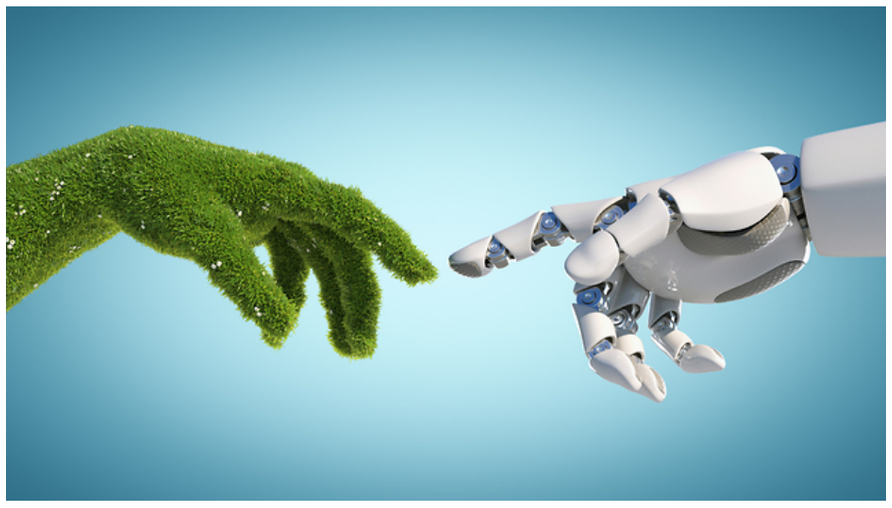
Green AI

Roy Schwartz

Allen Institute for AI/
University of Washington

Hebrew University
of Jerusalem



THE HEBREW
UNIVERSITY
OF JERUSALEM

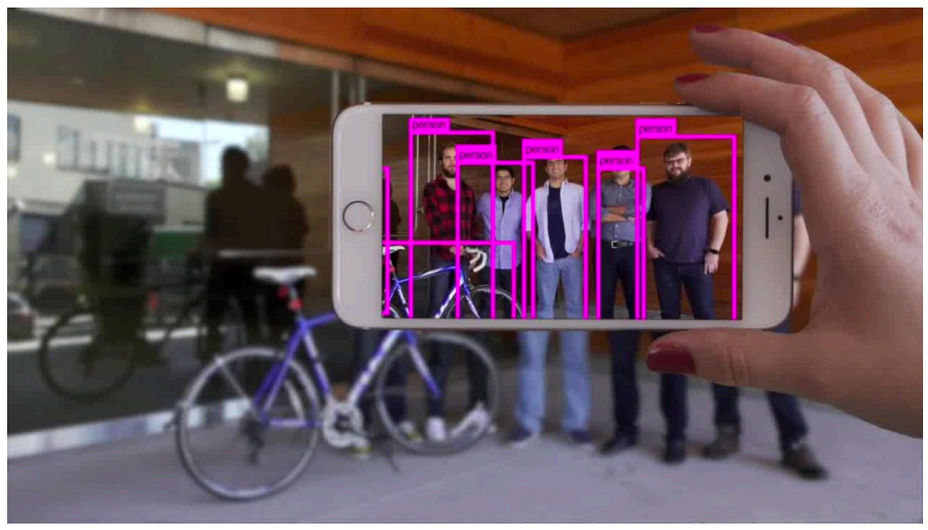
A Little about Me

Roy Schwartz

- Research Scientist at AI2 (AllenNLP) and UW
  - Will be joining the School of CS at the Hebrew U. of Jerusalem in 09/2020
- I study Natural Language Processing (NLP)
  - **Understanding** NLP models
  - Revealing artifacts and **biases** in datasets
  - Making models more **efficient**

AI Today

Microsoft Translator



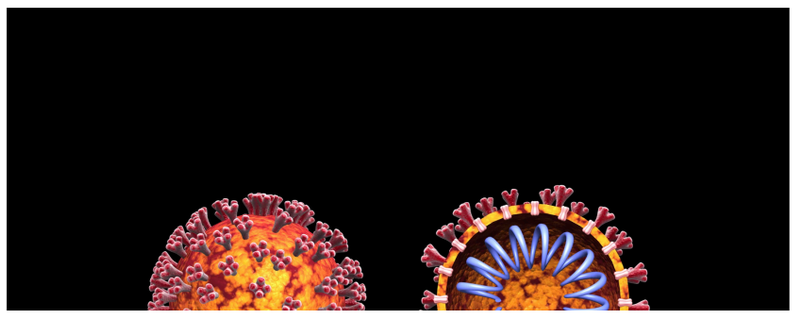
AI and COVID-19

WIRED BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY TRANSPORTATION SIGN IN | SUBSCRIBE

OREN ETZIONI NICOLE DECARIO IDEAS 03.28.2020 09:00 AM

AI Can Help Scientists Find a Covid-19 Vaccine

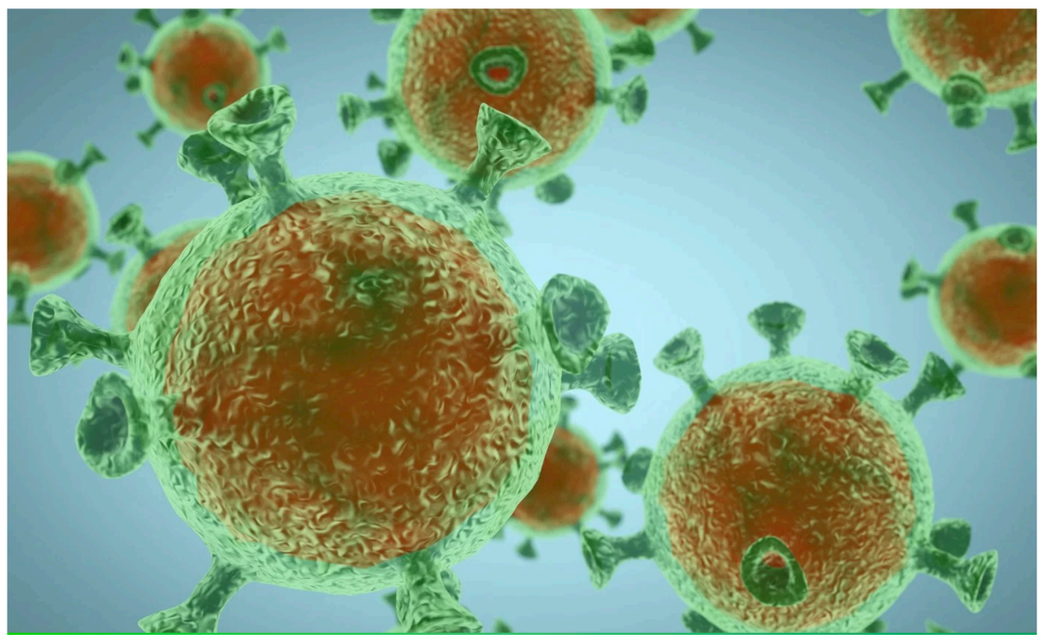
Artificial intelligence has already played a vital role in the outbreak since day 1—a reminder for the first time in a while that it can be a tool for good.



