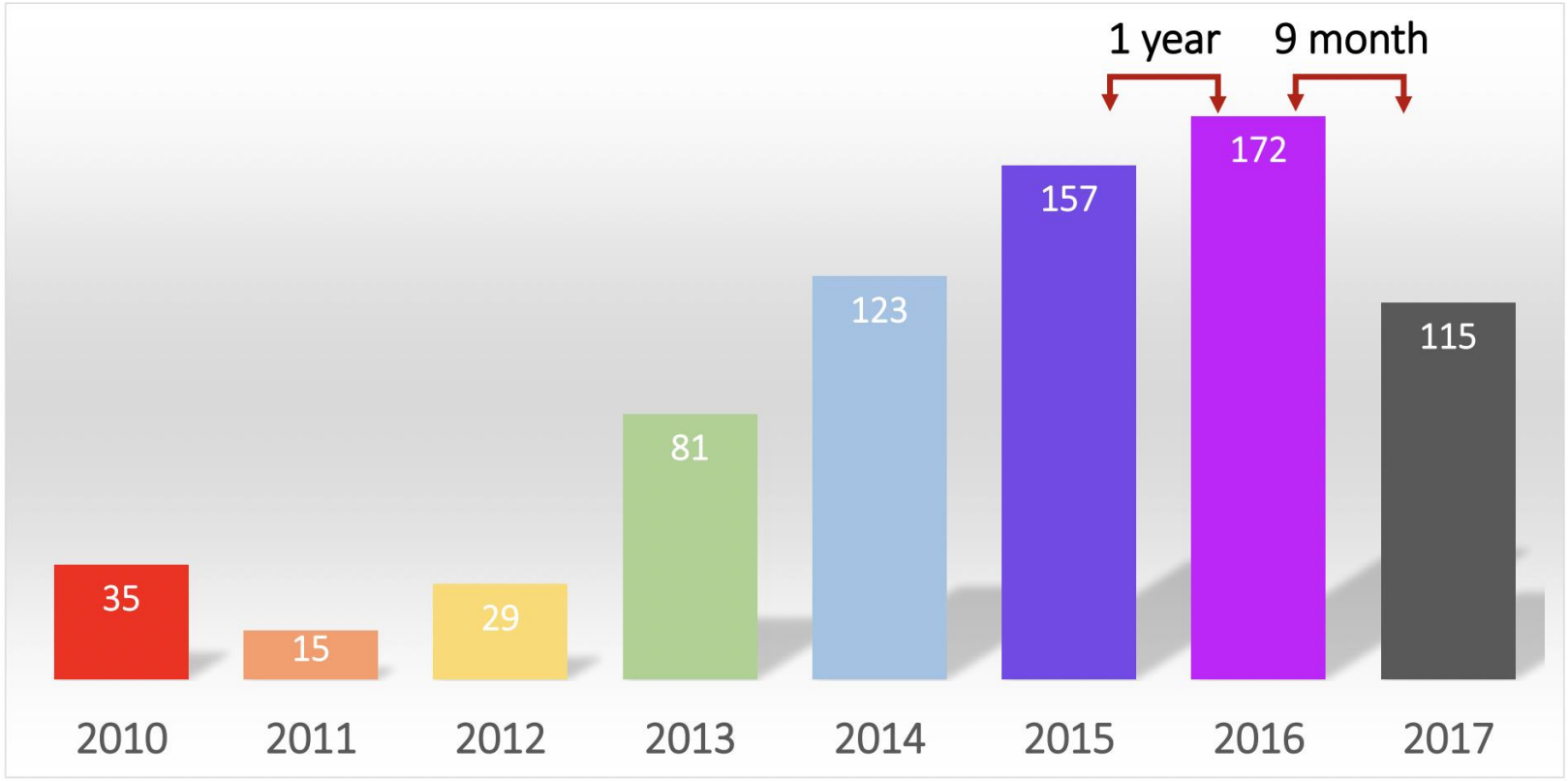
1. Dataset split into “training and validation sets” and a “test set”, with only the former shared publicly
2. Teams submit model predictions on the test set to evaluation servers that return performance score
3. Highest scoring model wins!

ImageNet: Website



ImageNet: Community Engagement by Announcing Winners at an Annual Workshop

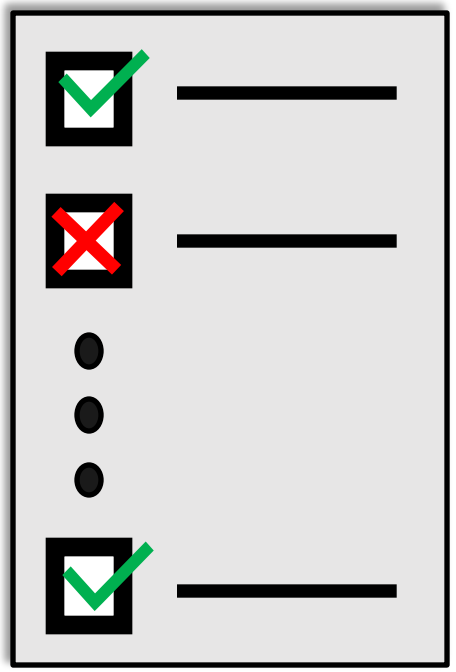
ImageNet: Community Engagement by Announcing Winners at an Annual Workshop



- 727 entries!! (plus an entry from Baidu that famously was kicked out in 2015 for cheating)
- Labor cost ~$110 million! : assuming 3 people contribute to each entry and $50k cost per person

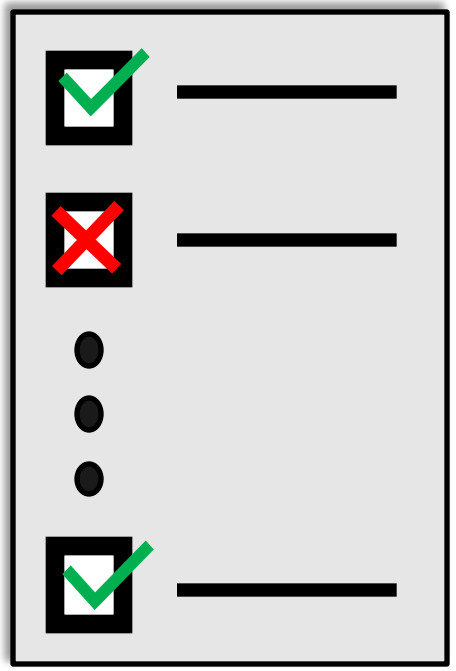
https://image-net.org/static_files/files/ILSVRC2017_overview.pdf

AI Progress Since 2012: Dataset Challenges



(Analogous to Tests in Schools, After Optionally Receiving Lessons)

AI Progress Since 2012: Dataset Challenges

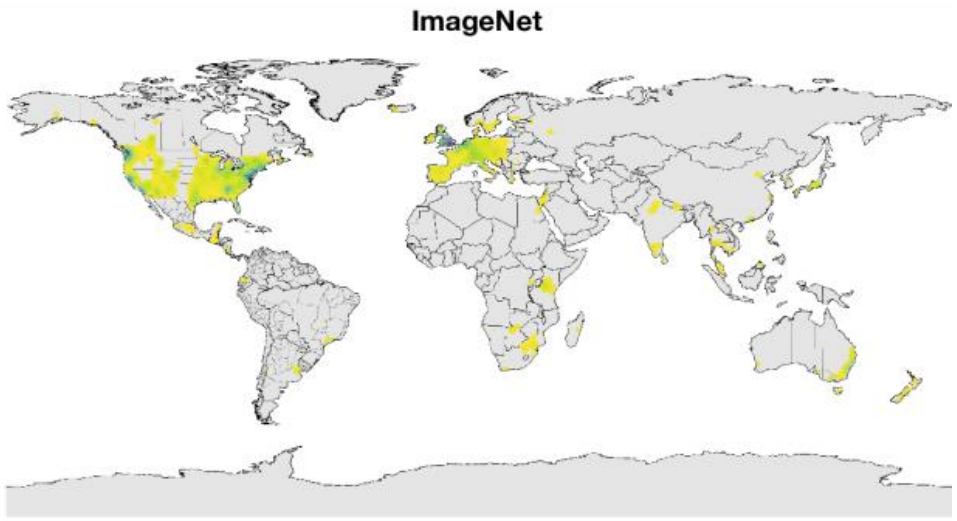


Key ingredients:

1. Test examples that includes target results
2. Metric for assessing the similarity between each model prediction and the target result
3. New challenges for the community to tackle, evidenced by dataset analysis and model benchmarking

e.g., Overcoming ImageNet's Limitations

Geographical origins of ImageNet images (based on Flickr metadata):



What's the limitation of ImageNet?

Dataset with greater diversity of countries & income levels:

Ground truth: Soap — **Nepal, 288 $/month**

Azure: food, cheese, bread, cake, sandwich
Clarifai: food, wood, cooking, delicious, healthy
Google: food, dish, cuisine, comfort food, spam
Amazon: food, confectionary, sweets, burger
Watson: food, food product, turmeric, seasoning
Tencent: food, dish, matter, fast food, nutriment

Ground truth: Soap — **UK, 1890 $/month**

Azure: toilet, design, art, sink
Clarifai: people, faucet, healthcare, lavatory, wash closet
Google: product, liquid, water, fluid, bathroom accessory
Amazon: sink, indoors, bottle, sink faucet
Watson: gas tank, storage tank, toiletry, dispenser, soap dispenser
Tencent: lotion, toiletry, soap dispenser, dispenser, after shave

Ground truth: Spices — **Phillipines, 262 $/month**

Azure: bottle, beer, counter, drink, open
Clarifai: container, food, bottle, drink, stock
Google: product, yellow, drink, bottle, plastic bottle
Amazon: beverage, beer, alcohol, drink, bottle
Watson: food, larder food supply, pantry, condiment, food seasoning
Tencent: condiment, sauce, flavorer, catsup, hot sauce

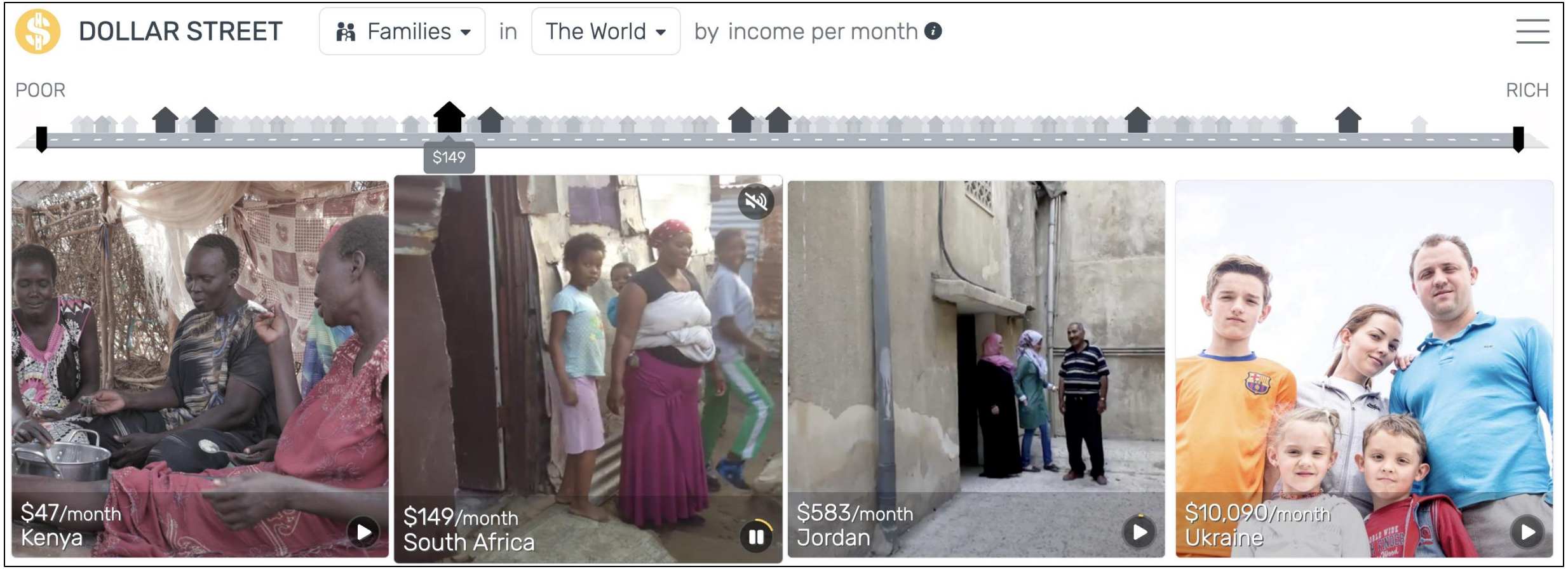
Ground truth: Spices — **USA, 4559 $/month**

Azure: bottle, wall, counter, food
Clarifai: container, food, can, medicine, stock
Google: seasoning, seasoned salt, ingredient, spice, spice rack
Amazon: shelf, tin, pantry, furniture, aluminium
Watson: tin, food, pantry, paint, can
Tencent: spice rack, chili sauce, condiment, canned food, rack

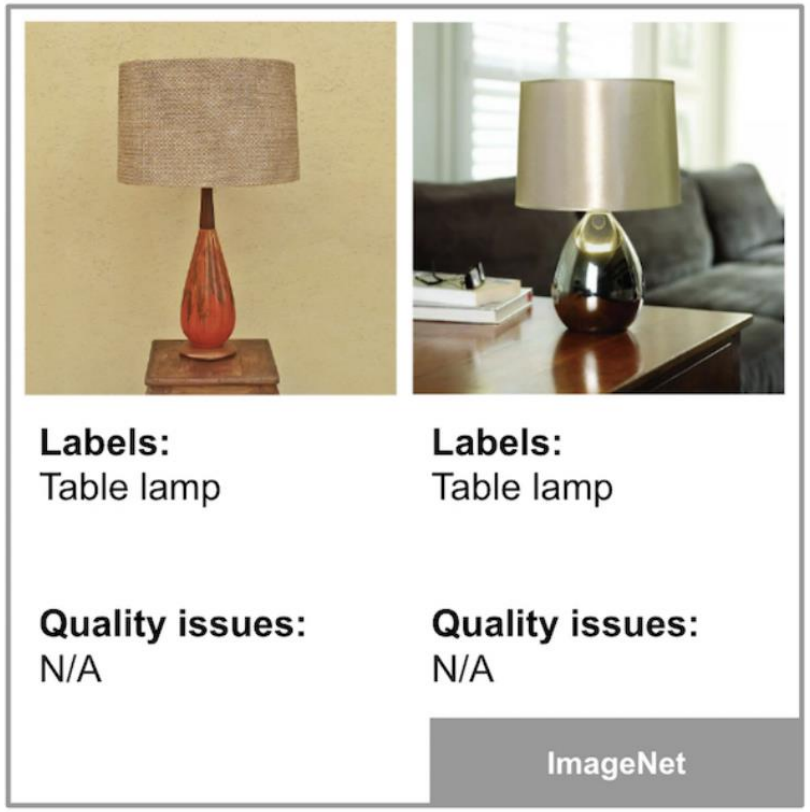
DeVries et al. Does Object Recognition Work for Everyone. CVPRW 2019

e.g., Overcoming ImageNet's Geographical/Income Limitations

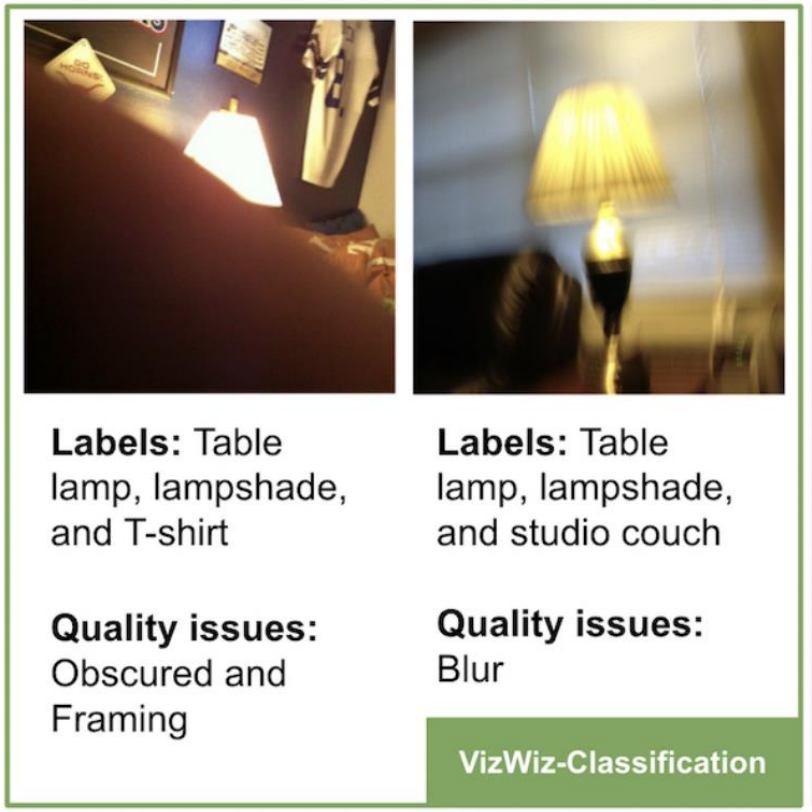
Browse the dataset: https://www.gapminder.org/dollar-street



e.g., Overcoming ImageNet's Photographer Bias



Who were the dominant photographers?