Potential new treatment for COVID-19 uncovered by BenevolentAI enters trials

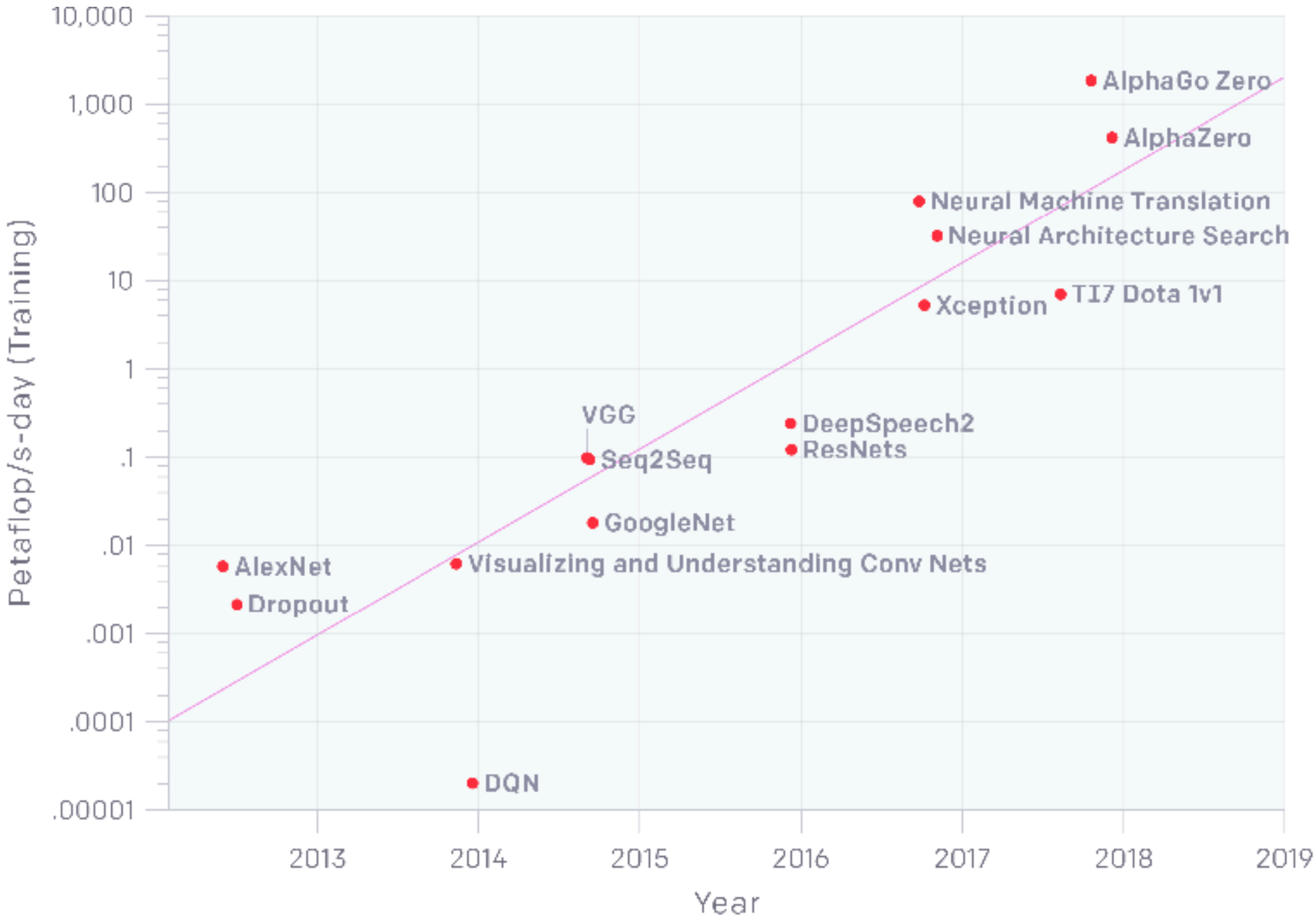
Mike Butcher @mikebutcher / 8:05 am PDT • April 14, 2020

Comment



Big Models

300,000X in 6 Years



Big and Expensive Models

Problems with Big Models

Inclusiveness

Synced

AI TECHNOLOGY & INDUSTRY REVIEW

FEATURE INDUSTRY TECHNOLOGY COMMUNITY ABOUT US REPORT CONTRIBUTE TO SYNCED REVIEW

https://syncedreview.com/2019/06/27/the-staggering-cost-of-training-sota-ai-models/

Problems with Big Models

Adoption

NCE MACHINE LEARNING PROGRAMMING VISUALIZATION AI PICKS MC

Too big to deploy: How GPT-2 is breaking servers

A look at the bottleneck around deploying massive models to production

Caleb Kaiser Follow

Jan 31 · 7 min read

https://towardsdatascience.com/too-big-to-deploy-how-gpt-2-is-breaking-production-63ab29f0897c

Problems with Big Models

Environment

Consumption	CO_2e (lbs)
Air travel, 1 person, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

Strubell et al. (2019)

Green AI

Schwartz*, Dodge*, Smith & Etzioni (2019)

- Red AI
  - Inclusiveness, adoption, environment
- Green AI
  - Enhance reporting of computational budgets
    - Add a *price-tag* for scientific results
  - Promote efficiency as a core evaluation for AI
    - **In addition to** accuracy

Outline

- Red AI

  - Why here? why now?
  - Big models, large datasets

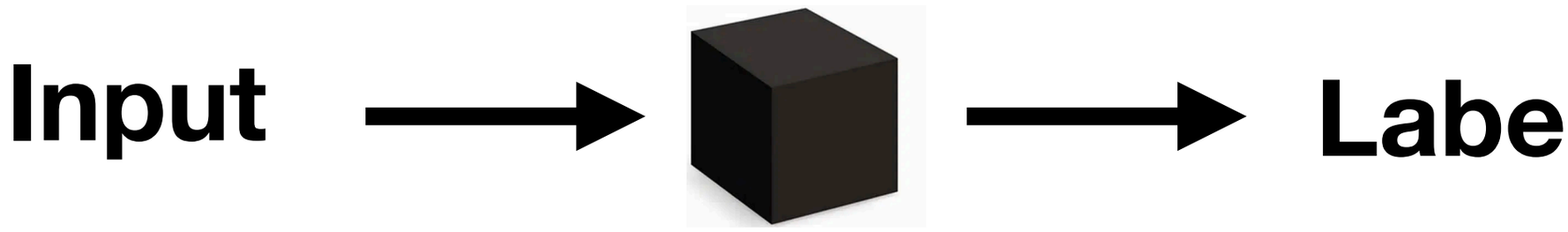
- Green AI

  - Reporting, efficiency

Red AI is Important

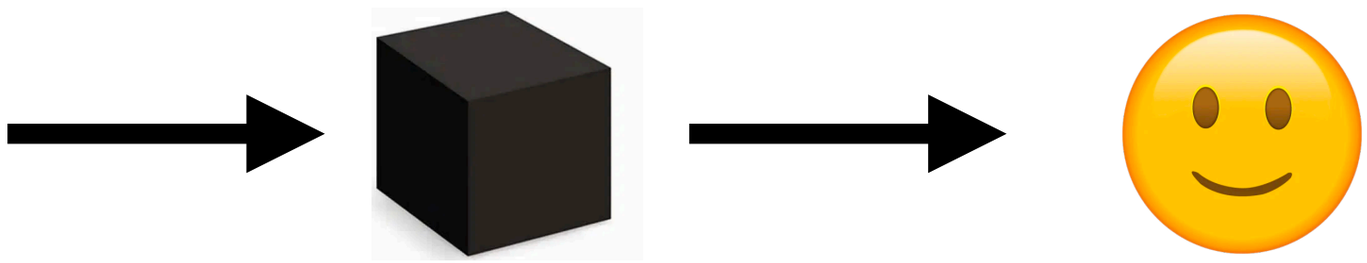
- Push the limits of the state of the art
- High training costs can be amortized by publicly releasing large models
- But, Red AI has **concerning side affects**
  - Inclusiveness, adoption, environment
- Our goal is to **mitigate these side affects**