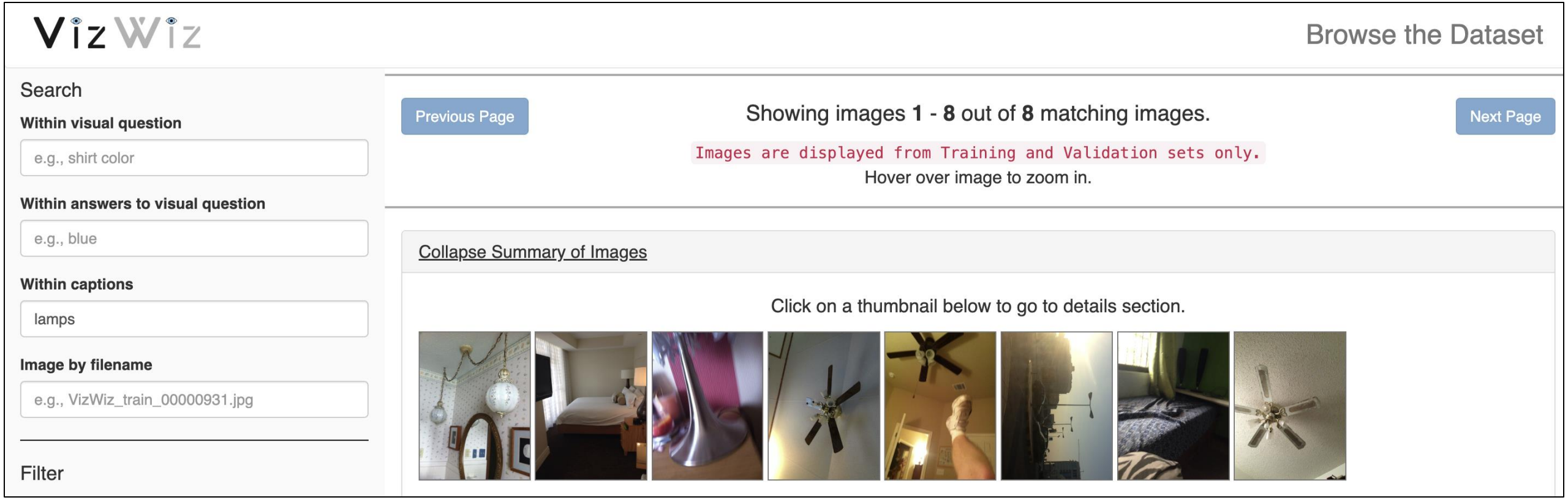


Dataset of same categories from blind photographers

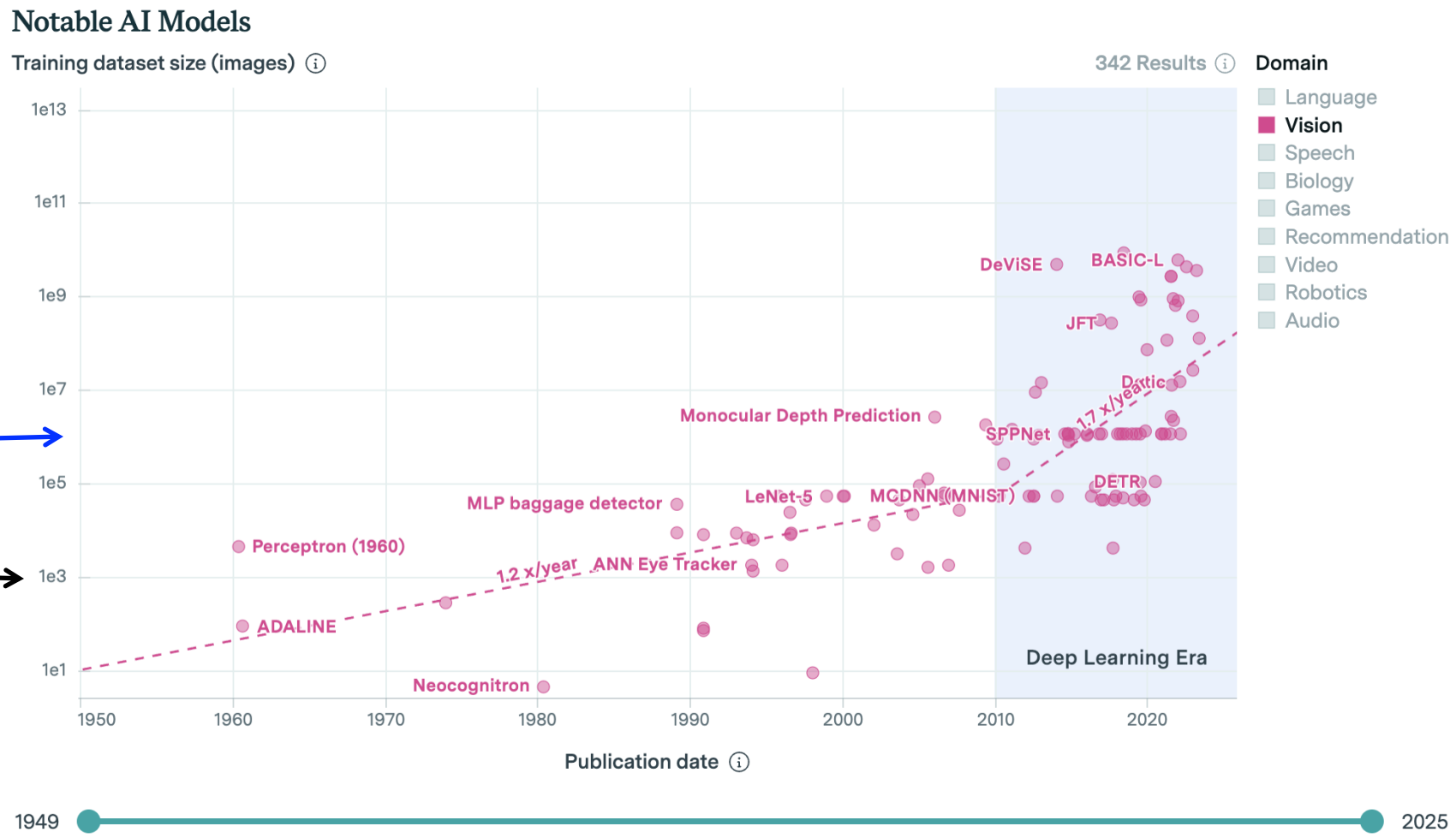
Akbarian Bafghi and Gurari. CVPR 2023

e.g., Overcoming ImageNet's Photographer Bias

Browse the dataset: https://vizwiz.cs.colorado.edu/VizWiz_visualization/view_dataset.php



e.g., Overcoming ImageNet's Limited Size (Scope: Models Trained from Scratch)



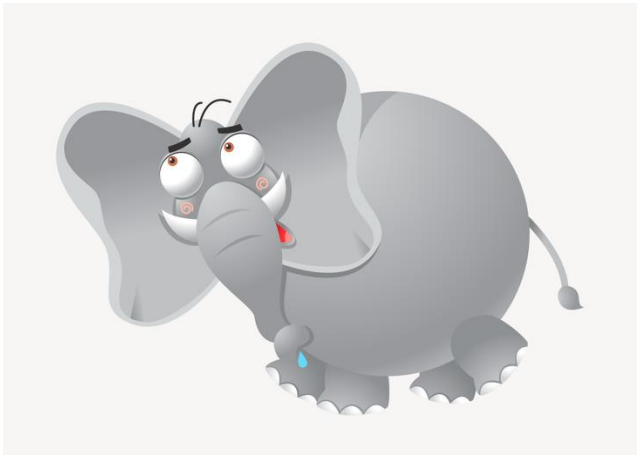
ImageNet's size →

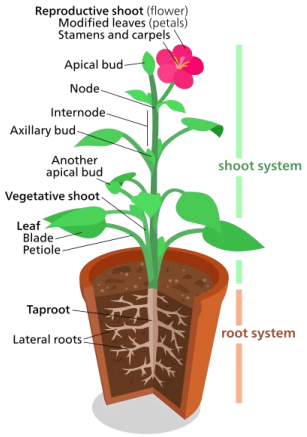
e.g., 1000 →

Interactive Demo: https://epoch.ai/data/notable-ai-models#explore-the-data

How to Create a Dataset? - Annotation

- What object is in the image? (i.e., object recognition task)



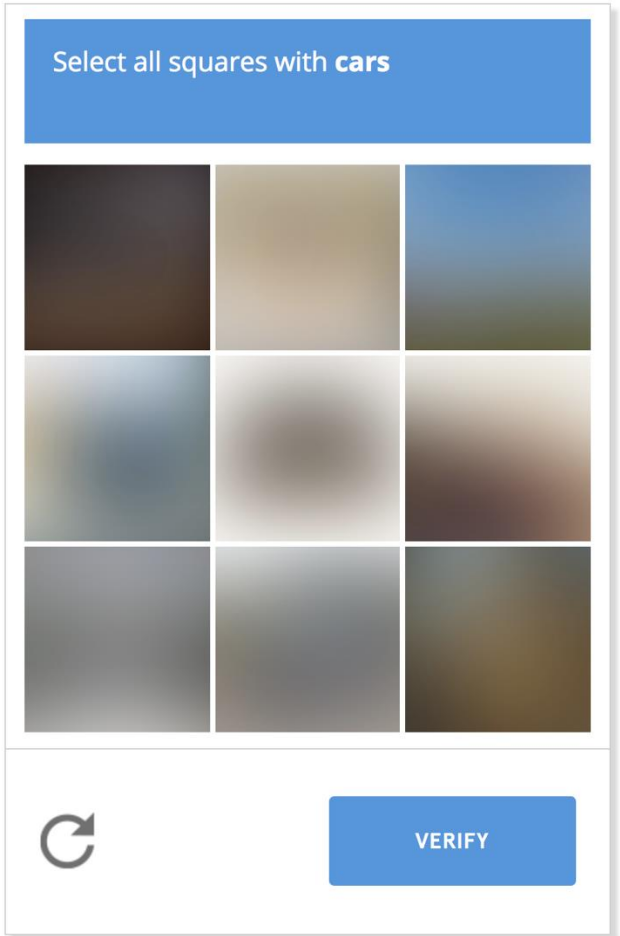


Data Annotation: How to Recruit Annotators?

- Security Questions
- Gamification
- Citizen Science
- Pay

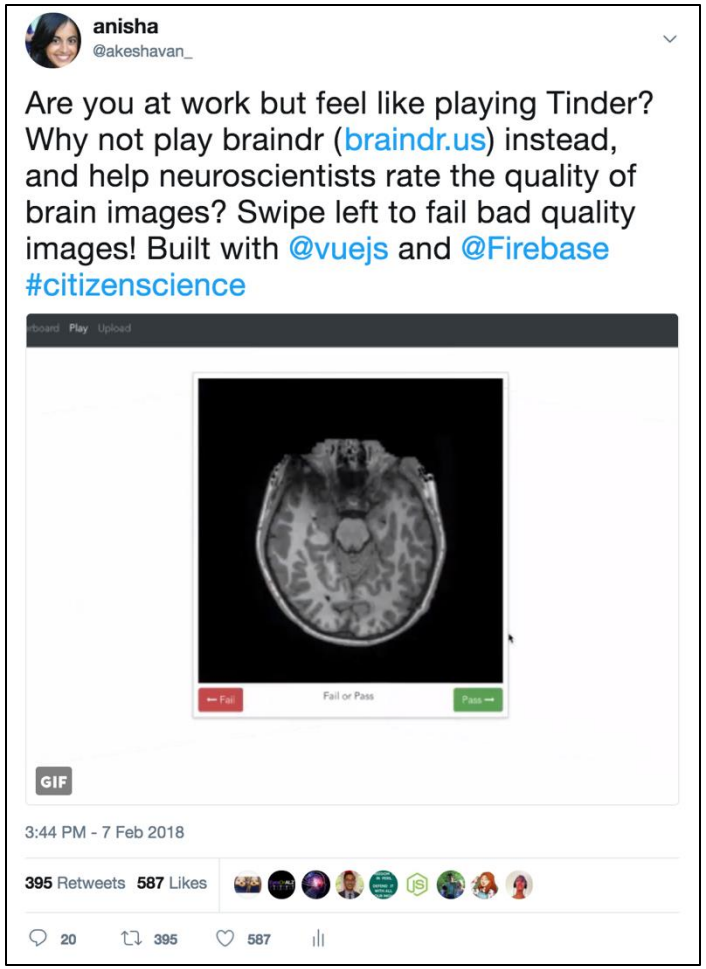
Data Annotation: How to Recruit Annotators?

- Security Questions
- Gamification
- Citizen Science
- Pay

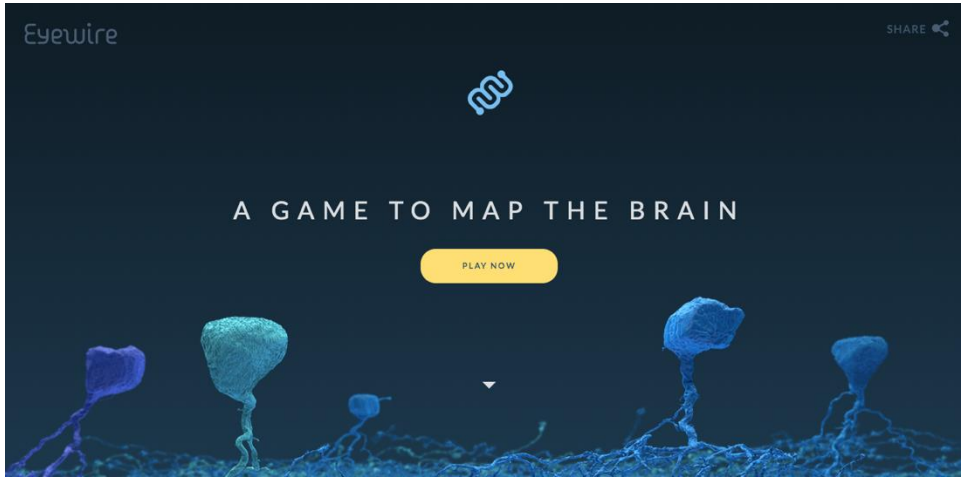


Data Annotation: How to Recruit Annotators?

- Security Questions
- Gamification
- Citizen Science
- Pay



[von Ahn & Dabbish; 2004]



Data Annotation: How to Recruit Annotators?

- Security Questions
- Gamification
- Citizen Science
- Pay

Data Annotation: How to Recruit Annotators?

- Security Questions
- Gamification
- Citizen Science
- Pay

Discussion: Which Annotators Would You Use When (and Why)?

- Security Questions
- Gamification
- Citizen Science
- Pay

Data Annotation: How To Collect High Quality Annotations?

Pre-Task

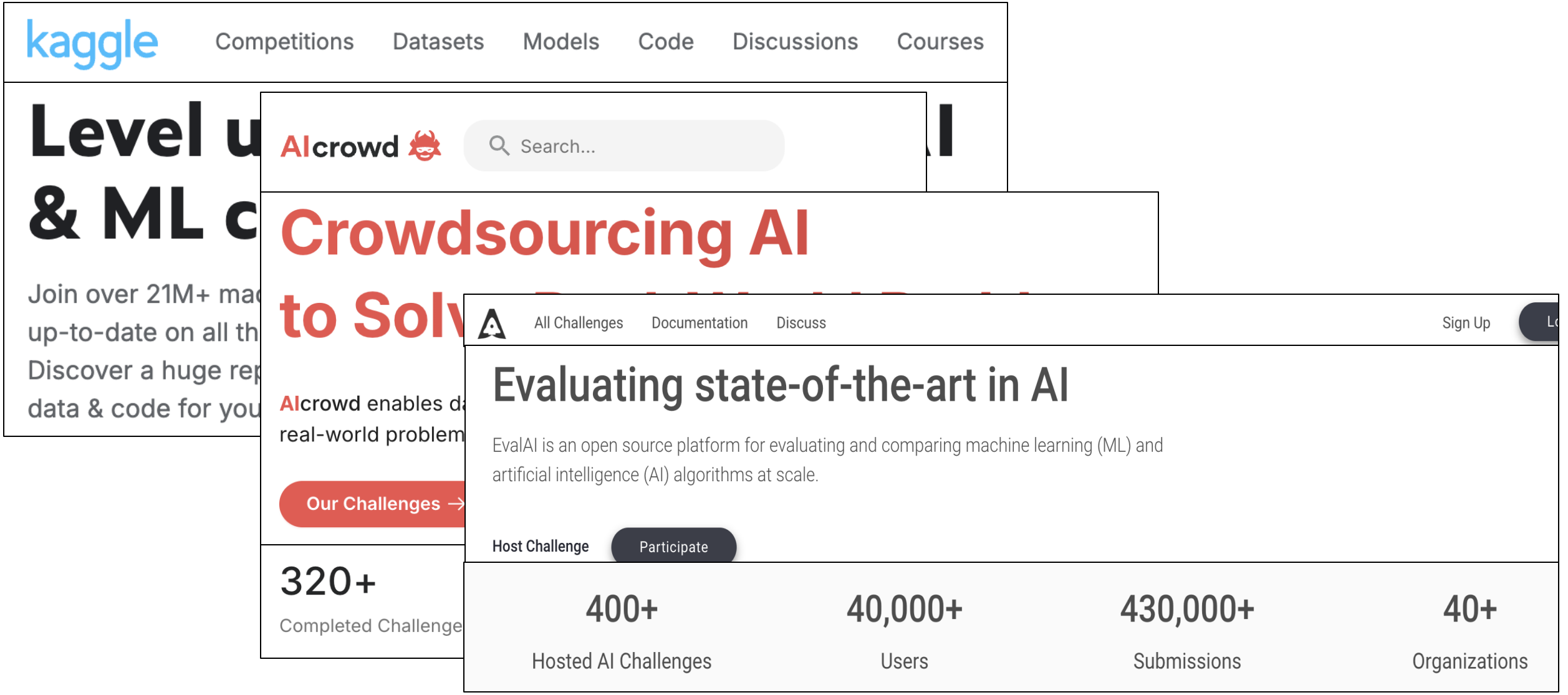
- Instructions
- Incentives
- Required qualification criteria
- Training/Qualification tests