Supervised Learning



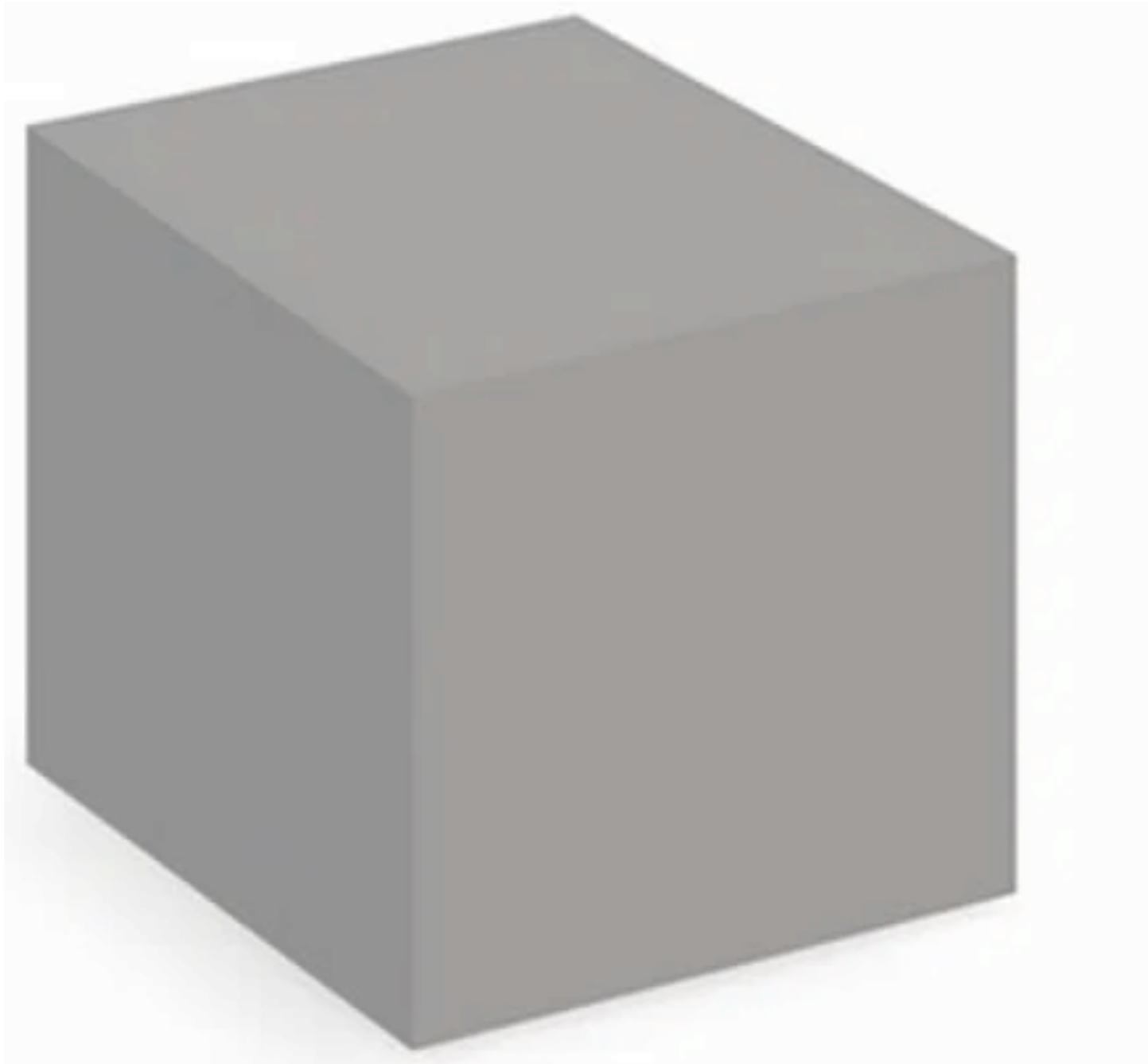
Example:
Sentiment Analysis

“What a great presentation”



Model

- A lot of numbers (***weights*** *or* ***parameters***)
- Computation over these weights



Learning

- Assigning values to the **weights** of the model

- Requires labeled data

Labeled Data

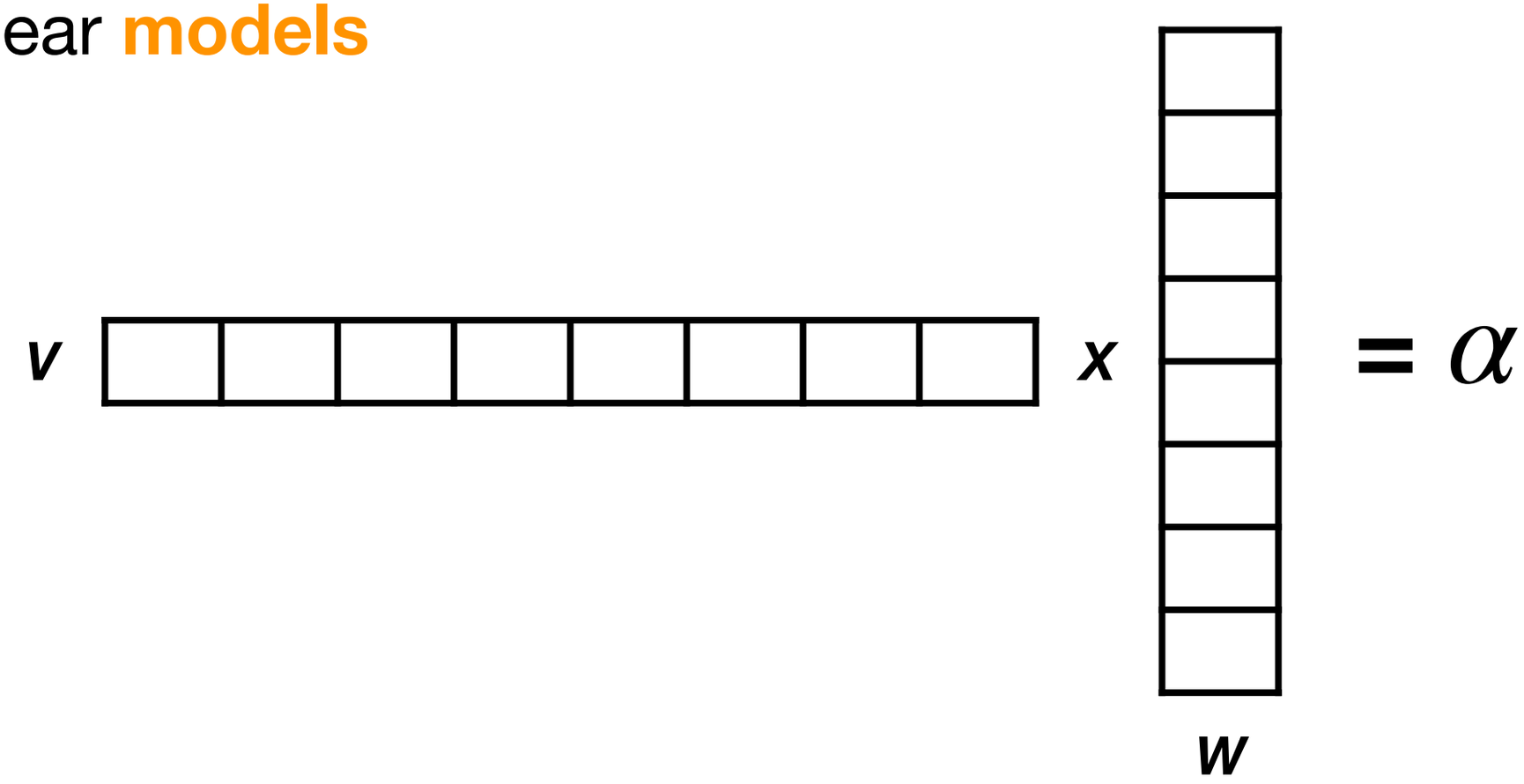
- *What a great presentation* 🙂
- *I didn’t like that movie* ☹️
- *Washington state seems to be flattening its curve* 🙂
- *…*