Post-Task

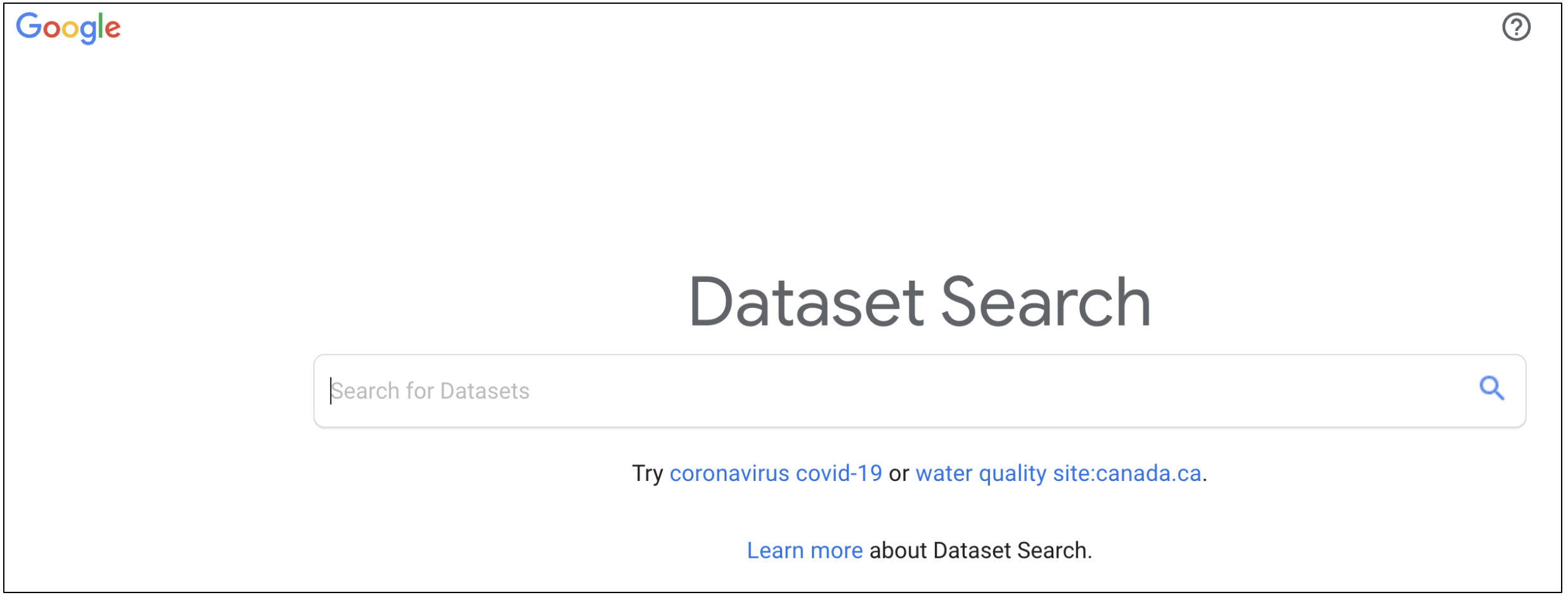
- Honey pot tasks
- Aggregate redundant responses
- Manual review
- Automated review

Chittilappilly et al. A Survey of General-Purpose Crowdsourcing Techniques. IEEE TKDE 2016.

How to Find Dataset Challenges (Which Include Evaluation Servers)



How to Find Dataset Challenges (Without Evaluation Servers); e.g.,



Summary: Steps to Launch a Dataset Challenge

1. Choose Problem: e.g., object recognition

2. Build Infrastructure: create dataset, establish evaluation metrics, and then host on evaluation server

3. Scale: market to target users, such as by offering a monetary prize

Today’s Topics

- Dataset challenges: before versus after 2012
- Hardware: before versus after 2012
- Programming tutorial

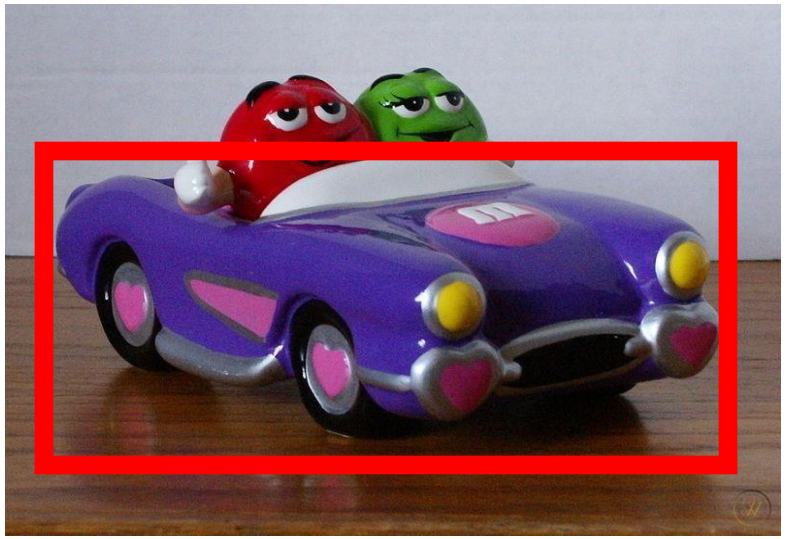
Recall: Key Ingredients for Training Success

An **algorithm** learns from **data**
on a **processor** the patterns that
will be used to make a prediction

Analogous to a Love Story of Partnering Up and Road Tripping Somewhere

Recall: Key Ingredients for Training Success

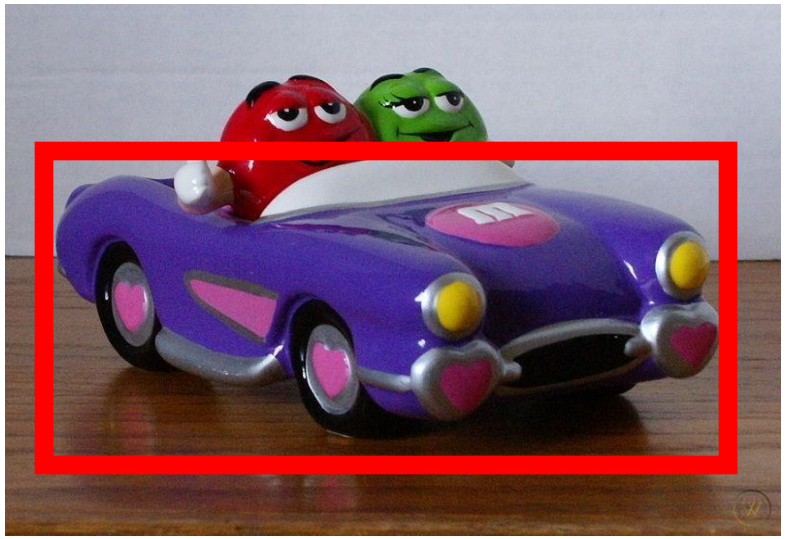
Key Issue: How Long Will It Take to Get There?



AlexNet Helped Inspire Use of Faster Hardware

e.g.,

GPUs (think Porsche) Instead of CPUs (think Golf Cart)



AlexNet Helped Inspire Use of Faster Hardware

Spot the CPU!
(central processing unit)

http://cs231n.stanford.edu/slides/2018/cs231n_2018_lecture08.pdf

AlexNet Helped Inspire Use of Faster Hardware

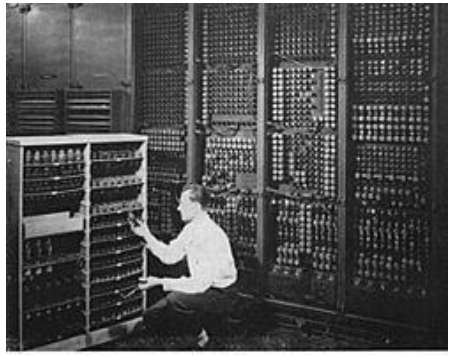
Spot the GPUs!
(graphics processing unit)

This image is in the public domain

http://cs231n.stanford.edu/slides/2018/cs231n_2018_lecture08.pdf

HW for Deep Learning: Historical Context

1945 — First machine: room-sized since components were based on large vacuum tubes



(ENIAC's vacuum tubes being changed)

1959 — First silicon-based "chip": silicon components wired together on a single "integrated circuit" (IC)

1965 — Moore's law established: # of transistors on IC doubles every 2 (initially 1) years

The transistors switch on and off to generate the 0s and 1s of binary notation

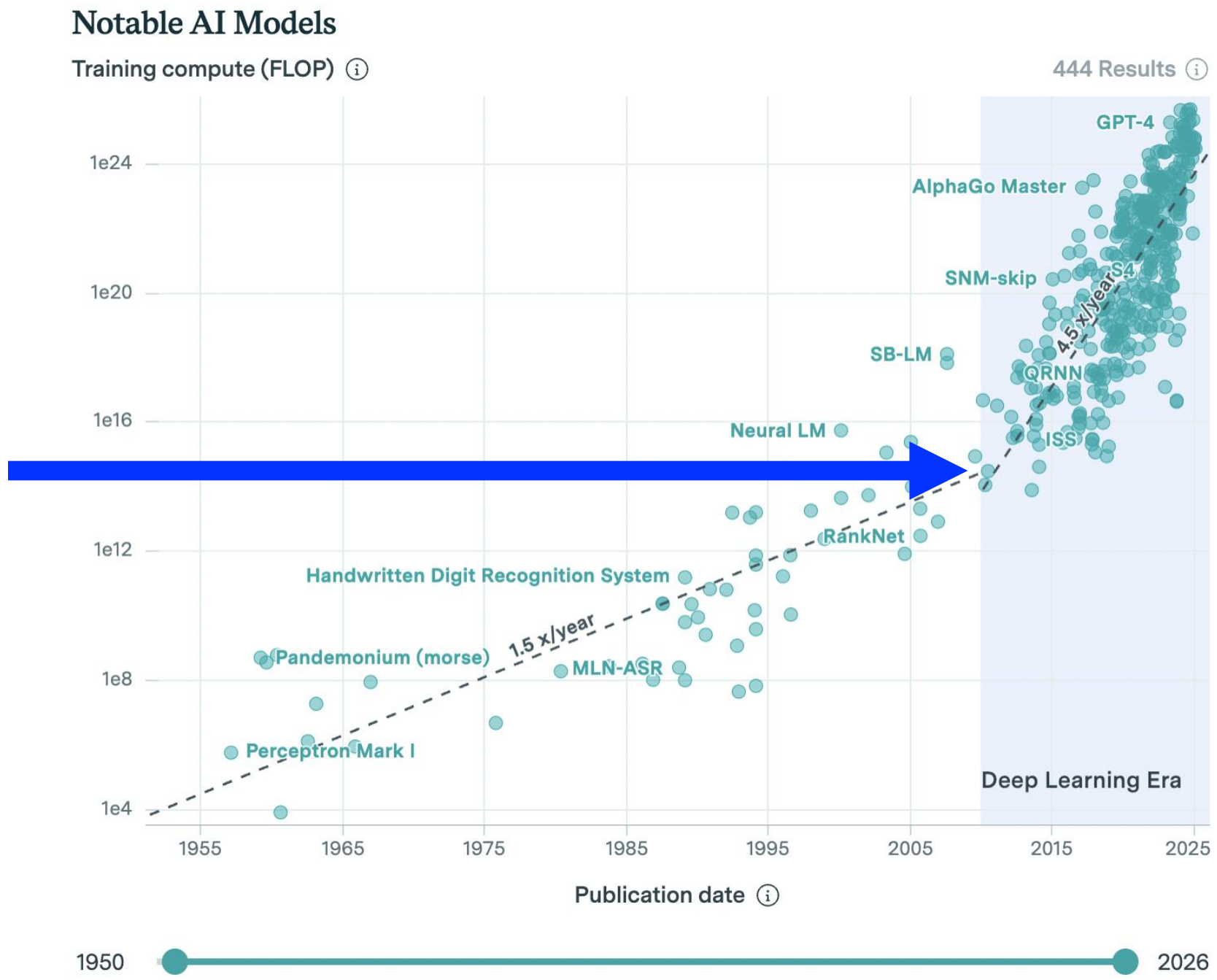
(e.g., 200 per mm^2 vs 150M per mm^2 in 1971 vs 2023 for Intel's MI300)

2010 — Computing for training AI models exceeds progress from Moore's law

https://en.wikipedia.org/wiki/Invention_of_the_integrated_circuit

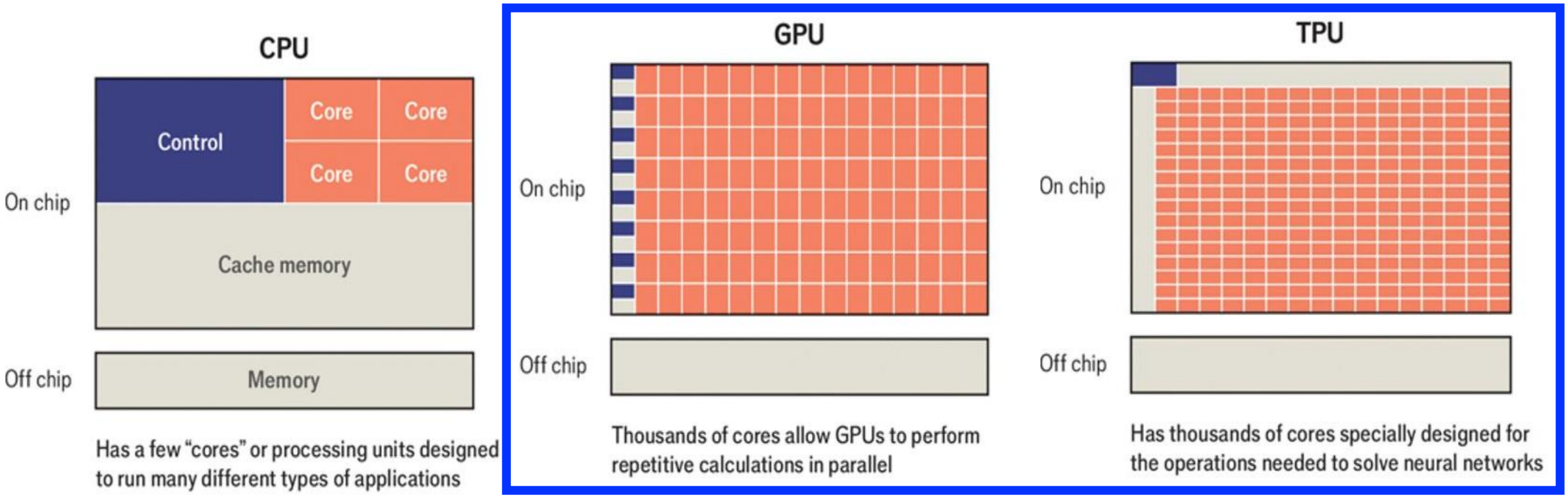
The Demand for Better HW

By 2012, the amount of computing needed to train AI models doubled every 6 months!



https://epoch.ai/blog/compute-trends

Rise of Specialized HW for Deep Learning; e.g.,

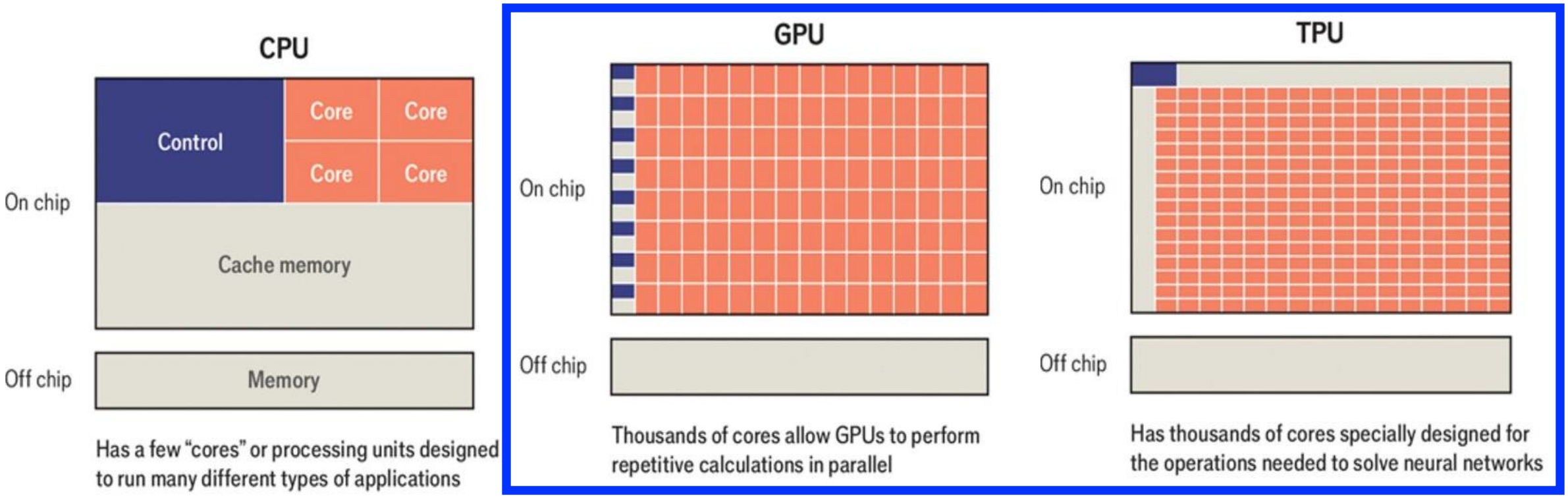


Graphics cards widely used for decades for rendering repurposed

Introduced by Google for internal use in 2015 and public use in 2018

Source: Tsung Tai Yeh, National Yang Ming Chiao Tung University; The Economist ; September 21, 2024