- Rule of thumb: the more weights the model has, the more data it needs

Classical Machine Learning

~1995-2010

- Small **models**
  - Relatively few weights
  - One set of computation
- E.g., linear **models**

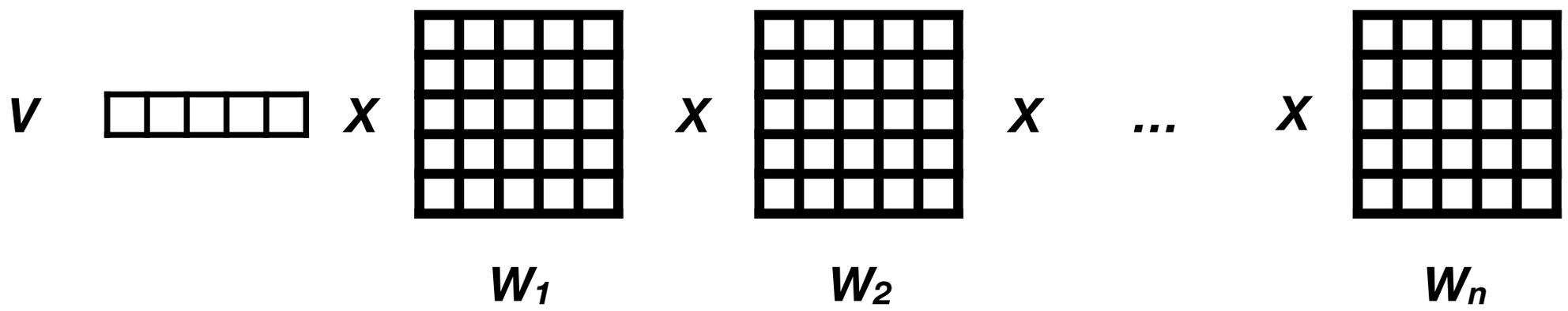


Deep Learning 1.0

~1980-2010

- **Deep** hidden representation

- Multiply multiple weight matrices



More Weights

- Higher model capacity
  - Potentially more accurate
- **But**
  - Performing many matrix multiplications is **slow**
  - The more weights the model has, the more data it needs

Deep Learning 2.0
Graphical Processing Units (GPUs)

- Originally designed for graphical displays
- **Efficient and parallel matrix multiplication**
- General purpose GPUs

Deep Learning 2.0

Crowd Sourcing

- **Cheap** and **fast** platform to collect **large amounts** of data

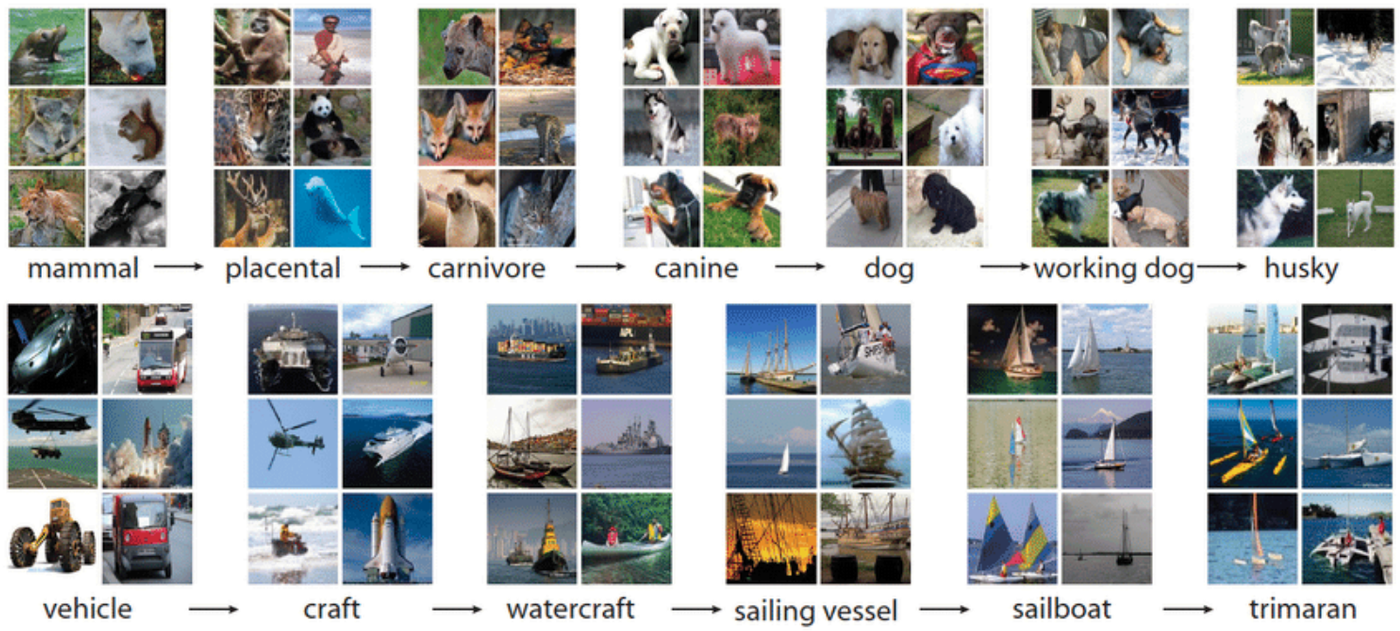
CrowdFlower

amazon mechanical turk

AlexNET

Krizhevsky, Sutskever & Hinton, 2012

IMAGENET



Best classical ML model

Best AlexNet deep model

Model	Top-1 (val)	Top-5 (val)	Top-5 (test)
SIFT + FVs [7]	—	—	*26.2%*
1 CNN	40.7%	18.2%	—
5 CNNs	38.1%	16.4%	**16.4%**
1 CNN*	39.0%	16.6%	—
7 CNNs*	36.7%	15.4%	**15.3%**