Rise of Specialized HW for Deep Learning; e.g.,

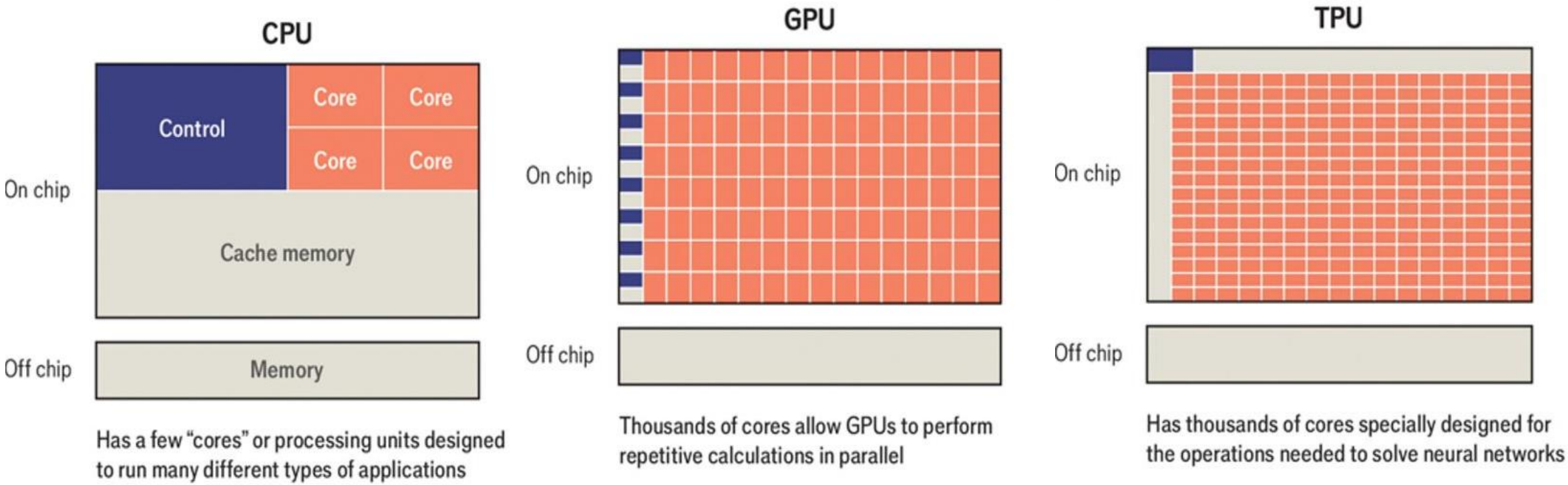


What are key architecture differences and their implications?

- Many more *parallel* processing cores accelerates workload
- Easier access to cache memory speeds up storing/retrieving intermediate computed results
- Reducing control increases speed while reducing ability to do various things

Source: Tsung Tai Yeh, National Yang Ming Chiao Tung University; The Economist ; September 21, 2024

Rise of Specialized HW for Deep Learning; e.g.,

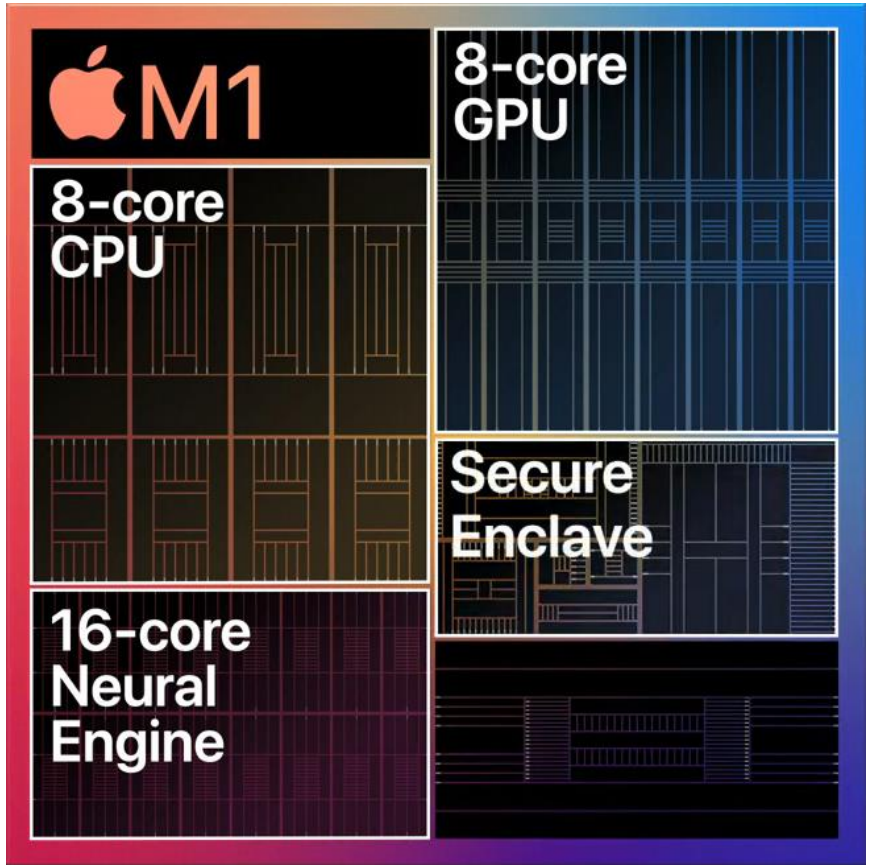


Source: Tsung Tai Yeh, National Yang Ming Chiao Tung University; The Economist

Rise of Specialized HW for Deep Learning; e.g.,

- Alternatively, one chip can do it all (e.g., “Neural Processing Unit” or NPU)

Added by Apple in 2017 to their chips, a dedicated block for NN operations to benefit users of their devices (e.g., phones, iPads, laptops)

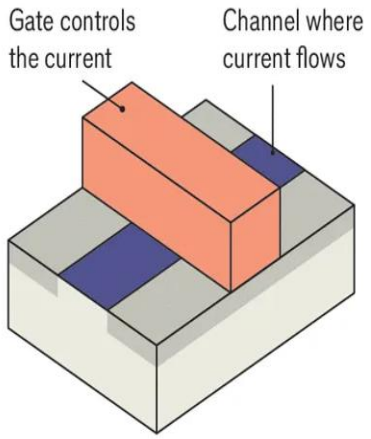


https://apple.fandom.com/wiki/Neural_Engine?file=Apple_M1_simplified_schematic.jpg

Rise of Specialized HW for Deep Learning

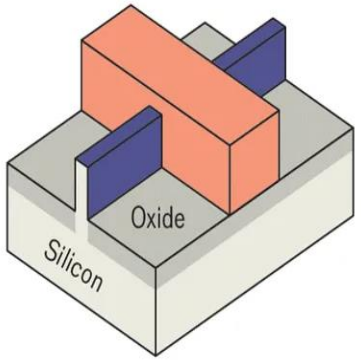
- Many companies are innovating to address the next key challenges:
  - Getting more transistors on a chip while minimizing energy consumption and waste heat
  - Less expensive factories to build chips (e.g., $20bn for one to be built in Arizona by TSMC)

Planar transistor, pre-2011



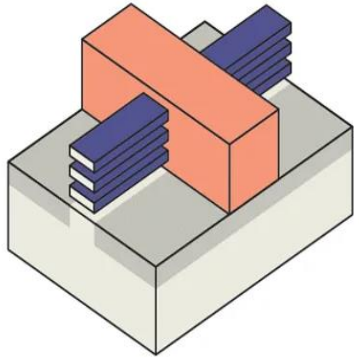
Issue: current leaks in small transistors (e.g., 90 nm) from quantum effects, even when they are off

FinFET, 2011



Introduced in 2011 by Intel, channels embedded in gates on 3 sides achieved better control of current flow and so less leakage

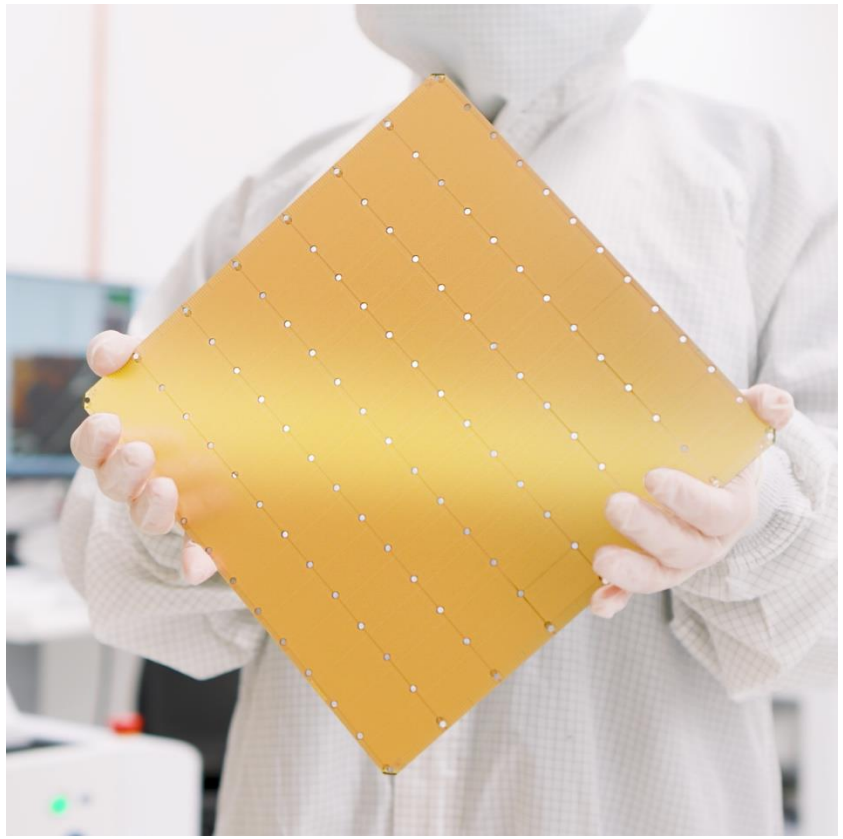
Gate-all-around, 2022



Introduced in 2022 by Samsung, gates surrounded channels on all sides to achieve better control of current flow

The Economist; September 21, 2024

State-of-the-Art in Chip (in the World!)



- Approximately size of a dinner plate, so all components can be on a single chip
- 900,000 cores
- 4 trillion transistors
- 44 GB of on-chip memory
- Rumored to cost $2-3 million

https://cerebras.ai/product-chip/

Blackwell: Nvidia's Best Chip, Released in 2025

- 208 billion transistors on 2 connected dies
- Each die surrounded by 4 blocks of memory, totaling at 192 GB
- Over 1 year, it would consume ~1/2 the energy of an average American household (5.2 MWhrs)

Do you think a Blackwell costs (A) $7,000, (B) $70,000, or (C) $700,000

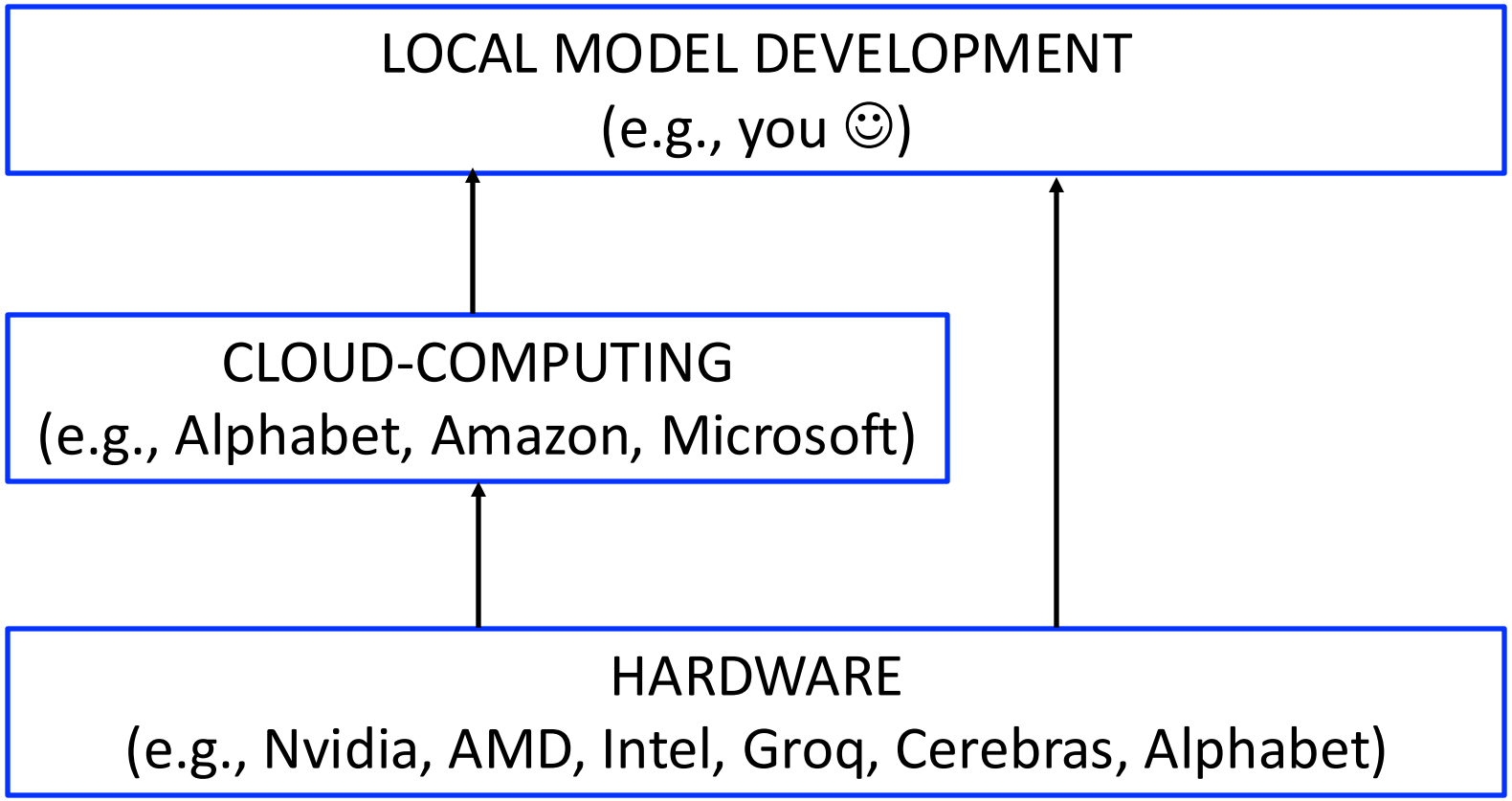
https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing

Compute Cost for Training Models from Scratch

e.g., $1000 →

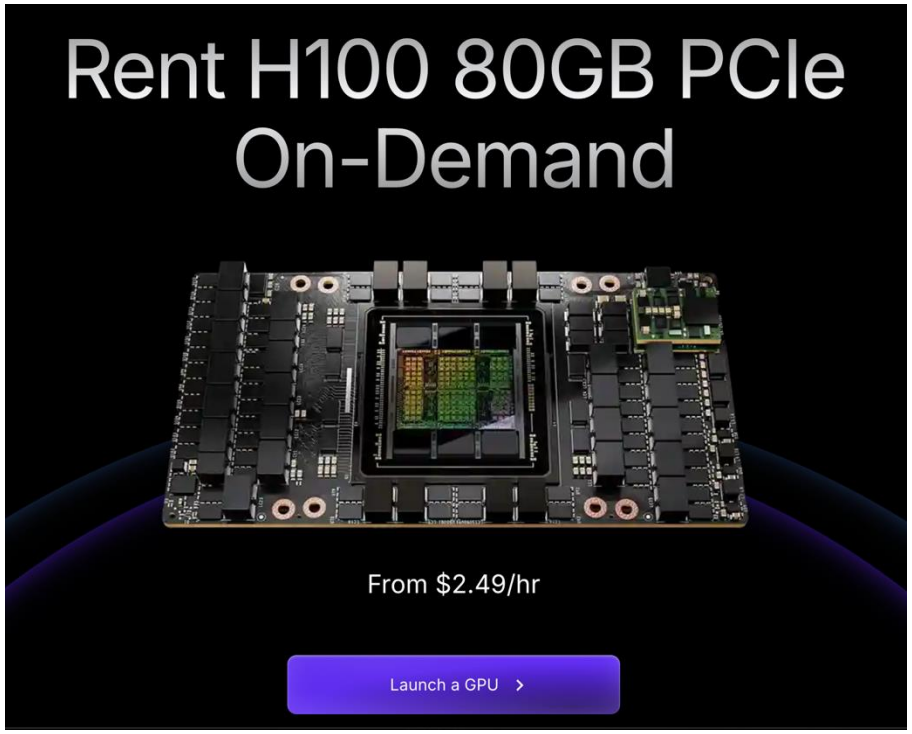
Interactive Demo: https://epoch.ai/data/notable-ai-models#explore-the-data

How to Obtain Hardware? – Rent Versus Buy



How to Obtain Hardware? – Rent Versus Buy

e.g., in 2025, buy H100 (predecessor to Nvidia's Blackwell) at ~$40k versus rent at $2.49 per hour (~$1800 per month of 24/7 use)



https://www.runpod.io/gpu/h100-pcie

Congratulations!

- By taking this class, you receive gifts of cloud credits worth over:

- Thanks to: Google Jetstream2

Today’s Topics

- Dataset challenges: before versus after 2012
- Hardware: before versus after 2012
- Programming tutorial

Today’s Topics

- Dataset challenges: before versus after 2012
- Hardware: before versus after 2012
- Programming tutorial


The End