AlexNET
Perspective

Usain Bolt, 100m finals, Berlin 2009,
WR **9.69** -> **9.58**

Steph Curry, Warriors vs. Knicks, Feb 2013,
54 points, 11/13 3-pointers

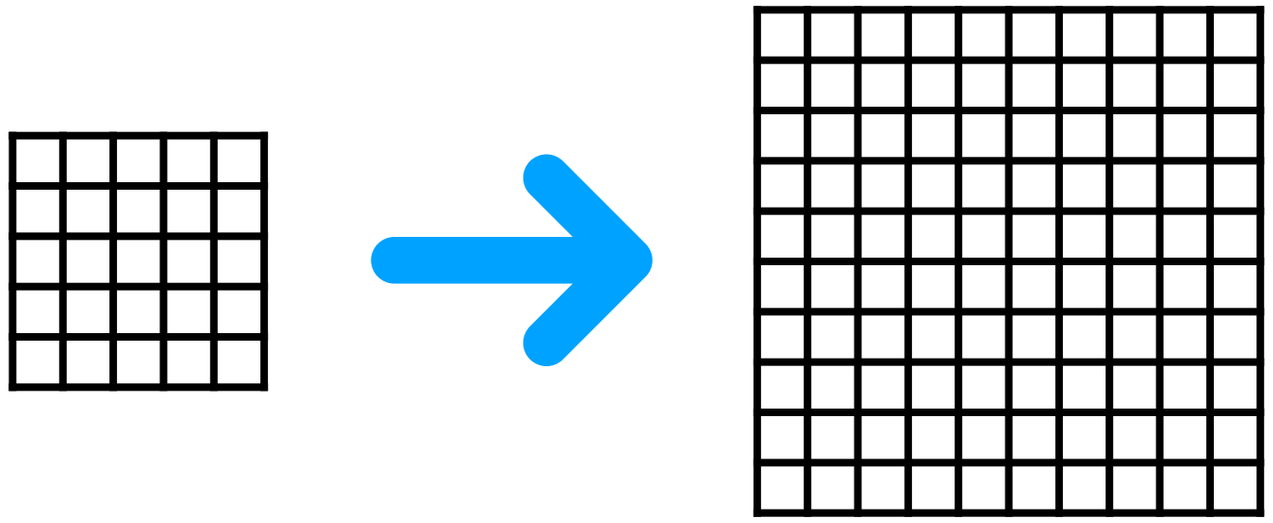
Deep Learning Takes Over

2012—now

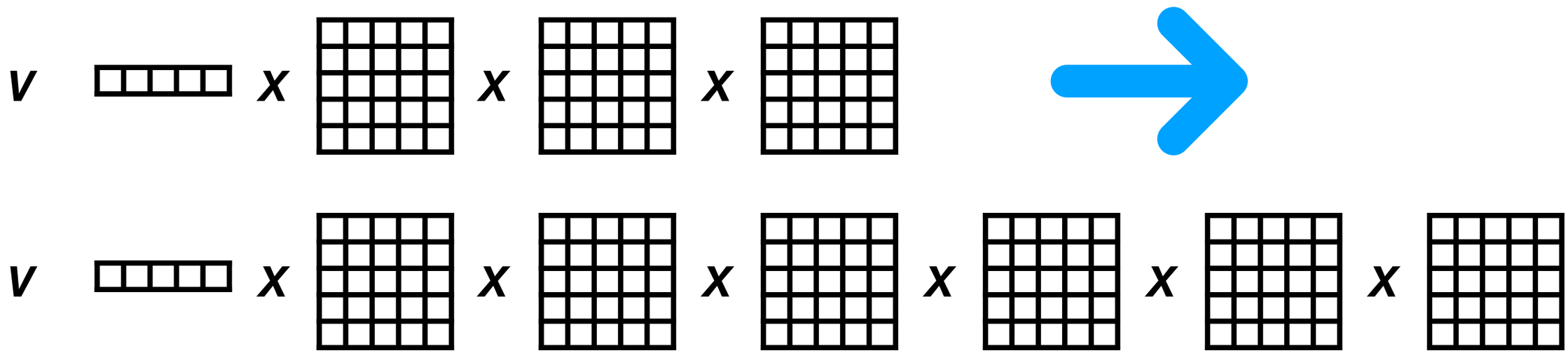
- Computer vision
- Speech recognition
- Natural language processing
- Bioinformatics
- recommendation systems
- …

Models are Getting Bigger

Wider



Deeper



Red AI

☑ Bigger models

Datasets are Getting Larger

- The more weights the model has, the more data it needs
- Large datasets released
  - Stanford Question Answering Dataset (SQUAD; Rajpurkar et al. 2016): **100,000** samples
  - CNN/DailyMail Dataset (Hermann et al., 2015): **300,000** samples
  - The Multi-Genre Natural Language Inference (MultiNLI, Williams et al., 2018): **400,000** samples

Red AI

☑ Bigger models

☑ Larger datasets

How Large can Datasets Grow?

- Datasets require manual annotation

- Even crowd sourcing **doesn’t scale**

Naturally Occurring Supervision

Language Modeling

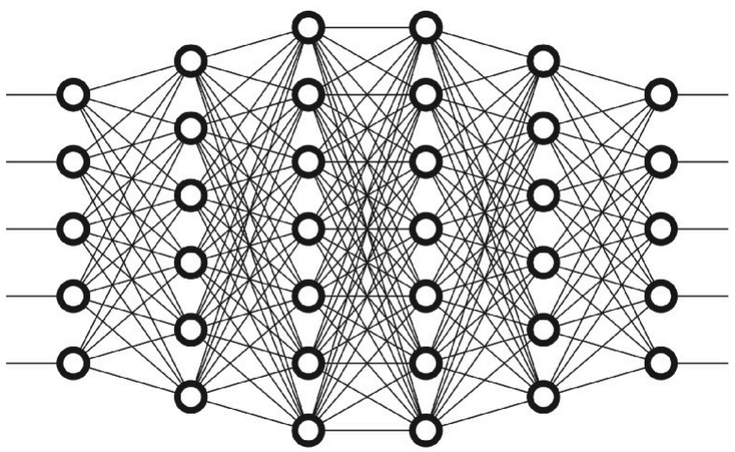
- The boy read a _____
  - book? story? ~~giraffe~~, ~~the~~, ~~apple~~
- The presentation is _____
  - interesting? funny? boring? long? ~~tree~~, ~~paper~~, ~~tall~~
- **No need for manual annotation!**

Deep Learning 3.0

Peters et al., 2018

Language model

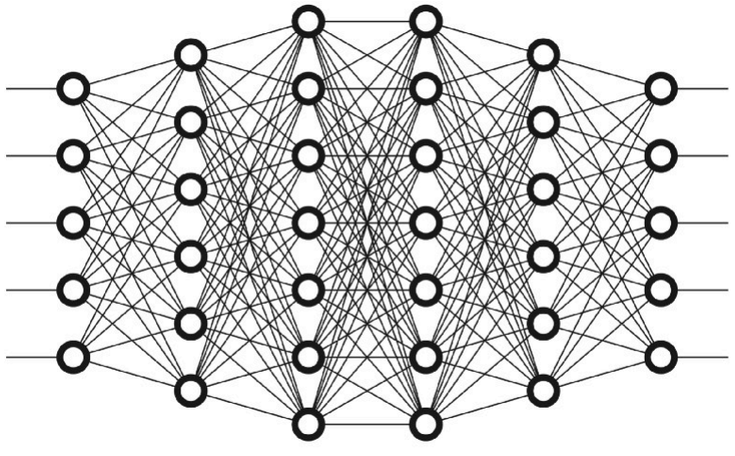
Part 1: Train a **big** language model

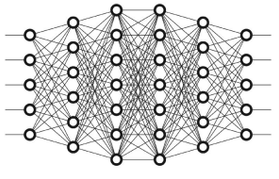


The more weights the model has, the more data it needs

Sentiment analysis

Part 2: use the weights to **initial** the weights of your model

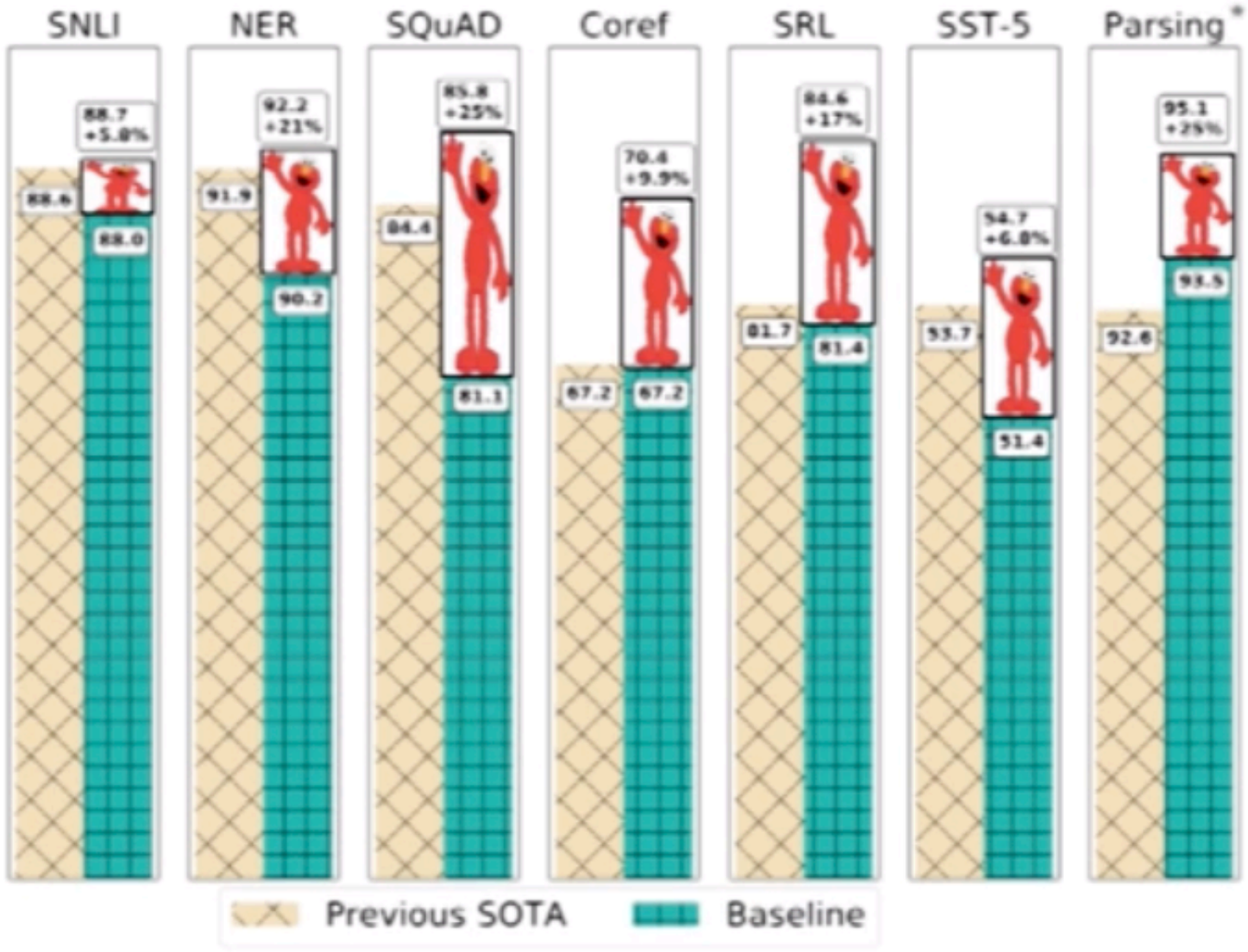




Contextual representations

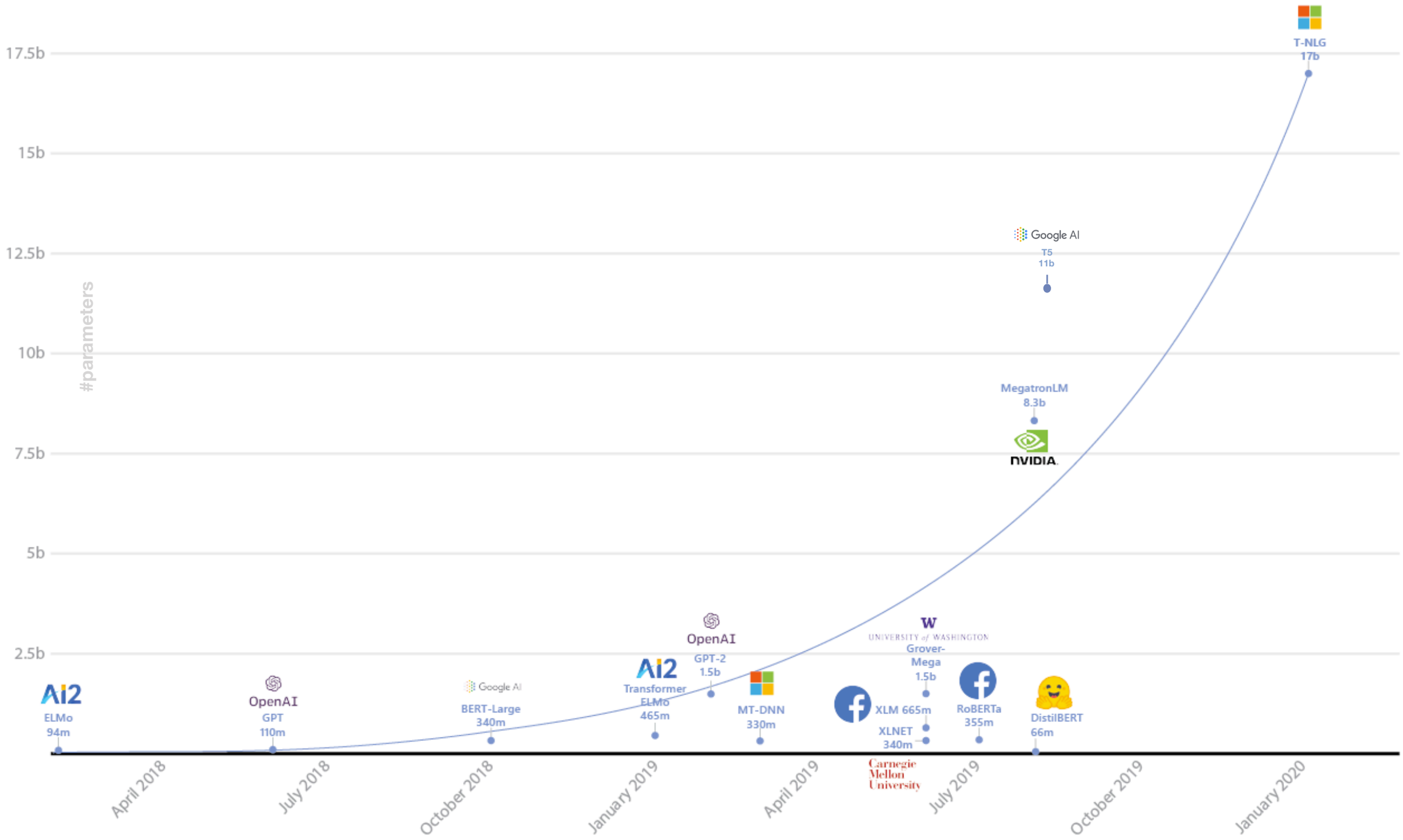
Peters et al., 2018

Michael Phelps, Beijing 2008,
8 gold medals



Big Models

170X in 2 Years

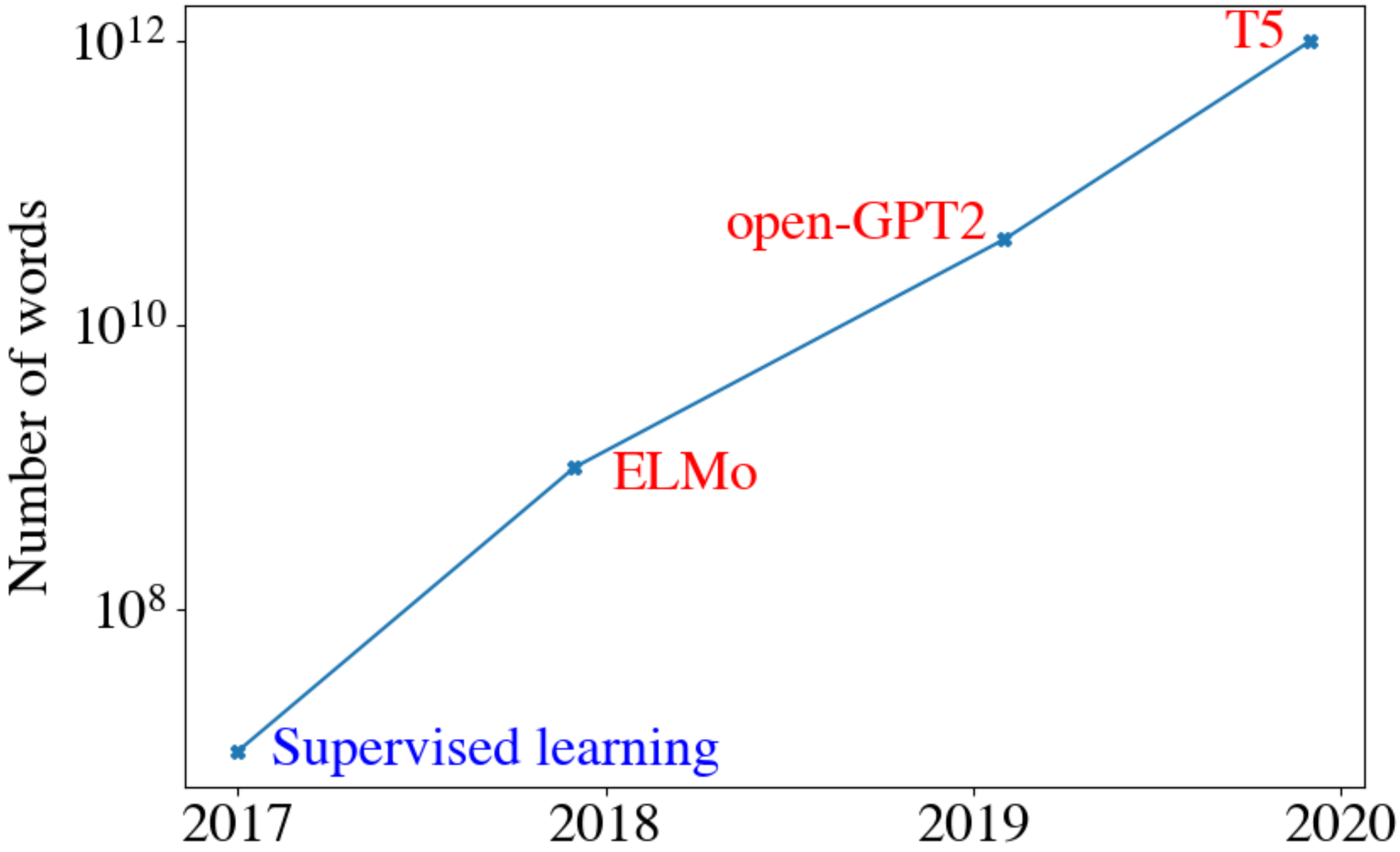


https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

https://medium.com/huggingface/distilbert-8cf3380435b5

More Data

100,000X in 3 Years!



Red AI

☑ ~~Bigger~~ **Huge** models

☑ ~~Larger~~ **XXL** datasets

Red AI is Expensive

- Grover (Zellers et al., 2019): $25,000
- XLNET (Yang et al., 2019): $60,000
- AlphaGO (Silver et al., 2016): $**35,000,000**

Outline

- Red AI

  - Why here? why now?
  - Big models, large datasets

- Green AI

  - Reporting, efficiency

Enhanced Reporting

- Better comparison
- More clarity and transparency
- Increased reproducibility

Evaluation in AI

- Run model against a collection of unseen test examples
- Compute some measure of performance
  - E.g., *accuracy*: proportion of examples the model got right

Model Comparison

Is **Model A > Model B**?

Is accuracy(A) > accuracy (B)

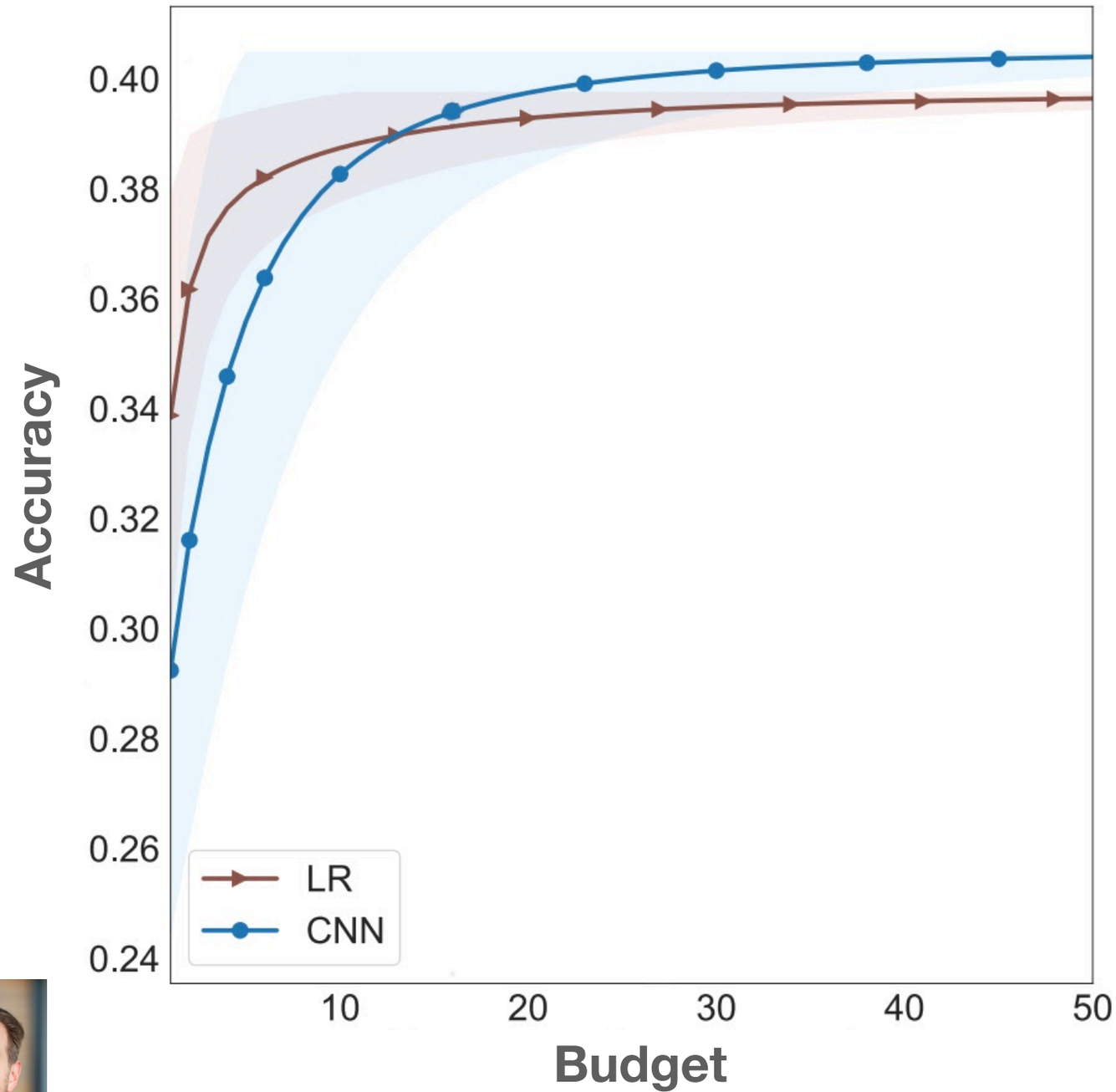
Better(?) Comparison

Is Model A > Model B?

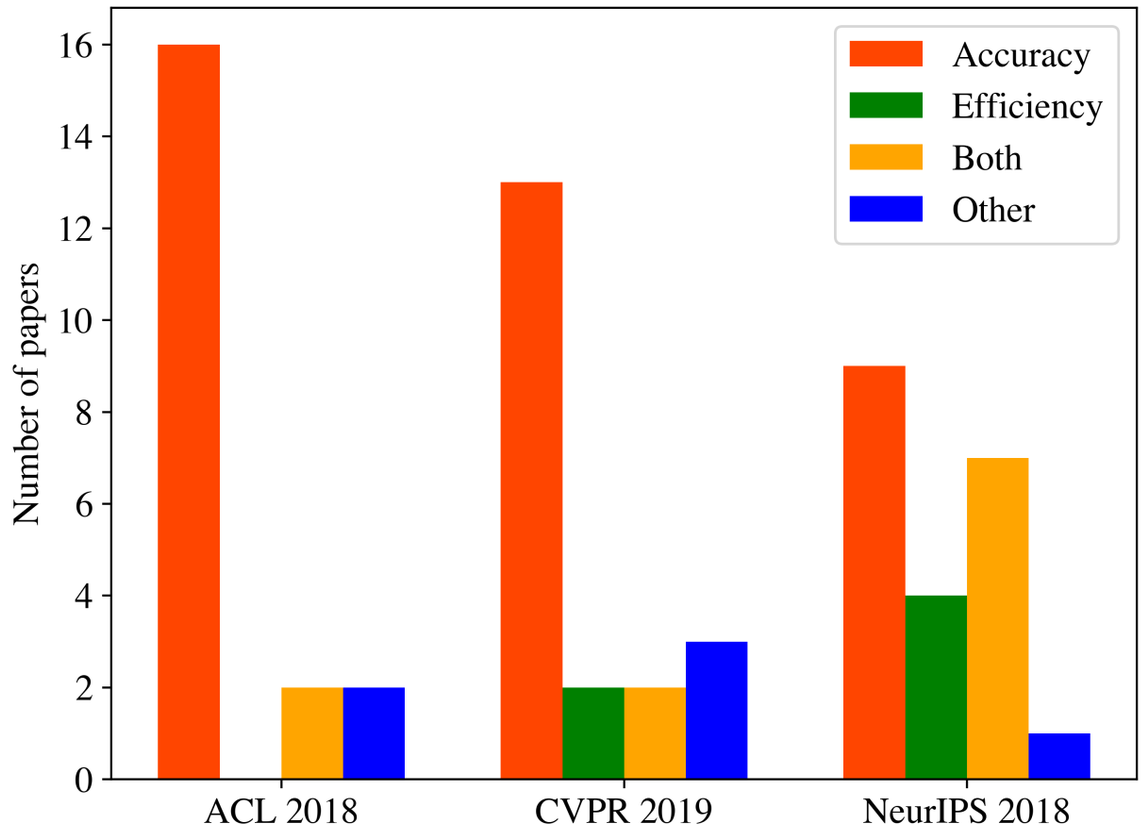
Is accuracy(A) > accuracy (B) ***given a certain Budget***

Budget-Aware Comparison

Dodge, Gururangan, Card, **Schwartz** & Smith, 2019



Accuracy or Efficiency?



Schwartz et al. (2019)

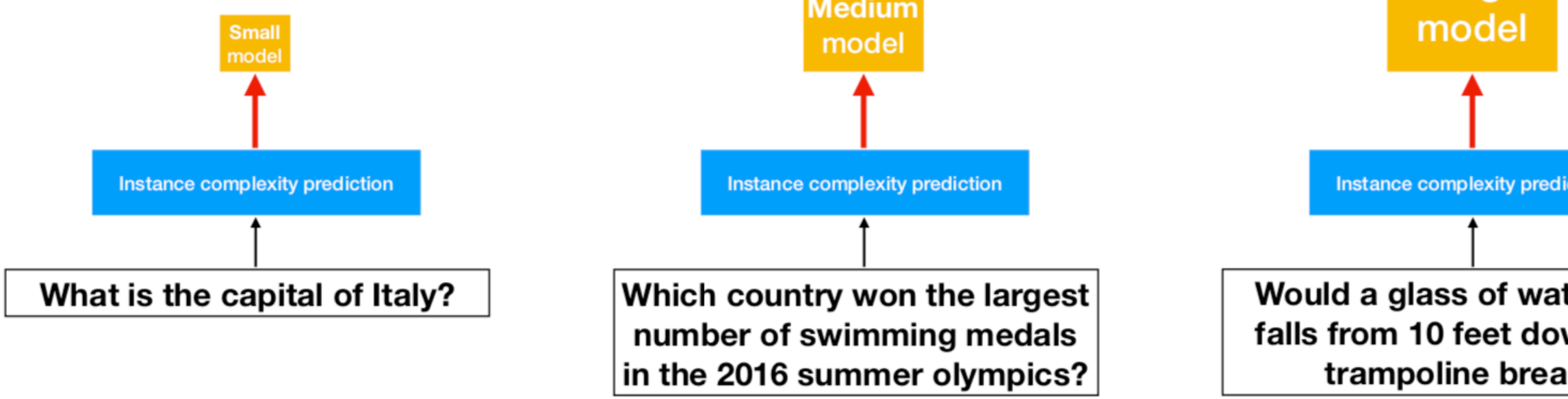
Efficiency 1:

Bigger -> Smaller

- Start by training a large model
- Use that model to train a smaller, more efficient model

Efficiency 2:

Schwartz et al., 2020



Open Research Questions

- Reporting

  - Can we predict how much will we gain by putting more compute?
  - What to report? ($, KW, CO_2, …)

- Efficiency

  - Training efficiency
  - Sample efficiency (train with less data)

Think Green

- Red AI
  - Big models, large datasets
  - Inclusiveness, adoption, environment
- Green AI

  - Enhance reporting of computational budgets
    - Add a *price-tag* for scientific results

  - Promote efficiency as a core evaluation for AI
    - **In addition to** accuracy